Gradient Descent with Random Initialization: Fast Global Convergence for Nonconvex Phase Retrieval

Yuxin Chen* Yuejie Chi† Jianqing Fan‡ Cong Ma‡

March 2018; Revised December 2018

Abstract

This paper considers the problem of solving systems of quadratic equations, namely, recovering an object of interest $\boldsymbol{x}^\natural \in \mathbb{R}^n$ from m quadratic equations / samples $y_i = (\boldsymbol{a}_i^\top \boldsymbol{x}^\natural)^2$, $1 \le i \le m$. This problem, also dubbed as phase retrieval, spans multiple domains including physical sciences and machine learning.

We investigate the efficacy of gradient descent (or Wirtinger flow) designed for the nonconvex least squares problem. We prove that under Gaussian designs, gradient descent — when randomly initialized — yields an ϵ-accurate solution in $O\big(\log n + \log(1/\epsilon)\big)$ iterations given nearly minimal samples, thus achieving near-optimal computational and sample complexities at once. This provides the first global convergence guarantee concerning vanilla gradient descent for phase retrieval, without the need of (i) carefully-designed initialization, (ii) sample splitting, or (iii) sophisticated saddle-point escaping schemes. All of these are achieved by exploiting the statistical models in analyzing optimization algorithms, via a leave-one-out approach that enables the decoupling of certain statistical dependency between the gradient descent iterates and the data.

Contents

Author names are sorted alphabetically.

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

†Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Email: `yuejiechi@cmu.edu`.

‡Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; Email: `{jqfan, congm}@princeton.edu`.

1 Introduction

Suppose we are interested in learning an unknown object $\boldsymbol{x}^{\natural} \in \mathbb{R}^n$, but only have access to a few quadratic equations of the form

$$y_i = \left(\boldsymbol{a}_i^{\top} \boldsymbol{x}^{\natural}\right)^2, \qquad 1 \leq i \leq m, \tag{1}$$

where y_i is the sample we collect and $\boldsymbol{a}_i$ is the design vector known *a priori*. Is it feasible to reconstruct $\boldsymbol{x}^{\natural}$ in an accurate and efficient manner?

The problem of solving systems of quadratic equations (1) is of fundamental importance and finds applications in numerous contexts. Perhaps one of the best-known applications is the so-called *phase retrieval* problem arising in physical sciences [CESV13, SEC+15]. In X-ray crystallography, due to the ultra-high frequency of the X-rays, the optical sensors and detectors are incapable of recording the phases of the diffractive waves; rather, only intensity measurements are collected. The phase retrieval problem comes down to reconstructing the specimen of interest given intensity-only measurements. If one thinks of $\boldsymbol{x}^{\natural}$ as the specimen under study and uses $\{y_i\}$ to represent the intensity measurements, then phase retrieval is precisely about inverting the quadratic system (1).

Moving beyond physical sciences, the above problem also spans various machine learning applications. One example is *mixed linear regression*, where one wishes to estimate two unknown vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ from unlabeled linear measurements [CYC14]. The acquired data $\{\boldsymbol{a}_i, b_i\}_{1\leq i\leq m}$ take the form of either $b_i \approx \boldsymbol{a}_i^{\top}\boldsymbol{\beta}_1$

or $b_i \approx \boldsymbol{a}_i^\top \boldsymbol{\beta}_2$, without knowing which of the two vectors generates the data. In a simple symmetric case with $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2 = \boldsymbol{x}^\natural$ (so that $b_i \approx \pm \boldsymbol{a}_i^\top \boldsymbol{x}^\natural$), the squared measurements $y_i = b_i^2 \approx (\boldsymbol{a}_i^\top \boldsymbol{x}^\natural)^2$ become the sufficient statistics, and hence mixed linear regression can be converted to learning $\boldsymbol{x}^\natural$ from $\{\boldsymbol{a}_i, y_i\}$. Furthermore, the quadratic measurement model in (1) allows to represent a single neuron associated with a quadratic activation function, where $\{\boldsymbol{a}_i, y_i\}$ are the data and $\boldsymbol{x}^\natural$ encodes the parameters to be learned. As described in [SJL17, LMZ17], *learning neural nets with quadratic activations* involves solving systems of quadratic equations.

1.1 Nonconvex optimization via gradient descent

A natural strategy for inverting the system of quadratic equations (1) is to solve the following nonconvex least squares estimation problem

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

Under Gaussian designs where $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, the solution to (2) is known to be exact — up to some global sign — with high probability, as soon as the number m of equations (samples) exceeds the order of the number n of unknowns [BCMN14]. However, the loss function in (2) is highly nonconvex, thus resulting in severe computational challenges. With this issue in mind, can we still hope to find the global minimizer of (2) via low-complexity algorithms which, ideally, run in time proportional to that taken to read the data?

Fortunately, in spite of nonconvexity, a variety of optimization-based methods are shown to be effective in the presence of proper statistical models. Arguably, one of the simplest algorithms for solving (2) is vanilla gradient descent (GD), which attempts recovery via the update rule

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f \left(\boldsymbol{x}^t \right), \qquad t = 0, 1, \cdots \tag{3}$$

with η_t being the stepsize / learning rate. The above iterative procedure is also dubbed *Wirtinger flow* for phase retrieval, which can accommodate the complex-valued case as well [CLS15]. This simple algorithm is remarkably efficient under Gaussian designs: in conjunction with carefully-designed initialization and stepsize rules, GD provably converges to the truth $\boldsymbol{x}^\natural$ at a linear rate[1], provided that the ratio m/n of the number of equations to the number of unknowns exceeds some logarithmic factor [CLS15, Sol14, MWCC17].

One crucial element in prior convergence analysis is initialization. In order to guarantee linear convergence, prior works typically recommend spectral initialization or its variants [CLS15, CC17, WGE17, ZZLC17, MWCC17, LL17, MM17]. Specifically, the spectral method forms an initial estimate $\boldsymbol{x}^0$ using the (properly scaled) leading eigenvector of a certain data matrix. Two important features are worth emphasizing:

- $\boldsymbol{x}^0$ falls within a local ℓ_2-ball surrounding $\boldsymbol{x}^\natural$ with a reasonably small radius, where $f(\cdot)$ enjoys strong convexity;
- $\boldsymbol{x}^0$ is incoherent with all the design vectors $\{\boldsymbol{a}_i\}$ — in the sense that $|\boldsymbol{a}_i^\top \boldsymbol{x}^0|$ is reasonably small for all $1 \leq i \leq m$ — and hence $\boldsymbol{x}^0$ falls within a region where $f(\cdot)$ enjoys desired smoothness conditions.

These two properties taken collectively allow gradient descent to converge rapidly from the very beginning.

1.2 Random initialization?

The enormous success of spectral initialization gives rise to a curious question: is carefully-designed initialization necessary for achieving fast convergence? Obviously, vanilla GD cannot start from arbitrary points, since it may get trapped in undesirable stationary points (e.g. saddle points). However, is there any *simpler* initialization approach that avoids such stationary points and works equally well as spectral initialization?

A strategy that practitioners often like to employ is to initialize GD randomly. The advantage is clear: compared with spectral methods, random initialization is model-agnostic and is usually more robust vis-a-vis model mismatch. Despite its wide use in practice, however, GD with random initialization is poorly

[1] An iterative algorithm is said to enjoy linear convergence if the iterates $\{\boldsymbol{x}^t\}$ converge geometrically fast to the minimizer $\boldsymbol{x}^\natural$.

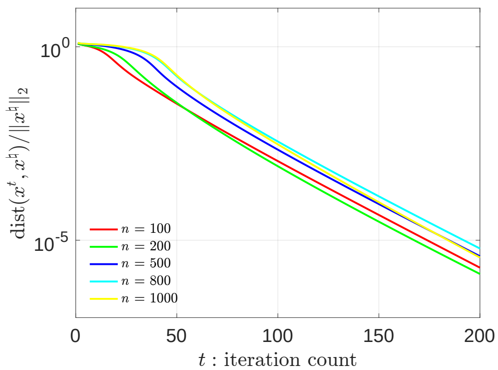


Figure 1: The relative ℓ_2 error vs. iteration count for GD with random initialization, plotted semi-logarithmically. The results are shown for $n = 100, 200, 500, 800, 1000$ with $m = 10n$ and $\eta_t \equiv 0.1$.

understood in theory. One way to study this method is through a geometric lens [SQW16]: under Gaussian designs, the loss function $f(\cdot)$ (cf. (2)) does not have any spurious local minima as long as the sample size m is on the order of $n \log^3 n$. Moreover, all saddle points are strict [GHJY15], meaning that the associated Hessian matrices have at least one negative eigenvalue if they are not local minima. Armed with these two conditions, the theory of Lee et al. [LSJR16] implies that vanilla GD converges *almost surely* to the truth. However, the convergence rate remains unsettled. In fact, we are not aware of any theory that guarantees polynomial-time convergence of vanilla GD for phase retrieval in the absence of carefully-designed initialization.

Motivated by this, we aim to pursue a formal understanding about the convergence properties of GD with random initialization. Before embarking on theoretical analyses, we first assess its practical efficiency through numerical experiments. Generate the true object $\boldsymbol{x}^\natural$ and the initial guess $\boldsymbol{x}^0$ randomly as

$$\boldsymbol{x}^\natural \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n) \qquad \text{and} \qquad \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n).$$

We vary the number n of unknowns (i.e. $n = 100, 200, 500, 800, 1000$), set $m = 10n$, and take a constant stepsize $\eta_t \equiv 0.1$. Here the measurement vectors are generated from Gaussian distributions, i.e. $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for $1 \leq i \leq m$. The relative ℓ_2 errors $\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural)/\|\boldsymbol{x}^\natural\|_2$ of the GD iterates in a random trial are plotted in Figure 1, where

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) := \min\left\{\|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2, \|\boldsymbol{x}^t + \boldsymbol{x}^\natural\|_2\right\} \tag{4}$$

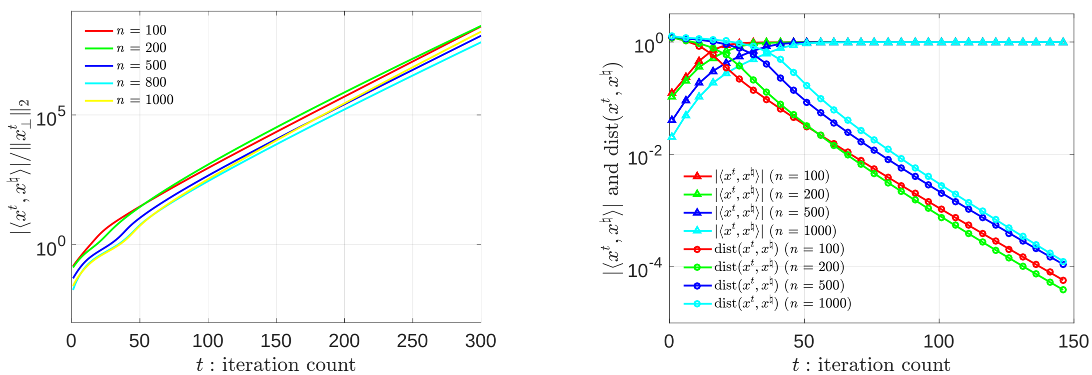
represents the ℓ_2 distance between $\boldsymbol{x}^t$ and $\boldsymbol{x}^\natural$ modulo the unrecoverable global sign.

In all experiments carried out in Figure 1, we observe two stages for GD: (1) Stage 1: the relative error of $\boldsymbol{x}^t$ stays nearly flat; (2) Stage 2: the relative error of $\boldsymbol{x}^t$ experiences geometric decay. Interestingly, Stage 1 lasts only for a few tens of iterations. These numerical findings taken together reveal appealing computational efficiency of GD in the presence of random initialization — it attains 5-digit accuracy within about 200 iterations!

To further illustrate this point, we take a closer inspection of the signal component $\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural$ and the orthogonal component $\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural$, where we normalize $\|\boldsymbol{x}^\natural\|_2 = 1$ for simplicity. Denote by $\|\boldsymbol{x}^t_\perp\|_2$ the ℓ_2 norm of the orthogonal component. We highlight two important and somewhat surprising observations that allude to why random initialization works.

- *The strength ratio of the signal to the orthogonal components grows exponentially.* The ratio, $|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle| \,/\, \|\boldsymbol{x}^t_\perp\|_2$, grows exponentially fast throughout the execution of the algorithm, as demonstrated in Figure 2(a). This metric $|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle| \,/\, \|\boldsymbol{x}^t_\perp\|_2$ in some sense captures the signal-to-noise ratio of the running iterates.

- *Exponential growth of the signal strength in Stage 1.* While the ℓ_2 estimation error of $\boldsymbol{x}^t$ may not drop significantly during Stage 1, the size $|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle|$ of the signal component increases exponentially fast and becomes the dominant component within several tens of iterations, as demonstrated in Figure 2(b). This helps explain why Stage 1 lasts only for a short duration.



(a) (b)

Figure 2: (a) The ratio $|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural\rangle| \,/\, \|\boldsymbol{x}^t_\perp\|_2$, and (b) the size $|\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural\rangle|$ of the signal component and the ℓ_2 error vs. iteration count, both plotted on semilogarithmic scales. The results are shown for $n = 100, 200, 500, 800, 1000$ with $m = 10n$, $\eta_t \equiv 0.1$, and $\|\boldsymbol{x}^\natural\|_2 = 1$.

The central question then amounts to whether one can develop a mathematical theory to interpret such intriguing numerical performance. In particular, how many iterations does Stage 1 encompass, and how fast can the algorithm converge in Stage 2?

1.3 Main findings

The objective of the current paper is to demystify the computational efficiency of GD with random initialization, thus bridging the gap between theory and practice. Assuming a tractable random design model in which $\boldsymbol{a}_i$'s follow Gaussian distributions, our main findings are summarized in the following theorem. Here and throughout, the notation $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ (resp. $f(n) \gtrsim g(n)$, $f(n) \asymp g(n)$) means that there exist constants $c_1, c_2 > 0$ such that $f(n) \leq c_1 g(n)$ (resp. $f(n) \geq c_2 g(n)$, $c_1 g(n) \leq f(n) \leq c_2 g(n)$).

Theorem 1. *Fix $\boldsymbol{x}^\natural \in \mathbb{R}^n$ with $\|\boldsymbol{x}^\natural\|_2 = 1$. Suppose that $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for $1 \leq i \leq m$, $\boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$, and $\eta_t \equiv \eta = c / \|\boldsymbol{x}^\natural\|_2^2$ for some sufficiently small constant $c > 0$. Then with probability approaching one, there exist some sufficiently small constant $0 < \gamma < 1$ and $T_\gamma \lesssim \log n$ such that the GD iterates (3) obey*

$$\mathrm{dist}\big(\boldsymbol{x}^t, \boldsymbol{x}^\natural\big) \leq \gamma(1-\rho)^{t - T_\gamma}, \qquad \forall\, t \geq T_\gamma$$

for some absolute constant $0 < \rho < 1$, provided that the sample size $m \gtrsim n \,\mathrm{poly}\log(m)$.

Remark 1. The readers are referred to Theorem 2 for a more general statement.

Here, the stepsize is taken to be a fixed constant throughout all iterations, and we reuse the same data across all iterations (i.e. no sample splitting is needed to establish this theorem). The GD trajectory is divided into 2 stages: (1) Stage 1 consists of the first T_γ iterations, corresponding to the first tens of iterations discussed in Section 1.2; (2) Stage 2 consists of all remaining iterations, where the estimation error contracts linearly. Several important implications / remarks follow immediately.

- *Stage 1 takes $O(\log n)$ iterations.* When seeded with a random initial guess, GD is capable of entering a local region surrounding $\boldsymbol{x}^\natural$ within $T_\gamma \lesssim \log n$ iterations, namely,

$$\mathrm{dist}\big(\boldsymbol{x}^{T_\gamma}, \boldsymbol{x}^\natural\big) \leq \gamma$$

  for some sufficiently small constant $\gamma > 0$. Even though Stage 1 may not enjoy linear convergence in terms of the estimation error, it is of fairly short duration.

- *Stage 2 takes $O(\log(1/\epsilon))$ iterations.* After entering the local region, GD converges linearly to the ground truth $\boldsymbol{x}^\natural$ with a contraction rate $1 - \rho$. This tells us that GD reaches ϵ-accuracy (in a relative sense) within $O(\log(1/\epsilon))$ iterations.

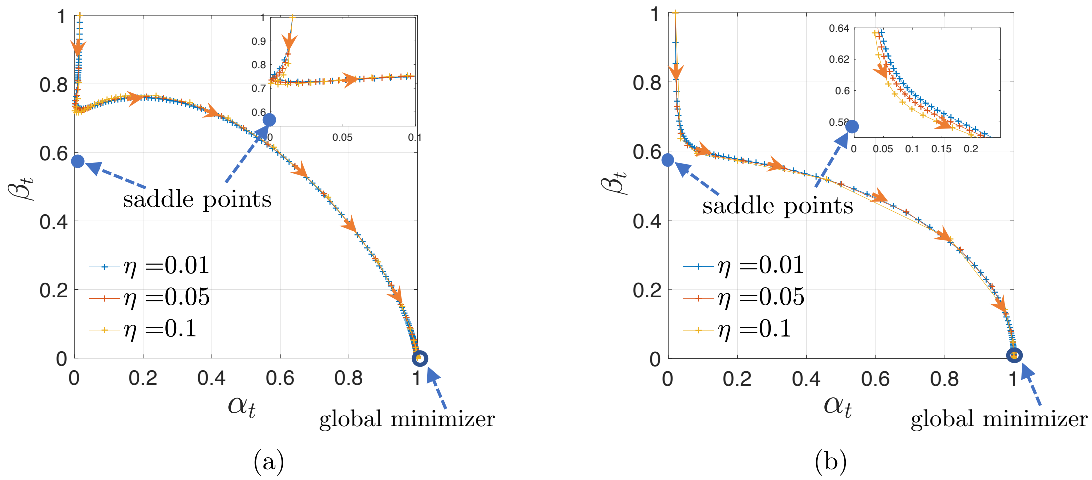


Figure 3: The trajectory of (α_t, β_t), where $\alpha_t = |\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle|$ and $\beta_t = \|\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural\|_2$ represent respectively the size of the signal component and that of the orthogonal component of the GD iterates (assume $\|\boldsymbol{x}^\natural\|_2 = 1$). (a) The results are shown for $n = 1000$ with $m = 10n$, and $\eta_t = 0.01, 0.05, 0.1$. (b) The results are shown for $n = 1000$ with m approaching infinity, and $\eta_t = 0.01, 0.05, 0.1$. The blue filled circles represent the population-level saddle points, and the orange arrows indicate the directions of increasing t.

- *Near linear-time computational complexity.* Taken collectively, these imply that the iteration complexity of GD with random initialization is
$$O\left(\log n + \log \frac{1}{\epsilon}\right).$$
Given that the cost of each iteration mainly lies in calculating the gradient $\nabla f(\boldsymbol{x}^t)$, the whole algorithm takes nearly linear time, namely, it enjoys a computational complexity proportional to the time taken to read the data (modulo some logarithmic factor).

- *Near-minimal sample complexity.* The preceding computational guarantees occur as soon as the sample size exceeds $m \gtrsim n \operatorname{poly}\log(m)$. Given that one needs at least n samples to recover n unknowns, the sample complexity of randomly initialized GD is optimal up to some logarithmic factor.

- *Saddle points?* The GD iterates never hit the saddle points (see Figure 3 for an illustration). In fact, after a constant number of iterations at the very beginning, GD will follow a path that increasingly distances itself from the set of saddle points as the algorithm progresses. There is no need to adopt sophisticated saddle-point escaping schemes developed in generic optimization theory (e.g. cubic regularization [NP06], perturbed GD [JGN+17]).

- *Weak dependency w.r.t. the design vectors.* As we will elaborate in Section 4, the statistical dependency between the GD iterates $\{\boldsymbol{x}^t\}$ and certain components of the design vectors $\{\boldsymbol{a}_i\}$ stays at an exceedingly weak level. Consequently, the GD iterates $\{\boldsymbol{x}^t\}$ proceed *as if* fresh samples were employed in each iteration. This statistical observation plays a crucial role in characterizing the dynamics of the algorithm without the need of sample splitting.

It is worth emphasizing that the entire trajectory of GD is automatically confined within a certain region enjoying favorable geometry. For example, the GD iterates are always incoherent with the design vectors, stay sufficiently away from any saddle point, and exhibit desired smoothness conditions, which we will formalize in Section 4. Such delicate geometric properties underlying the GD trajectory are not explained by prior papers. In light of this, convergence analysis based on global geometry [SQW16] — which provides valuable insights into algorithm designs with *arbitrary* initialization — results in suboptimal (or even pessimistic) computational guarantees when analyzing a specific algorithm like GD. In contrast, the current paper establishes near-optimal performance guarantees by paying particular attention to finer dynamics of the algorithm. As will be seen later, this is accomplished by heavily exploiting the statistical properties in each iterative update.

2 Why random initialization works?

Before diving into the proof of the main theorem, we pause to develop intuitions regarding why gradient descent with random initialization is expected to work. We will build our understanding step by step: (i) we first investigate the dynamics of the population gradient sequence (the case where we have infinite samples); (ii) we then turn to the finite-sample case and present a heuristic argument assuming independence between the iterates and the design vectors; (iii) finally, we argue that the true trajectory is remarkably close to the one heuristically analyzed in the previous step, which arises from a key property concerning the "near-independence" between $\{\boldsymbol{x}^t\}$ and the design vectors $\{\boldsymbol{a}_i\}$.

Without loss of generality, we assume $\boldsymbol{x}^\natural = \boldsymbol{e}_1$ throughout this section, where $\boldsymbol{e}_1$ denotes the first standard basis vector. For notational simplicity, we denote by

$$x_\parallel^t := x_1^t \qquad \text{and} \qquad \boldsymbol{x}_\perp^t := [x_i^t]_{2\le i\le n} \tag{5}$$

the first entry and the 2nd through the nth entries of $\boldsymbol{x}^t$, respectively. Since $\boldsymbol{x}^\natural = \boldsymbol{e}_1$, it is easily seen that

$$\underbrace{x_\parallel^t \boldsymbol{e}_1 = \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural\rangle \boldsymbol{x}^\natural}_{\text{signal component}} \qquad \text{and} \qquad \underbrace{\begin{bmatrix} 0 \\ \boldsymbol{x}_\perp^t \end{bmatrix} = \boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural\rangle \boldsymbol{x}^\natural}_{\text{orthogonal component}} \tag{6}$$

represent respectively the components of $\boldsymbol{x}^t$ along and orthogonal to the signal direction. In what follows, we focus our attention on the following two quantities that reflect the sizes of the preceding two components[2]

$$\alpha_t := x_\parallel^t \qquad \text{and} \qquad \beta_t := \left\|\boldsymbol{x}_\perp^t\right\|_2. \tag{7}$$

Without loss of generality, assume that $\alpha_0 > 0$.

2.1 Population dynamics

To start with, we consider the unrealistic case where the iterates $\{\boldsymbol{x}^t\}$ are constructed using the population gradient (or equivalently, the gradient when the sample size m approaches infinity), i.e.

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t).$$

Here, $\nabla F(\boldsymbol{x})$ represents the population gradient given by

$$\nabla F(\boldsymbol{x}) := (3\|\boldsymbol{x}\|_2^2 - 1)\boldsymbol{x} - 2(\boldsymbol{x}^{\natural\top}\boldsymbol{x})\boldsymbol{x}^\natural,$$

which can be computed by $\nabla F(\boldsymbol{x}) = \mathbb{E}[\nabla f(\boldsymbol{x})] = \mathbb{E}\big[\{(\boldsymbol{a}_i^\top \boldsymbol{x})^2 - (\boldsymbol{a}_i^\top \boldsymbol{x}^\natural)^2\}\boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{x}\big]$ assuming that $\boldsymbol{x}$ and the $\boldsymbol{a}_i$'s are independent. Simple algebraic manipulation reveals the dynamics for both the signal and the orthogonal components:

$$x_\parallel^{t+1} = \left\{1 + 3\eta\left(1 - \|\boldsymbol{x}^t\|_2^2\right)\right\} x_\parallel^t; \tag{8a}$$

$$\boldsymbol{x}_\perp^{t+1} = \left\{1 + \eta\left(1 - 3\|\boldsymbol{x}^t\|_2^2\right)\right\} \boldsymbol{x}_\perp^t. \tag{8b}$$

Assuming that η is sufficiently small and recognizing that $\|\boldsymbol{x}^t\|_2^2 = \alpha_t^2 + \beta_t^2$, we arrive at the following population-level state evolution for both α_t and β_t (cf. (7)):

$$\alpha_{t+1} = \left\{1 + 3\eta\left[1 - \left(\alpha_t^2 + \beta_t^2\right)\right]\right\} \alpha_t; \tag{9a}$$

$$\beta_{t+1} = \left\{1 + \eta\left[1 - 3\left(\alpha_t^2 + \beta_t^2\right)\right]\right\} \beta_t. \tag{9b}$$

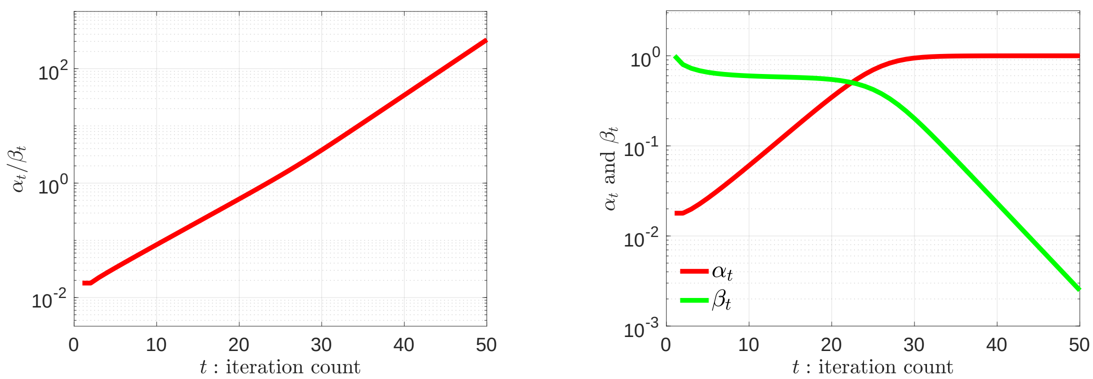
This recursive system has three *fixed points*:

$$(\alpha, \beta) = (1, 0), \qquad (\alpha, \beta) = (0, 0), \qquad \text{and} \qquad (\alpha, \beta) = (0, 1/\sqrt{3}),$$

which correspond to the global minimizer, the local maximizer, and the saddle points, respectively, of the population objective function.

We make note of the following key observations in the presence of a randomly initialized $\boldsymbol{x}^0$, which will be formalized later in Lemma 1:

[2]Here, we do not take the absolute value of $x_\parallel^t$. As we shall see later, the $x_\parallel^t$'s are of the same sign throughout the execution of the algorithm.



(a) α_t/β_t (b) α_t and β_t

Figure 4: Population-level state evolution, plotted semilogarithmically: (a) the ratio α_t/β_t vs. iteration count, and (b) α_t and β_t vs. iteration count. The results are shown for $n = 1000$, $\eta_t \equiv 0.1$, and $\boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$ (assuming $\alpha_0 > 0$ though).

- the ratio α_t/β_t of the size of the signal component to that of the orthogonal component increases exponentially fast;
- the size α_t of the signal component keeps growing until it plateaus around 1;
- the size β_t of the orthogonal component eventually drops towards zero.

In other words, when randomly initialized, (α^t, β^t) converges to $(1, 0)$ rapidly, thus indicating rapid convergence of $\boldsymbol{x}^t$ to the truth $\boldsymbol{x}^\natural$, without getting stuck at any undesirable saddle points. We also illustrate these phenomena numerically. Set $n = 1000$, $\eta_t \equiv 0.1$ and $\boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$. Figure 4 displays the dynamics of α_t/β_t, α_t, and β_t, which are precisely as discussed above.

2.2 Finite-sample analysis: a heuristic treatment

We now move on to the finite-sample regime, and examine how many samples are needed in order for the population dynamics to be reasonably accurate. Notably, the arguments in this subsection are heuristic in nature, but they are useful in developing insights into the true dynamics of the GD iterates.

Rewrite the gradient update rule (3) as

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) = \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t) - \eta \underbrace{\big(\nabla f(\boldsymbol{x}^t) - \nabla F(\boldsymbol{x}^t)\big)}_{:=\boldsymbol{r}(\boldsymbol{x}^t)}, \tag{10}$$

where $\nabla f(\boldsymbol{x}) = m^{-1}\sum_{i=1}^m [(\boldsymbol{a}_i^\top \boldsymbol{x})^2 - (\boldsymbol{a}_i^\top \boldsymbol{x}^\natural)^2]\boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{x}$. Assuming (unreasonably) that the iterate $\boldsymbol{x}^t$ is *independent of* $\{\boldsymbol{a}_i\}$, the central limit theorem (CLT) allows us to control the size of the fluctuation term $\boldsymbol{r}(\boldsymbol{x}^t)$. Take the signal component as an example: simple calculations give

$$x_{\|}^{t+1} = x_{\|}^t - \eta \left(\nabla F(\boldsymbol{x}^t)\right)_1 - \eta r_1(\boldsymbol{x}^t),$$

where

$$r_1(\boldsymbol{x}) := \frac{1}{m}\sum_{i=1}^m \left[\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^3 - a_{i,1}^2\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)\right] a_{i,1} - \mathbb{E}\left[\left\{\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^3 - a_{i,1}^2\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)\right\} a_{i,1}\right] \tag{11}$$

with $a_{i,1}$ the first entry of $\boldsymbol{a}_i$. Owing to the preceding independence assumption, r_1 is the sum of m i.i.d. zero-mean random variables. Assuming that $\boldsymbol{x}^t$ never blows up so that $\|\boldsymbol{x}^t\|_2 = O(1)$, one can apply the CLT to demonstrate that

$$|r_1(\boldsymbol{x}^t)| \lesssim \sqrt{\mathrm{Var}(r_1(\boldsymbol{x}^t))\,\mathrm{poly}\log(m)} \lesssim \sqrt{\frac{\mathrm{poly}\log(m)}{m}} \tag{12}$$

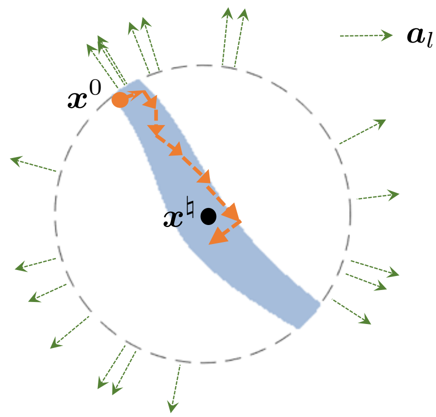


Figure 5: Illustration of the region satisfying the "near-independence" property. Here, the green arrows represent the directions of $\{\boldsymbol{a}_i\}_{1\leq i\leq 20}$, and the blue region consists of all points such that the first entry $r_1(\boldsymbol{x})$ of the fluctuation $\boldsymbol{r}(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \nabla F(\boldsymbol{x})$ is bounded above in magnitude by $|x_{||}^t|/5$ (or $|\langle \boldsymbol{x}, \boldsymbol{x}^\natural\rangle|/5$).

with high probability, which is often negligible compared to the other terms. For instance, for the random initial guess $\boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$ one has $\left|x_{||}^0\right| \gtrsim 1/\sqrt{n\log n}$ with probability approaching one, telling us that

$$|r_1(\boldsymbol{x}^0)| \lesssim \sqrt{\frac{\operatorname{poly}\log(m)}{m}} \ll |x_{||}^0|$$

as long as $m \gtrsim n \operatorname{poly}\log(m)$. This combined with the fact that $|x_{||}^0 - \eta(\nabla F(\boldsymbol{x}^0))_1| \asymp |x_{||}^0|$ reveals $|r_1(\boldsymbol{x}^0)| \lesssim |x_{||}^0 - \eta(\nabla F(\boldsymbol{x}^0))_1|$. Similar observations hold true for the orthogonal component $\boldsymbol{x}_\perp^t$.

In summary, by assuming independence between $\boldsymbol{x}^t$ and $\{\boldsymbol{a}_i\}$, we arrive at an approximate state evolution for the finite-sample regime:

$$\alpha_{t+1} \approx \left\{1 + 3\eta\left[1 - \left(\alpha_t^2 + \beta_t^2\right)\right]\right\}\alpha_t; \tag{13a}$$

$$\beta_{t+1} \approx \left\{1 + \eta\left[1 - 3\left(\alpha_t^2 + \beta_t^2\right)\right]\right\}\beta_t, \tag{13b}$$

with the proviso that $m \gtrsim n \operatorname{poly}\log(m)$.

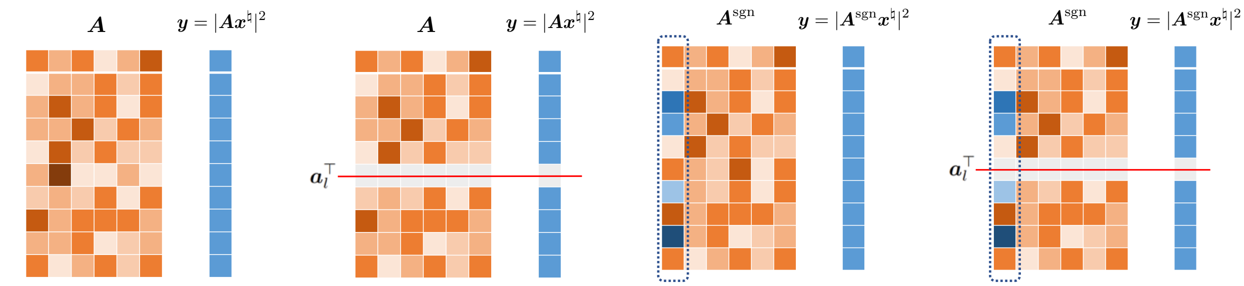
2.3 Key analysis ingredients: near-independence and leave-one-out tricks

The preceding heuristic argument justifies the approximate validity of the population dynamics, under an independence assumption that never holds unless we use fresh samples in each iteration. On closer inspection, what we essentially need is the fluctuation term $\boldsymbol{r}(\boldsymbol{x}^t)$ (cf. (10)) being well-controlled. For instance, when focusing on the signal component, one needs $|r_1(\boldsymbol{x}^t)| \ll \left|x_{||}^t\right|$ for all $t \geq 0$. In particular, in the beginning iterations, $|x_{||}^t|$ is as small as $O(1/\sqrt{n})$. Without the independence assumption, the CLT types of results fail to hold due to the complicated dependency between $\boldsymbol{x}^t$ and $\{\boldsymbol{a}_i\}$. In fact, one can easily find many points that result in much larger remainder terms (as large as $O(1)$) and that violate the approximate state evolution (13). See Figure 5 for a caricature of the region where the fluctuation term $\boldsymbol{r}(\boldsymbol{x}^t)$ is well-controlled. As can be seen, it only occupies a tiny fraction of the neighborhood of $\boldsymbol{x}^\natural$.

Fortunately, despite the complicated dependency across iterations, one can provably guarantee that $\boldsymbol{x}^t$ always stays within the preceding desirable region in which $\boldsymbol{r}(\boldsymbol{x}^t)$ is well-controlled. The key idea is to exploit a certain "near-independence" property between $\{\boldsymbol{x}^t\}$ and $\{\boldsymbol{a}_i\}$. Towards this, we make use of a leave-one-out trick proposed in [MWCC17] for analyzing nonconvex iterative methods. In particular, we construct auxiliary sequences that are

1. independent of *certain components* of the design vectors $\{\boldsymbol{a}_i\}$; and

2. extremely close to the original gradient sequence $\{\boldsymbol{x}^t\}_{t\geq 0}$.



(a) $\boldsymbol{x}^t$ (b) $\boldsymbol{x}^{t,(l)}$ (c) $\boldsymbol{x}^{t,\mathrm{sgn}}$ (d) $\boldsymbol{x}^{t,\mathrm{sgn},(l)}$

Figure 6: Illustration of the leave-one-out and random-sign sequences. (a) $\{\boldsymbol{x}^t\}$ is constructed using all data $\{\boldsymbol{a}_i, y_i\}$; (b) $\{\boldsymbol{x}^{t,(l)}\}$ is constructed by discarding the lth sample $\{\boldsymbol{a}_l, y_l\}$; (c) $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ is constructed by using auxiliary design vectors $\{\boldsymbol{a}_i^{\mathrm{sgn}}\}$, where $\boldsymbol{a}_i^{\mathrm{sgn}}$ is obtained by randomly flipping the sign of the first entry of $\boldsymbol{a}_i$; (d) $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}$ is constructed by discarding the lth sample $\{\boldsymbol{a}_l^{\mathrm{sgn}}, y_l\}$.

As it turns out, we need to construct several auxiliary sequences $\{\boldsymbol{x}^{t,(l)}\}_{t\geq 0}$, $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}_{t\geq 0}$ and $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}_{t\geq 0}$, where $\{\boldsymbol{x}^{t,(l)}\}_{t\geq 0}$ is independent of the lth sampling vector $\boldsymbol{a}_l$, $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}_{t\geq 0}$ is independent of the sign information of the first entries of all $\boldsymbol{a}_i$'s, and $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}$ is independent of both. In addition, these auxiliary sequences are constructed by slightly perturbing the original data (see Figure 6 for an illustration), and hence one can expect all of them to stay close to the original sequence throughout the execution of the algorithm. Taking these two properties together, one can propagate the above statistical independence underlying each auxiliary sequence to the true iterates $\{\boldsymbol{x}^t\}$, which in turn allows us to obtain near-optimal control of the fluctuation term $\boldsymbol{r}(\boldsymbol{x}^t)$. The details are postponed to Section 4.

3 Related work

Solving systems of quadratic equations, or phase retrieval, has been studied extensively in the recent literature; see [SEC+15] for an overview. One popular method is convex relaxation (e.g. *PhaseLift* [CSV13]), which is guaranteed to work as long as m/n exceeds some large enough constant [CL14, DH14, CCG15, CZ15, KRT17]. However, the resulting semidefinite program is computationally prohibitive for solving large-scale problems. To address this issue, [CLS15] proposed the Wirtinger flow algorithm with spectral initialization, which provides the first convergence guarantee for nonconvex methods without sample splitting. Both the sample and computation complexities were further improved by [CC17] with an adaptive truncation strategy. Other nonconvex phase retrieval methods include [NJS13, CLM16, Sol17, WGE17, ZZLC17, WGSC17, CL16, DR17, GX16, CFL15, Wei15, BEB17, TV17, CLW17, ZWGC17, QZEW17, ZCL16, YYF+17, CWZG17, Zha17, MXM18, CLC18]. Almost all of these nonconvex methods require carefully-designed initialization to guarantee a sufficiently accurate initial point. One exception is the approximate message passing algorithm proposed in [MXM18], which works as long as the correlation between the truth and the initial signal is bounded away from zero. This, however, does not accommodate the case when the initial signal strength is vanishingly small (like random initialization). Other works [Zha17, LGL15] explored the global convergence of alternating minimization / projection with random initialization which, however, require fresh samples at least in each of the first $O(\log n)$ iterations in order to enter the local basin. In addition, [LMZ17] explored low-rank recovery from quadratic measurements with near-zero initialization. Using a truncated least-squares objective, [LMZ17] established approximate (but non-exact) recovery of over-parametrized GD. Notably, if we do not over-parametrize the phase retrieval problem, then GD with near-zero initialization is (nearly) equivalent to running the power method for spectral initialization[3], which can be understood using prior theory.

Another related line of research is the design of generic saddle-point escaping algorithms, where the goal is to locate a second-order stationary point (i.e. the point with a vanishing gradient and a positive-semidefinite Hessian). As mentioned earlier, it has been shown by [SQW16] that as soon as $m \gg n\log^3 n$, all local minima

[3]More specifically, the GD update $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - m^{-1}\eta_t \sum_{i=1}^m \left[(\boldsymbol{a}_i^\top \boldsymbol{x}^t)^2 - y_i\right]\boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{x}_t \approx (\boldsymbol{I} + m^{-1}\eta_t \sum_{i=1}^m y_i \boldsymbol{a}_i\boldsymbol{a}_i^\top)\boldsymbol{x}_t$ when $\boldsymbol{x}_t \approx \boldsymbol{0}$, which is equivalent to a power iteration (without normalization) w.r.t. the data matrix $\boldsymbol{I} + m^{-1}\eta_t \sum_{i=1}^m y_i \boldsymbol{a}_i\boldsymbol{a}_i^\top$.

are global and all the saddle points are strict. With these two geometric properties in mind, saddle-point escaping algorithms are guaranteed to converge globally for phase retrieval. Existing saddle-point escaping algorithms include but are not limited to Hessian-based methods [NP06, SQW16] (see also [AAZB+16, AZ17, JGN+17] for some reviews), noisy stochastic gradient descent [GHJY15], perturbed gradient descent [JGN+17], and normalized gradient descent [MSK17]. On the one hand, the results developed in these works are fairly general: they establish polynomial-time convergence guarantees under a few generic geometric conditions. On the other hand, the iteration complexity derived therein may be pessimistic when specialized to a particular problem.

Take phase retrieval and the perturbed gradient descent algorithm [JGN+17] as an example. It has been shown in [JGN+17, Theorem 5] that for an objective function that is L-gradient Lipschitz, ρ-Hessian Lipschitz, (θ,γ,ζ)-strict saddle, and also locally α-strongly convex and β-smooth (see definitions in [JGN+17]), it takes[4]

$$O\left(\frac{L}{\left[\min\left(\theta,\gamma^2/\rho\right)\right]^2}+\frac{\beta}{\alpha}\log\frac{1}{\epsilon}\right)=O\left(n^3+n\log\frac{1}{\epsilon}\right)$$

iterations (ignoring logarithmic factors) for perturbed gradient descent to converge to ϵ-accuracy. In fact, even with Nesterov's accelerated scheme [JNJ17], the iteration complexity for entering the local region is at least

$$O\left(\frac{L^{1/2}\rho^{1/4}}{\left[\min\left(\theta,\gamma^2/\rho\right)\right]^{7/4}}\right)=O\left(n^{2.5}\right).$$

Both of them are much larger than the $O\big(\log n+\log(1/\epsilon)\big)$ complexity established herein. This is primarily due to the following facts: (i) the Lipschitz constants of both the gradients and the Hessians are quite large, i.e. $L\asymp n$ and $\rho\asymp n$ (ignoring log factors), which are, however, treated as dimension-independent constants in the aforementioned papers; (ii) the local condition number is also large, i.e. $\beta/\alpha\asymp n$. In comparison, as suggested by our theory, the GD iterates with random initialization are always confined within a restricted region enjoying much more benign geometry than the worst-case / global characterization.

Furthermore, the above saddle-escaping first-order methods are often more complicated than vanilla GD. Despite its algorithmic simplicity and wide use in practice, the convergence rate of GD with random initialization remains largely unknown. In fact, Du et al. [DJL+17] demonstrated that there exist non-pathological functions such that GD can take exponential time to escape the saddle points when initialized randomly. In contrast, as we have demonstrated, saddle points are not an issue for phase retrieval; the GD iterates with random initialization never get trapped in the saddle points.

Finally, the leave-one-out arguments have been invoked to analyze other high-dimensional statistical inference problems including robust M-estimators [EKBB+13, EK15], and maximum likelihood theory for logistic regression [SCC18], etc. In addition, [ZB17, CFMW17, AFWZ17] made use of the leave-one-out trick to derive entrywise perturbation bounds for eigenvectors resulting from certain spectral methods. The techniques have also been applied by [MWCC17, LMCC18] to establish local linear convergence of vanilla GD for nonconvex statistical estimation problems in the presence of proper spectral initialization.

4 Analysis

In this section, we first provide a more general version of Theorem 1 as follows. It spells out exactly the conditions on $\boldsymbol{x}^0$ in order for vanilla GD with random initialization to succeed.

Theorem 2. *Fix $\boldsymbol{x}^\natural\in\mathbb{R}^n$. Suppose $\boldsymbol{a}_i\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)$ $(1\le i\le m)$ and $m\ge Cn\log^{13}m$ for some sufficiently large constant $C>0$. Assume that the initialization $\boldsymbol{x}^0$ is independent of $\{\boldsymbol{a}_i\}$ and obeys*

$$\frac{\left|\langle\boldsymbol{x}^0,\boldsymbol{x}^\natural\rangle\right|}{\|\boldsymbol{x}^\natural\|_2^2}\ge\frac{1}{\sqrt{n\log n}}\qquad\text{and}\qquad\left(1-\frac{1}{\log n}\right)\|\boldsymbol{x}^\natural\|_2\le\|\boldsymbol{x}^0\|_2\le\left(1+\frac{1}{\log n}\right)\|\boldsymbol{x}^\natural\|_2,\tag{14}$$

and that the stepsize satisfies $\eta_t\equiv\eta=c/\|\boldsymbol{x}^\natural\|_2^2$ for some sufficiently small constant $c>0$. Then there exist a sufficiently small absolute constant $0<\gamma<1$ and $T_\gamma\lesssim\log n$ such that with probability at least $1-O(m^2e^{-1.5n})-O(m^{-9})$,

[4] When applied to phase retrieval with $m\asymp n\,\mathrm{poly}\log n$, one has $L\asymp n$, $\rho\asymp n$, $\theta\asymp\gamma\asymp 1$ (see [SQW16, Theorem 2.2]), $\alpha\asymp 1$, and $\beta\gtrsim n$ (ignoring logarithmic factors).

1. *the GD iterates (3) converge linearly to $\boldsymbol{x}^\natural$ after $t \geq T_\gamma$, namely,*

$$\operatorname{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \leq \left(1 - \frac{\eta}{2}\left\|\boldsymbol{x}^\natural\right\|_2^2\right)^{t-T_\gamma} \cdot \gamma \left\|\boldsymbol{x}^\natural\right\|_2, \qquad \forall\, t \geq T_\gamma;$$

2. *the strength ratio of the signal component $\frac{\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle}{\|\boldsymbol{x}^\natural\|_2^2}\boldsymbol{x}^\natural$ to the orthogonal component $\boldsymbol{x}^t - \frac{\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle}{\|\boldsymbol{x}^\natural\|_2^2}\boldsymbol{x}^\natural$ obeys*

$$\frac{\left\| \frac{\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle}{\|\boldsymbol{x}^\natural\|_2^2}\boldsymbol{x}^\natural \right\|_2}{\left\| \boldsymbol{x}^t - \frac{\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle}{\|\boldsymbol{x}^\natural\|_2^2}\boldsymbol{x}^\natural \right\|_2} \gtrsim \frac{1}{\sqrt{n \log n}}(1 + c_1\eta^2)^t, \qquad t = 0, 1, \cdots \tag{15}$$

for some constant $c_1 > 0$.

Several remarks regarding Theorem 2 are in order.

- Our current sample complexity reads $m \gtrsim n \log^{13} m$, which is optimal up to logarithmic factors. It is possible to further reduce the logarithmic factors using more refined probabilistic tools, which we leave for future work.
- We can also prove similar performance guarantees for noisy phase retrieval. For brevity, we do not provide the exact theorem and the detailed proofs. The readers will find them in the last author's Ph.D. thesis.
- The random initialization $\boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\|\boldsymbol{x}^\natural\|_2^2 \boldsymbol{I}_n)$ obeys the condition (14) with probability exceeding $1 - O(1/\sqrt{\log n})$, which in turn establishes Theorem 1.
- Theorem 2 requires an initialization $\boldsymbol{x}^0$ which is independent of the data and the knowledge of $\|\boldsymbol{x}^\natural\|$, which is not practical. One possible method is to estimate it from the data, which results in an initial value that depends on the data. The following theorem demonstrate both independent initial value and known $\|\boldsymbol{x}^\natural\|$ are not necessary, resulting a practical algorithm.

Theorem 3. *Let*

$$\boldsymbol{x}^0 = \sqrt{\frac{1}{m}\sum_{i=1}^m y_i} \cdot \boldsymbol{u},$$

where $\boldsymbol{u}$ is uniformly distributed over the unit sphere. With probability at least $1 - O(1/\sqrt{\log n})$ all the claims in Theorem 2 continue to hold.

Proof. The proof is very similar to that of Theorem 2, with only a few changes. See Appendix N for detailed explanations. □

The remainder of this section is then devoted to proving Theorem 2. Without loss of generality[5], we will assume throughout that

$$\boldsymbol{x}^\natural = \boldsymbol{e}_1 \qquad \text{and} \qquad x_1^0 > 0. \tag{16}$$

Given this, one can decompose

$$\boldsymbol{x}^t = x_\parallel^t \boldsymbol{e}_1 + \begin{bmatrix} 0 \\ \boldsymbol{x}_\perp^t \end{bmatrix} \tag{17}$$

where $x_\parallel^t = x_1^t$ and $\boldsymbol{x}_\perp^t = [x_i^t]_{2 \leq i \leq n}$ as introduced in Section 2. For notational simplicity, we define

$$\alpha_t := x_\parallel^t \qquad \text{and} \qquad \beta_t := \|\boldsymbol{x}_\perp^t\|_2. \tag{18}$$

Intuitively, α_t represents the size of the signal component, whereas β_t measures the size of the component orthogonal to the signal direction. In view of (16), we have $\alpha_0 > 0$.

[5]This is because of the rotational invariance of Gaussian distributions.

4.1 Outline of the proof

To begin with, it is easily seen that if α_t and β_t (cf. (18)) obey $|\alpha_t - 1| \leq \gamma/2$ and $\beta_t \leq \gamma/2$, then

$$\mathrm{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \leq \|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2 \leq \left|\alpha_t - 1\right| + \left|\beta_t\right| \leq \gamma.$$

Therefore, our first step — which is concerned with proving $\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma$ — comes down to the following two steps.

1. Show that if α_t and β_t satisfy the approximate state evolution (see (13)), then there exists some $T_\gamma = O(\log n)$ such that
$$\left|\alpha_{T_\gamma} - 1\right| \leq \gamma/2 \qquad \text{and} \qquad \beta_{T_\gamma} \leq \gamma/2, \tag{19}$$
which would immediately imply that
$$\mathrm{dist}\left(\boldsymbol{x}^{T_\gamma}, \boldsymbol{x}^\natural\right) \leq \gamma.$$
Along the way, we will also show that the ratio α_t/β_t grows exponentially fast.

2. Justify that α_t and β_t satisfy the approximate state evolution with high probability, using (some variants of) leave-one-out arguments.

After $t \geq T_\gamma$, we can invoke prior theory [MWCC17] concerning local convergence to show that with high probability,

$$\mathrm{dist}\left(\boldsymbol{x}^t, \boldsymbol{x}^\natural\right) \leq (1-\rho)^{t-T_\gamma} \|\boldsymbol{x}^{T_\gamma} - \boldsymbol{x}^\natural\|_2, \qquad \forall\, t > T_\gamma$$

for some constant $0 < \rho < 1$ independent of n and m.

4.2 Dynamics of approximate state evolution

This subsection formalizes our intuition in Section 2: as long as the approximate state evolution holds, then one can find $T_\gamma \lesssim \log n$ obeying condition (19). In particular, the approximate state evolution is given by

$$\alpha_{t+1} = \left\{1 + 3\eta\left[1 - \left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\zeta_t\right\}\alpha_t, \tag{20a}$$
$$\beta_{t+1} = \left\{1 + \eta\left[1 - 3\left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\rho_t\right\}\beta_t, \tag{20b}$$

where $\{\zeta_t\}$ and $\{\rho_t\}$ represent the perturbation terms. Our result is this:

Lemma 1. *Let $\gamma > 0$ be some sufficiently small constant, and consider the approximate state evolution (20). Suppose the initial point obeys*

$$\alpha_0 \geq \frac{1}{\sqrt{n \log n}} \qquad \textit{and} \qquad 1 - \frac{1}{\log n} \leq \sqrt{\alpha_0^2 + \beta_0^2} \leq 1 + \frac{1}{\log n}. \tag{21}$$

and the perturbation terms satisfy

$$\max\left\{\left|\zeta_t\right|, \left|\rho_t\right|\right\} \leq \frac{c_3}{\log n}, \qquad t = 0, 1, \cdots$$

for some sufficiently small constant $c_3 > 0$.

(a) Let

$$T_\gamma := \min\left\{t : |\alpha_t - 1| \leq \gamma/2 \text{ and } \beta_t \leq \gamma/2\right\}. \tag{22}$$

Then for any sufficiently large n and m and any sufficiently small constant $\eta > 0$, one has

$$T_\gamma \lesssim \log n, \tag{23}$$

and there exist some constants $c_5, c_{10} > 0$ independent of n and m such that

$$\frac{1}{2\sqrt{n \log n}} \leq \alpha_t \leq 2, \quad c_5 \leq \beta_t \leq 1.5 \quad \textit{and} \quad \frac{\alpha_{t+1}/\alpha_t}{\beta_{t+1}/\beta_t} \geq 1 + c_{10}\eta^2, \qquad 0 \leq t \leq T_\gamma. \tag{24}$$

(b) If we define

$$T_0 := \min\left\{t : \alpha_{t+1} \geq c_6/\log^5 m\right\}, \tag{25}$$
$$T_1 := \min\left\{t : \alpha_{t+1} > c_4\right\}, \tag{26}$$

for some arbitrarily small constants $c_4, c_6 > 0$, then

1) $T_0 \leq T_1 \leq T_\gamma \lesssim \log n$; $T_1 - T_0 \lesssim \log\log m$; $T_\gamma - T_1 \lesssim 1$;

2) For $T_0 < t \leq T_\gamma$, one has $\alpha_t \geq c_6/\log^5 m$.

Proof. See Appendix B. □

Remark 2. Recall that γ is sufficiently small and $(\alpha, \beta) = (1, 0)$ represents the global minimizer. Since $|\alpha_0 - 1| \approx 1$, one has $T_\gamma > 0$, which denotes the first time when the iterates enter the local region surrounding the global minimizer. In addition, the fact that $\alpha_0 \lesssim 1/\sqrt{n}$ gives $T_0 > 0$ and $T_1 > 0$, both of which indicate the first time when the signal strength is sufficiently large.

Lemma 1 makes precise that under the approximate state evolution, the first stage enjoys a fairly short duration $T_\gamma \lesssim \log n$. Moreover, the size of the signal component grows faster than that of the orthogonal component for any iteration $t < T_\gamma$, thus confirming the exponential growth of α_t/β_t.

In addition, Lemma 1 identifies two midpoints T_0 and T_1 when the sizes of the signal component α_t become sufficiently large. These are helpful in our subsequent analysis. In what follows, we will divide Stage 1 (which consists of all iterations up to T_γ) into two phases:

- *Phase I*: consider the duration $0 \leq t \leq T_0$;
- *Phase II*: consider all iterations with $T_0 < t \leq T_\gamma$.

We will justify the approximate state evolution (20) for these two phases separately.

4.3 Motivation of the leave-one-out approach

As we have alluded in Section 2.3, the main difficulty in establishing the approximate state evolution (20) lies in controlling the perturbation terms to the desired orders (i.e. $|\zeta_t|, |\rho_t| \ll 1/\log n$ in Lemma 1). To achieve this, we advocate the use of (some variants of) leave-one-out sequences to help establish certain "near-independence" between $\boldsymbol{x}^t$ and certain components of $\{\boldsymbol{a}_i\}$.

We begin by taking a closer look at the perturbation terms. Regarding the signal component, it is easily seen from (11) that

$$x_{\|}^{t+1} = \left\{1 + 3\eta\left(1 - \|\boldsymbol{x}^t\|_2^2\right)\right\} x_{\|}^t - \eta r_1(\boldsymbol{x}^t),$$

where the perturbation term $r_1(\boldsymbol{x}^t)$ obeys

$$\begin{aligned} r_1(\boldsymbol{x}^t) = &\underbrace{\left[1 - \left(x_{\|}^t\right)^2\right] x_{\|}^t \left(\frac{1}{m}\sum_{i=1}^m a_{i,1}^4 - 3\right)}_{:=I_1} + \underbrace{\left[1 - 3\left(x_{\|}^t\right)^2\right] \frac{1}{m}\sum_{i=1}^m a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t}_{:=I_2} \\ &\underbrace{- 3x_{\|}^t \left(\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 a_{i,1}^2 - \left\|\boldsymbol{x}_\perp^t\right\|_2^2\right)}_{:=I_3} - \underbrace{\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^3 a_{i,1}}_{:=I_4}. \end{aligned} \tag{27}$$

Here and throughout the paper, for any vector $\boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{v}_\perp \in \mathbb{R}^{n-1}$ denotes the 2nd through the nth entries of $\boldsymbol{v}$. Due to the dependency between $\boldsymbol{x}^t$ and $\{\boldsymbol{a}_i\}$, it is challenging to obtain sharp control of some of these terms.

In what follows, we use the term I_4 to explain and motivate our leave-one-out approach. As discussed in Section 2.3, I_4 needs to be controlled to the level $O(1/(\sqrt{n}\,\mathrm{poly}\log(n)))$. This precludes us from seeking a uniform bound on the function $h(\boldsymbol{x}) := m^{-1}\sum_{i=1}^m (\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp)^3 a_{i,1}$ over all $\boldsymbol{x}$ (or even all $\boldsymbol{x}$ within the set $\mathcal{C}$

Algorithm 1 The lth leave-one-out sequence

Input: $\{\boldsymbol{a}_i\}_{1\le i\le m, i\neq l}$, $\{y_i\}_{1\le i\le m, i\neq l}$, and $\boldsymbol{x}^0$.
Gradient updates: **for** $t = 0, 1, 2, \dots, T-1$ **do**

$$\boldsymbol{x}^{t+1,(l)} = \boldsymbol{x}^{t,(l)} - \eta_t \nabla f^{(l)}(\boldsymbol{x}^{t,(l)}), \tag{29}$$

where $\boldsymbol{x}^{0,(l)} = \boldsymbol{x}^0$ and $f^{(l)}(\boldsymbol{x}) = (1/4m)\cdot\sum_{i:i\neq l}[(\boldsymbol{a}_i^\top \boldsymbol{x})^2 - (\boldsymbol{a}_i^\top \boldsymbol{x}^\natural)^2]^2$.

---

incoherent with $\{\boldsymbol{a}_i\}$), since the uniform bound $\sup_{\boldsymbol{x}\in\mathcal{C}} |h(\boldsymbol{x})|$ can be $O(\sqrt{n}/\text{poly}\log(n))$ times larger than the desired order.

In order to control I_4 to the desirable order, one strategy is to approximate it by a sum of independent variables and then invoke the CLT. Specifically, we first rewrite I_4 as

$$I_4 = \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^3 |a_{i,1}|\, \xi_i$$

with $\xi_i := \text{sgn}(a_{i,1})$. Here $\text{sgn}(\cdot)$ denotes the usual sign function. To exploit the statistical independence between ξ_i and $\{|a_{i,1}|, \boldsymbol{a}_{i,\perp}\}$, we would like to identify some vector independent of ξ_i that well approximates $\boldsymbol{x}^t$. If this can be done, then one may treat I_4 as a weighted independent sum of $\{\xi_i\}$. Viewed in this light, our plan is the following:

1. Construct a sequence $\{\boldsymbol{x}^{t,\text{sgn}}\}$ independent of $\{\xi_i\}$ obeying $\boldsymbol{x}^{t,\text{sgn}} \approx \boldsymbol{x}^t$, so that

$$I_4 \approx \frac{1}{m}\sum_{i=1}^m \underbrace{\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\text{sgn}}\right)^3 |a_{i,1}|}_{:=w_i} \xi_i.$$

   One can then apply standard concentration results (e.g. the Bernstein inequality) to control I_4, as long as none of the weight w_i is exceedingly large.

2. Demonstrate that the weight w_i is well-controlled, or equivalently, $\left|\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\text{sgn}}\right|$ $(1 \le i \le m)$ is not much larger than its typical size. This can be accomplished by identifying another sequence $\{\boldsymbol{x}^{t,(i)}\}$ independent of $\boldsymbol{a}_i$ such that $\boldsymbol{x}^{t,(i)} \approx \boldsymbol{x}^t \approx \boldsymbol{x}^{t,\text{sgn}}$, followed by the argument:

$$\left|\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\text{sgn}}\right| \approx \left|\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t}\right| \approx \left|\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,(i)}\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}_\perp^{t,(i)}\right\|_2 \approx \sqrt{\log m}\left\|\boldsymbol{x}_\perp^{t}\right\|_2. \tag{28}$$

   Here, the inequality follows from standard Gaussian tail bounds and the independence between $\boldsymbol{a}_i$ and $\boldsymbol{x}^{t,(i)}$. This explains why we would like to construct $\{\boldsymbol{x}^{t,(i)}\}$ for each $1 \le i \le m$.

As we will detail in the next subsection, such auxiliary sequences are constructed by leaving out a small amount of relevant information from the collected data before running the GD algorithm, which is a variant of the "leave-one-out" approach rooted in probability theory and random matrix theory.

4.4 Leave-one-out and random-sign sequences

We now describe how to design auxiliary sequences to help establish certain independence properties between the gradient iterates $\{\boldsymbol{x}^t\}$ and the design vectors $\{\boldsymbol{a}_i\}$. In the sequel, we formally define the three sets of auxiliary sequences $\{\boldsymbol{x}^{t,(l)}\}, \{\boldsymbol{x}^{t,\text{sgn}}\}, \{\boldsymbol{x}^{t,\text{sgn},(l)}\}$ as introduced in Section 2.3 and Section 4.3.

- *Leave-one-out sequences* $\{\boldsymbol{x}^{t,(l)}\}_{t\ge 0}$. For each $1 \le l \le m$, we introduce a sequence $\{\boldsymbol{x}^{t,(l)}\}$, which drops the lth sample and runs GD w.r.t. the auxiliary objective function

$$f^{(l)}(\boldsymbol{x}) = \frac{1}{4m}\sum_{i:i\neq l}\left[\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2\right]^2. \tag{32}$$

Algorithm 2 The random-sign sequence

Input: $\{|a_{i,1}|\}_{1\leq i\leq m}$, $\{\boldsymbol{a}_{i,\perp}\}_{1\leq i\leq m}$, $\{\xi_i^{\mathrm{sgn}}\}_{1\leq i\leq m}$, $\{y_i\}_{1\leq i\leq m}$, $\boldsymbol{x}^0$.
Gradient updates: **for** $t=0,1,2,\ldots,T-1$ **do**

$$\boldsymbol{x}^{t+1,\mathrm{sgn}} = \boldsymbol{x}^{t,\mathrm{sgn}} - \eta_t \nabla f^{\mathrm{sgn}}(\boldsymbol{x}^{t,\mathrm{sgn}}), \tag{30}$$

where $\boldsymbol{x}^{0,\mathrm{sgn}} = \boldsymbol{x}^0$, $f^{\mathrm{sgn}}(\boldsymbol{x}) = \frac{1}{4m}\sum_{i=1}^m [(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x})^2 - (\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^\natural)^2]^2$ with $\boldsymbol{a}_i^{\mathrm{sgn}} := \begin{bmatrix} \xi_i^{\mathrm{sgn}}|a_{i,1}| \\ \boldsymbol{a}_{i,\perp} \end{bmatrix}$.

Algorithm 3 The lth leave-one-out and random-sign sequence

Input:$\{|a_{i,1}|\}_{1\leq i\leq m, i\neq l}$, $\{\boldsymbol{a}_{i,\perp}\}_{1\leq i\leq m, i\neq l}$, $\{\xi_i^{\mathrm{sgn}}\}_{1\leq i\leq m,, i\neq l}$, $\{y_i\}_{1\leq i\leq m, i\neq l}$, $\boldsymbol{x}^0$.
Gradient updates: **for** $t=0,1,2,\ldots,T-1$ **do**

$$\boldsymbol{x}^{t+1,\mathrm{sgn},(l)} = \boldsymbol{x}^{t,\mathrm{sgn},(l)} - \eta_t \nabla f^{\mathrm{sgn},(l)}(\boldsymbol{x}^{t,\mathrm{sgn},(l)}), \tag{31}$$

where $\boldsymbol{x}^{0,\mathrm{sgn},(l)} = \boldsymbol{x}^0$, $f^{\mathrm{sgn},(l)}(\boldsymbol{x}) = \frac{1}{4m}\sum_{i:i\neq l}\left[\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^\natural\right)^2\right]^2$ with $\boldsymbol{a}_i^{\mathrm{sgn}} := \begin{bmatrix} \xi_i^{\mathrm{sgn}}|a_{i,1}| \\ \boldsymbol{a}_{i,\perp} \end{bmatrix}$.

See Algorithm 1 for details and also Figure 6(a) for an illustration. One of the most important features of $\{\boldsymbol{x}^{t,(l)}\}$ is that all of its iterates are statistically independent of $(\boldsymbol{a}_l, y_l)$, and hence are incoherent with $\boldsymbol{a}_l$ with high probability, in the sense that $\left|\boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)}\right| \lesssim \sqrt{\log m}\|\boldsymbol{x}^{t,(l)}\|_2$. Such incoherence properties further allow us to control both $\left|\boldsymbol{a}_l^\top \boldsymbol{x}^t\right|$ and $\left|\boldsymbol{a}_l^\top \boldsymbol{x}^{t,\mathrm{sgn}}\right|$ (see (28)), which is crucial for controlling the size of the residual terms (e.g. $r_1(\boldsymbol{x}^t)$ as defined in (11)). Notably, the sequence $\{\boldsymbol{x}^{t,(l)}\}$ has also been applied by [MWCC17] to justify the success of GD with spectral initialization for several nonconvex statistical estimation problems.

- *Random-sign sequence* $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}_{t\geq 0}$. Introduce a collection of auxiliary design vectors $\{\boldsymbol{a}_i^{\mathrm{sgn}}\}_{1\leq i\leq m}$ defined as

$$\boldsymbol{a}_i^{\mathrm{sgn}} := \begin{bmatrix} \xi_i^{\mathrm{sgn}}\,|a_{i,1}| \\ \boldsymbol{a}_{i,\perp} \end{bmatrix}, \tag{33}$$

where $\{\xi_i^{\mathrm{sgn}}\}_{1\leq i\leq m}$ is a set of Rademacher random variables independent of $\{\boldsymbol{a}_i\}$, i.e.

$$\xi_i^{\mathrm{sgn}} \overset{\text{i.i.d.}}{=} \begin{cases} 1, & \text{with probability } 1/2, \\ -1, & \text{else,} \end{cases} \qquad 1\leq i\leq m. \tag{34}$$

In words, $\boldsymbol{a}_i^{\mathrm{sgn}}$ is generated by randomly flipping the sign of the first entry of $\boldsymbol{a}_i$. To simplify the notations hereafter, we also denote

$$\xi_i = \mathrm{sgn}(a_{i,1}). \tag{35}$$

As a result, $\boldsymbol{a}_i$ and $\boldsymbol{a}_i^{\mathrm{sgn}}$ differ only by a single bit of information. With these auxiliary design vectors in place, we generate a sequence $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ by running GD w.r.t. the auxiliary loss function

$$f^{\mathrm{sgn}}(\boldsymbol{x}) = \frac{1}{4m}\sum_{i=1}^m \left[\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^\natural\right)^2\right]^2. \tag{36}$$

One simple yet important feature associated with these new design vectors is that it produces the same measurements as $\{\boldsymbol{a}_i\}$:

$$\left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2 = \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^\natural\right)^2 = |a_{i,1}|^2, \qquad 1\leq i\leq m. \tag{37}$$

See Figure 6(b) for an illustration and Algorithm 2 for the detailed procedure. This sequence is introduced in order to "randomize" certain Gaussian polynomials (e.g. I_4 in (27)), which in turn enables optimal control of these quantities. This is particularly crucial at the initial stage of the algorithm.

- *Leave-one-out and random-sign sequences* $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}_{t\geq 0}$. Furthermore, we also need to introduce another collection of sequences $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}$ by simultaneously employing the new design vectors $\{\boldsymbol{a}_i^{\mathrm{sgn}}\}$ and discarding a single sample $(\boldsymbol{a}_l^{\mathrm{sgn}}, y_l^{\mathrm{sgn}})$. This enables us to propagate the kinds of independence properties across the above two sets of sequences, which is useful in demonstrating that $\boldsymbol{x}^t$ is jointly "nearly-independent" of both $\boldsymbol{a}_l$ and $\{\mathrm{sgn}(a_{i,1})\}$. See Algorithm 3 and Figure 6(c).

As a remark, all of these leave-one-out and random-sign procedures are assumed to start from the same initial point as the original sequence, namely,

$$\boldsymbol{x}^0 = \boldsymbol{x}^{0,(l)} = \boldsymbol{x}^{0,\mathrm{sgn}} = \boldsymbol{x}^{0,\mathrm{sgn},(l)}, \qquad 1 \leq l \leq m. \tag{38}$$

4.5 Justification of approximate state evolution for Phase I of Stage 1

Recall that Phase I consists of the iterations $0 \leq t \leq T_0$, where

$$T_0 = \min\left\{ t : \alpha_{t+1} \geq \frac{c_6}{\log^5 m} \right\}. \tag{39}$$

Our goal here is to show that the approximate state evolution (20) for both the size α_t of the signal component and the size β_t of the orthogonal component holds true throughout Phase I. Our proof will be inductive in nature. Specifically, we will first identify a set of induction hypotheses that are helpful in proving the validity of the approximate state evolution (20), and then proceed by establishing these hypotheses via induction.

4.5.1 Induction hypotheses

For the sake of clarity, we first list all the induction hypotheses.

$$\max_{1\leq l\leq m} \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \leq \beta_t \left(1+\frac{1}{\log m}\right)^t C_1 \frac{\sqrt{n\log^5 m}}{m}, \tag{40a}$$

$$\max_{1\leq l\leq m} \left|x_{\|}^t - x_{\|}^{t,(l)}\right| \leq \alpha_t \left(1+\frac{1}{\log m}\right)^t C_2 \frac{\sqrt{n\log^{12} m}}{m}, \tag{40b}$$

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 \leq \alpha_t \left(1+\frac{1}{\log m}\right)^t C_3 \sqrt{\frac{n\log^5 m}{m}}, \tag{40c}$$

$$\max_{1\leq l\leq m} \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2 \leq \alpha_t \left(1+\frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n\log^9 m}}{m}, \tag{40d}$$

$$c_5 \leq \left\|\boldsymbol{x}_{\perp}^t\right\|_2 \leq \left\|\boldsymbol{x}^t\right\|_2 \leq C_5, \tag{40e}$$

$$\left\|\boldsymbol{x}^t\right\|_2 \leq 4\alpha_t \sqrt{n\log m}, \tag{40f}$$

where $C_1, \cdots, C_5$ and c_5 are some absolute positive constants.

Now we are ready to prove an immediate consequence of the induction hypotheses (40): if (40) hold for the t^{th} iteration, then α_{t+1} and β_{t+1} follow the approximate state evolution (see (20)). This is justified in the following lemma.

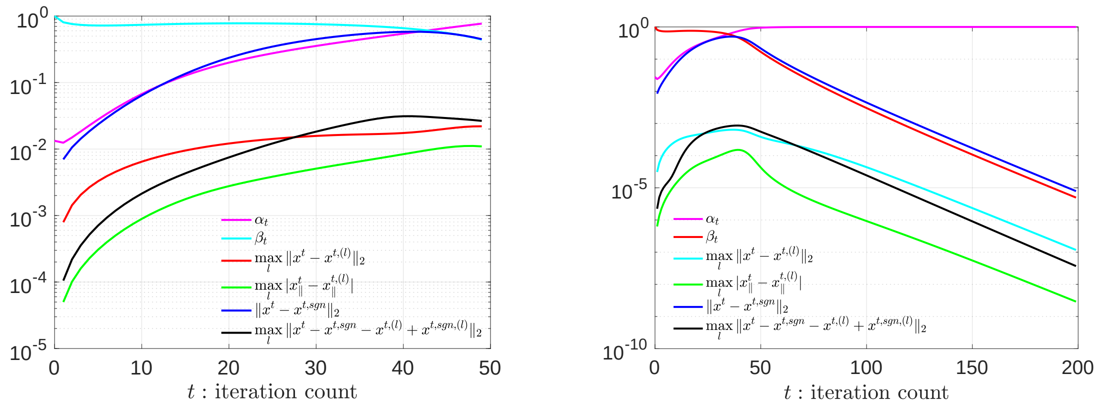
Lemma 2. *Suppose $m \geq Cn\log^{11} m$ for some sufficiently large constant $C > 0$. For any $0 \leq t \leq T_0$ (cf. (39)), if the t^{th} iterates satisfy the induction hypotheses (40), then with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$,*

$$\alpha_{t+1} = \left\{1 + 3\eta\left[1 - \left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\zeta_t\right\}\alpha_t; \tag{41a}$$

$$\beta_{t+1} = \left\{1 + \eta\left[1 - 3\left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\rho_t\right\}\beta_t \tag{41b}$$

hold for some $|\zeta_t| \ll 1/\log m$ and $|\rho_t| \ll 1/\log m$.

Proof. See Appendix C. □



(a) Stage 1 (b) Stage 1 and Stage 2

Figure 7: Illustration of the differences among leave-one-out and original sequences vs. iteration count, plotted semilogarithmically. The results are shown for $n = 1000$ with $m = 10n$, $\eta_t \equiv 0.1$, and $\|\boldsymbol{x}^\natural\|_2 = 1$. (a) The four differences increases in Stage 1. From the induction hypotheses (40), our upper bounds on $|x_\|^t - x_\|^{t,(l)}|$, $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\|_2$ and $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\|_2$ scale linearly with α_t, whereas the upper bound on $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\|_2$ is proportional to β_t. In addition, $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,(l)}\|_2 \lesssim 1/\sqrt{m}$, $|x_\|^1 - x_\|^{1,(l)}| \lesssim 1/m$, $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,\mathrm{sgn}}\|_2 \lesssim 1/\sqrt{m}$ and $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,\mathrm{sgn}} - \boldsymbol{x}^{1,(l)} + \boldsymbol{x}^{1,\mathrm{sgn},(l)}\|_2 \lesssim 1/m$. (b) The four differences converge to zero geometrically fast in Stage 2, as all the (variants of) leave-one-out sequences and the original sequence converge to the truth $\boldsymbol{x}^\natural$.

It remains to inductively show that the hypotheses hold for all $0 \leq t \leq T_0$. Before proceeding to this induction step, it is helpful to first develop more understanding about the preceding hypotheses.

1. In words, (40a), (40b), (40c) specify that the leave-one-out sequences $\{\boldsymbol{x}^{t,(l)}\}$ and $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ are exceedingly close to the original sequence $\{\boldsymbol{x}^t\}$. Similarly, the difference between $\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}$ and $\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}$ is extremely small, as asserted in (40d). The hypothesis (40e) says that the norm of the iterates $\{\boldsymbol{x}^t\}$ is always bounded from above and from below in Phase I. The last one (40f) indicates that the size α_t of the signal component is never too small compared with $\|\boldsymbol{x}^t\|_2$.

2. Another property that is worth mentioning is the growth rate (with respect to t) of the quantities appeared in the induction hypotheses (40). For instance, $\left|x_\|^t - x_\|^{t,(l)}\right|$, $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\|_2$ and $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\|_2$ grow more or less at the same rate as α_t (modulo some $(1 + 1/\log m)^{T_0}$ factor). In contrast, $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\|_2$ shares the same growth rate with β_t (modulo the $(1 + 1/\log m)^{T_0}$ factor). See Figure 7 for an illustration. The difference in the growth rates turns out to be crucial in establishing the advertised result.

3. Last but not least, we emphasize the sizes of the quantities of interest in (40) for $t = 1$ under the Gaussian initialization. Ignoring all of the $\log m$ terms and recognizing that $\alpha_1 \asymp 1/\sqrt{n}$ and $\beta_1 \asymp 1$, one sees that $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,(l)}\|_2 \lesssim 1/\sqrt{m}$, $|x_\|^1 - x_\|^{1,(l)}| \lesssim 1/m$, $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,\mathrm{sgn}}\|_2 \lesssim 1/\sqrt{m}$ and $\|\boldsymbol{x}^1 - \boldsymbol{x}^{1,\mathrm{sgn}} - \boldsymbol{x}^{1,(l)} + \boldsymbol{x}^{1,\mathrm{sgn},(l)}\|_2 \lesssim 1/m$. See Figure 7 for an illustration of the trends of the above four quantities.

Several consequences of (40) regarding the incoherence between $\{\boldsymbol{x}^t\}$, $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ and $\{\boldsymbol{a}_i\}$, $\{\boldsymbol{a}_i^{\mathrm{sgn}}\}$ are immediate, as summarized in the following lemma.

Lemma 3. *Suppose that $m \geq Cn \log^6 m$ for some sufficiently large constant $C > 0$ and the t^{th} iterates satisfy the induction hypotheses (40) for $t \leq T_0$, then with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$,*

$$\max_{1 \leq l \leq m} \left|\boldsymbol{a}_l^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log m} \left\|\boldsymbol{x}^t\right\|_2;$$

$$\max_{1 \leq l \leq m} \left|\boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^t\right| \lesssim \sqrt{\log m} \left\|\boldsymbol{x}_\perp^t\right\|_2;$$

$$\max_{1 \leq l \leq m} \left|\boldsymbol{a}_l^\top \boldsymbol{x}^{t,\mathrm{sgn}}\right| \lesssim \sqrt{\log m} \left\|\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2;$$

$$\max_{1\leq l\leq m}\left|\boldsymbol{a}_{l,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right|\lesssim\sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2;$$
$$\max_{1\leq l\leq m}\left|\boldsymbol{a}_{l}^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right|\lesssim\sqrt{\log m}\left\|\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2.$$

Proof. These incoherence conditions typically arise from the independence between $\{\boldsymbol{x}^{t,(l)}\}$ and $\boldsymbol{a}_l$. For instance, the first line follows since

$$\left|\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t}\right|\approx\left|\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t,(l)}\right|\lesssim\sqrt{\log m}\|\boldsymbol{x}^{t,(l)}\|_2\asymp\sqrt{\log m}\|\boldsymbol{x}^{t}\|_2.$$

See Appendix M for detailed proofs. □

4.5.2 Induction step

We then turn to showing that the induction hypotheses (40) hold throughout Phase I, i.e. for $0\leq t\leq T_0$. The base case can be easily verified because of the identical initial points (38). Now we move on to the inductive step, i.e. we aim to show that if the hypotheses (40) are valid up to the t^{th} iteration for some $t\leq T_0$, then they continue to hold for the $(t+1)^{\mathrm{th}}$ iteration.

The first lemma concerns the difference between the leave-one-out sequence $\boldsymbol{x}^{t+1,(l)}$ and the true sequence $\boldsymbol{x}^{t+1}$ (see (40a)).

Lemma 4. *Suppose $m\geq Cn\log^5 m$ for some sufficiently large constant $C>0$. If the induction hypotheses (40) hold true up to the t^{th} iteration for some $t\leq T_0$, then with probability at least $1-O(me^{-1.5n})-O(m^{-10})$,*

$$\max_{1\leq l\leq m}\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\right\|_2\leq\beta_{t+1}\left(1+\frac{1}{\log m}\right)^{t+1}C_1\frac{\sqrt{n\log^5 m}}{m}\tag{43}$$

holds as long as $\eta>0$ is a sufficiently small constant and $C_1>0$ is sufficiently large.

Proof. See Appendix D. □

The next lemma characterizes a finer relation between $\boldsymbol{x}^{t+1}$ and $\boldsymbol{x}^{t+1,(l)}$ when projected onto the signal direction (cf. (40b)).

Lemma 5. *Suppose $m\geq Cn\log^6 m$ for some sufficiently large constant $C>0$. If the induction hypotheses (40) hold true up to the t^{th} iteration for some $t\leq T_0$, then with probability at least $1-O(me^{-1.5n})-O(m^{-10})$,*

$$\max_{1\leq l\leq m}\left|x_{\|}^{t+1}-x_{\|}^{t+1,(l)}\right|\leq\alpha_{t+1}\left(1+\frac{1}{\log m}\right)^{t+1}C_2\frac{\sqrt{n\log^{12} m}}{m}\tag{44}$$

holds as long as $\eta>0$ is a sufficiently small constant and $C_2\gg C_4$.

Proof. See Appendix E. □

Regarding the difference between $\boldsymbol{x}^{t}$ and $\boldsymbol{x}^{t,\mathrm{sgn}}$ (see (40c)), we have the following result.

Lemma 6. *Suppose $m\geq Cn\log^5 m$ for some sufficiently large constant $C>0$. If the induction hypotheses (40) hold true up to the t^{th} iteration for some $t\leq T_0$, then with probability at least $1-O(me^{-1.5n})-O\left(m^{-10}\right)$,*

$$\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,\mathrm{sgn}}\right\|_2\leq\alpha_{t+1}\left(1+\frac{1}{\log m}\right)^{t+1}C_3\sqrt{\frac{n\log^5 m}{m}}\tag{45}$$

holds as long as $\eta>0$ is a sufficiently small constant and C_3 is a sufficiently large positive constant.

Proof. See Appendix F. □

We are left with the double difference $\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,\mathrm{sgn}}-\boldsymbol{x}^{t+1,(l)}+\boldsymbol{x}^{t+1,\mathrm{sgn},(l)}$ (cf. (40d)), for which one has the following lemma.

Lemma 7. *Suppose $m \geq Cn\log^8 m$ for some sufficiently large constant $C > 0$. If the induction hypotheses (40) hold true up to the t^{th} iteration for some $t \leq T_0$, then with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$,*

$$\max_{1\leq l\leq m} \left\| \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,\mathrm{sgn}} - \boldsymbol{x}^{t+1,(l)} + \boldsymbol{x}^{t+1,\mathrm{sgn},(l)} \right\|_2 \leq \alpha_{t+1} \left(1 + \frac{1}{\log m}\right)^{t+1} C_4 \frac{\sqrt{n\log^9 m}}{m} \tag{46}$$

holds as long as $\eta > 0$ is a sufficiently small constant and $C_4 > 0$ is sufficiently large.

Proof. See Appendix G. □

Assuming the induction hypotheses (40) hold up to the t^{th} iteration for some $t \leq T_0$, we know from Lemma 2 that the approximate state evolution for both α_t and β_t (see (20)) holds up to $t+1$. As a result, the last two hypotheses (40e) and (40f) for the $(t+1)^{\text{th}}$ iteration can be easily verified.

4.6 Justification of approximate state evolution for Phase II of Stage 1

Recall from Lemma 1 that Phase II refers to the iterations $T_0 < t \leq T_\gamma$ (see the definition of T_0 in Lemma 1), for which one has

$$\alpha_t \geq \frac{c_6}{\log^5 m} \tag{47}$$

as long as the approximate state evolution (20) holds. Here $c_6 > 0$ is the same constant as in Lemma 1. Similar to Phase I, we invoke an inductive argument to prove that the approximate state evolution (20) continues to hold for $T_0 < t \leq T_\gamma$.

4.6.1 Induction hypotheses

In Phase I, we rely on the leave-one-out sequences and the random-sign sequences $\{\boldsymbol{x}^{t,(l)}\}, \{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ and $\{\boldsymbol{x}^{t,\mathrm{sgn},(l)}\}$ to establish certain "near-independence" between $\{\boldsymbol{x}^t\}$ and $\{\boldsymbol{a}_l\}$, which in turn allows us to obtain sharp control of the residual terms $\boldsymbol{r}(\boldsymbol{x}^t)$ (cf. (10)) and $r_1(\boldsymbol{x}^t)$ (cf. (11)). As it turns out, once the size α_t of the signal component obeys $\alpha_t \gtrsim 1/\mathrm{poly}\log(m)$, then $\{\boldsymbol{x}^{t,(l)}\}$ alone is sufficient for our purpose to establish the "near-independence" property. More precisely, in Phase II we only need to impose the following induction hypotheses.

$$\max_{1\leq l\leq m} \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} \right\|_2 \leq \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_6 \frac{\sqrt{n\log^{15} m}}{m}; \tag{48a}$$

$$c_5 \leq \left\| \boldsymbol{x}_\perp^t \right\|_2 \leq \left\| \boldsymbol{x}^t \right\|_2 \leq C_5. \tag{48b}$$

A direct consequence of (48) is the incoherence between $\boldsymbol{x}^t$ and $\{\boldsymbol{a}_l\}$, namely,

$$\max_{1\leq l\leq m} \left| \boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^t \right| \lesssim \sqrt{\log m} \left\| \boldsymbol{x}_\perp^t \right\|_2; \tag{49a}$$

$$\max_{1\leq l\leq m} \left| \boldsymbol{a}_l^\top \boldsymbol{x}^t \right| \lesssim \sqrt{\log m} \left\| \boldsymbol{x}^t \right\|_2. \tag{49b}$$

To see this, one can use the triangle inequality to show that

$$\begin{aligned}
\left| \boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^t \right| &\leq \left| \boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^{t,(l)} \right| + \left| \boldsymbol{a}_{l,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} \right) \right| \\
&\overset{\text{(i)}}{\lesssim} \sqrt{\log m} \left\| \boldsymbol{x}_\perp^{t,(l)} \right\|_2 + \sqrt{n} \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} \right\|_2 \\
&\lesssim \sqrt{\log m} \left(\left\| \boldsymbol{x}_\perp^t \right\|_2 + \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} \right\|_2 \right) + \sqrt{n} \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} \right\|_2 \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{\log m} + \frac{\sqrt{n\log^{15} m}}{m} \sqrt{n} \lesssim \sqrt{\log m},
\end{aligned}$$

where (i) follows from the independence between $\boldsymbol{a}_l$ and $\boldsymbol{x}^{t,(l)}$ and the Cauchy-Schwarz inequality, and the last line (ii) arises from $(1+1/\log m)^t \lesssim 1$ for $t \leq T_\gamma \lesssim \log n$ and $m \gg n \log^{15/2} m$. This combined with the fact that $\|\boldsymbol{x}_\perp^t\|_2 \geq c_5/2$ results in

$$\max_{1\leq l\leq m} \left|\boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^t\right| \lesssim \sqrt{\log m}\, \left\|\boldsymbol{x}_\perp^t\right\|_2. \tag{50}$$

The condition (49b) follows using nearly identical arguments, which are omitted here.

As in Phase I, we need to justify the approximate state evolution (20) for both α_t and β_t, given that the t^{th} iterates satisfy the induction hypotheses (48). This is stated in the following lemma.

Lemma 8. *Suppose $m \geq Cn\log^{13} m$ for some sufficiently large constant $C > 0$. If the t^{th} iterates satisfy the induction hypotheses (48) for $T_0 < t < T_\gamma$, then with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$,*

$$\alpha_{t+1} = \left\{1 + 3\eta\left[1 - \left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\zeta_t\right\}\alpha_t; \tag{51a}$$

$$\beta_{t+1} = \left\{1 + \eta\left[1 - 3\left(\alpha_t^2 + \beta_t^2\right)\right] + \eta\rho_t\right\}\beta_t, \tag{51b}$$

for some $|\zeta_t| \ll 1/\log m$ and $\rho_t \ll 1/\log m$.

Proof. See Appendix H for the proof of (51a). The proof of (51b) follows exactly the same argument as in proving (41b), and is hence omitted. □

4.6.2 Induction step

We proceed to complete the induction argument. Towards this end, one has the following lemma in regard to the induction on $\max_{1\leq l\leq m} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)}\|_2$ (see (48a)).

Lemma 9. *Suppose $m \geq Cn\log^5 m$ for some sufficiently large constant $C > 0$, and consider any $T_0 < t < T_\gamma$. If the induction hypotheses (40) are valid throughout Phase I and (48) are valid from the T_0th to the t^{th} iterations, then with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$,*

$$\max_{1\leq l\leq m} \left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)}\right\|_2 \leq \alpha_{t+1}\left(1 + \frac{1}{\log m}\right)^{t+1} C_6 \frac{\sqrt{n\log^{13} m}}{m}$$

holds as long as $\eta > 0$ is sufficiently small and $C_6 > 0$ is sufficiently large.

Proof. See Appendix I. □

As in Phase I, since we assume the induction hypotheses (40) (resp. (48)) hold for all iterations up to the T_0th iteration (resp. between the T_0th and the t^{th} iteration), we know from Lemma 8 that the approximate state evolution for both α_t and β_t (see (20)) holds up to $t+1$. The last induction hypothesis (48b) for the $(t+1)^{\text{th}}$ iteration can be easily verified from Lemma 1.

It remains to check the case when $t = T_0 + 1$. It can be seen from the analysis in Phase I that

$$\begin{aligned}\max_{1\leq l\leq m} \left\|\boldsymbol{x}^{T_0+1} - \boldsymbol{x}^{T_0+1,(l)}\right\|_2 &\leq \beta_{T_0+1}\left(1 + \frac{1}{\log m}\right)^{T_0+1} C_1 \frac{\sqrt{n\log^5 m}}{m} \\ &\leq \alpha_{T_0+1}\left(1 + \frac{1}{\log m}\right)^{T_0+1} C_6 \frac{\sqrt{n\log^{15} m}}{m},\end{aligned}$$

for some constant condition $C_6 \gg 1$, where the second line holds since $\beta_{T_0+1} \leq C_5$, $\alpha_{T_0+1} \geq c_6/\log^5 m$.

4.7 Analysis for Stage 2

Combining the analyses in *Phase I* and *Phase II*, we finish the proof of Theorem 2 for Stage 1, i.e. $t \leq T_\gamma$. In addition to $\mathrm{dist}\left(\boldsymbol{x}^{T_\gamma}, \boldsymbol{x}^\natural\right) \leq \gamma$, we can also see from (49b) that

$$\max_{1\leq i\leq m} \left|\boldsymbol{a}_i^\top \boldsymbol{x}^{T_\gamma}\right| \lesssim \sqrt{\log m},$$

which in turn implies that

$$\max_{1\le i\le m}\left|\boldsymbol{a}_i^\top\left(\boldsymbol{x}^{T_\gamma}-\boldsymbol{x}^\natural\right)\right|\lesssim\sqrt{\log m}.$$

Armed with these properties, one can apply the arguments in [MWCC17, Section 6] to prove that for $t\ge T_\gamma+1$,

$$\operatorname{dist}\left(\boldsymbol{x}^t,\boldsymbol{x}^\natural\right)\le\left(1-\frac{\eta}{2}\right)^{t-T_\gamma}\operatorname{dist}\left(\boldsymbol{x}^{T_\gamma},\boldsymbol{x}^\natural\right)\le\left(1-\frac{\eta}{2}\right)^{t-T_\gamma}\cdot\gamma. \tag{52}$$

Notably, the theorem therein [MWCC17, Theorem 1] works under the stepsize $\eta_t\equiv\eta\asymp c/\log n$ when $m\gg n\log n$. Nevertheless, as remarked by the authors, when the sample complexity exceeds $m\gg n\log^3 m$, a constant stepsize is allowed.

We are left with proving (15) for Stage 2. Note that we have already shown that the ratio α_t/β_t increases exponentially fast in Stage 1. Therefore,

$$\frac{\alpha_{T_1}}{\beta_{T_1}}\ge\frac{1}{\sqrt{2n\log n}}(1+c_{10}\eta^2)^{T_1}$$

and, by the definition of T_1 (see (26)) and Lemma 1, one has $\alpha_{T_1}\asymp\beta_{T_1}\asymp 1$ and hence

$$\frac{\alpha_{T_1}}{\beta_{T_1}}\asymp 1. \tag{53}$$

When it comes to $t>T_\gamma$, in view of (52), one has

$$\begin{aligned}
\frac{\alpha_t}{\beta_t}&\ge\frac{1-\operatorname{dist}\left(\boldsymbol{x}^t,\boldsymbol{x}^\natural\right)}{\operatorname{dist}\left(\boldsymbol{x}^t,\boldsymbol{x}^\natural\right)}\ge\frac{1-\gamma}{\left(1-\frac{\eta}{2}\right)^{t-T_\gamma}\cdot\gamma}\\
&\ge\frac{1-\gamma}{\gamma}\left(1+\frac{\eta}{2}\right)^{t-T_\gamma}\overset{\text{(i)}}{\asymp}\frac{\alpha_{T_1}}{\beta_{T_1}}\left(1+\frac{\eta}{2}\right)^{t-T_\gamma}\\
&\gtrsim\frac{1}{\sqrt{n\log n}}\left(1+c_{10}\eta^2\right)^{T_1}\left(1+\frac{\eta}{2}\right)^{t-T_\gamma}\\
&\overset{\text{(ii)}}{\asymp}\frac{1}{\sqrt{n\log n}}\left(1+c_{10}\eta^2\right)^{T_\gamma}\left(1+\frac{\eta}{2}\right)^{t-T_\gamma}\\
&\gtrsim\frac{1}{\sqrt{n\log n}}\left(1+c_{10}\eta^2\right)^{t},
\end{aligned}$$

where (i) arises from (53) and the fact that γ is a constant, (ii) follows since $T_\gamma-T_1\asymp 1$ according to Lemma 1, and the last line holds as long as $c_{10}>0$ and η are sufficiently small. This concludes the proof regarding the lower bound on α_t/β_t.

5 Discussions

The current paper justifies the fast global convergence of gradient descent with random initialization for phase retrieval. Specifically, we demonstrate that GD with random initialization takes only $O\big(\log n+\log(1/\epsilon)\big)$ iterations to achieve a relative ϵ-accuracy in terms of the estimation error. It is likely that such fast global convergence properties also arise in other nonconvex statistical estimation problems. The technical tools developed herein may also prove useful for other settings. We conclude our paper with a few directions worthy of future investigation.

- *Sample complexity and phase transition.* We have proved in Theorem 2 that GD with random initialization enjoys fast convergence, with the proviso that $m\gg n\log^{13} m$. It is possible to improve the sample complexity via more sophisticated arguments. In addition, it would be interesting to examine the phase transition phenomenon of GD with random initialization.
- *Other nonconvex statistical estimation problems.* We use the phase retrieval problem to showcase the efficiency of GD with random initialization. It is certainly interesting to investigate whether this fast global

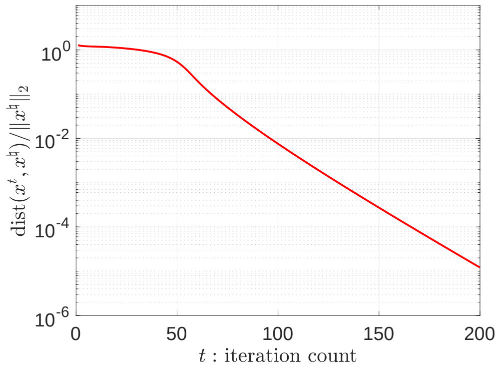


Figure 8: The relative ℓ_2 error vs. iteration count for GD with random initialization, plotted semilogarithmically. The results are shown for $n = 1000$ with $m = 10n$ and $\eta_t \equiv 0.1$. Here the entries of the sampling vectors $\boldsymbol{a}_i$ are drawn *i.i.d.* from a Rademacher distribution.

convergence carries over to other nonconvex statistical estimation problems including *low-rank matrix and tensor recovery* [KMO10, SL16, CW15, TBS+16, ZL16, ZWL15, MWCC17, CL17, CC18, CCF18, HZC18], *blind deconvolution* [LLSW18, MWCC17, HH17] and *neural networks* [SJL17, LMZ17, FCL18]. The leave-one-out sequences and the "near-independence" property introduced / identified in this paper might be useful in proving efficiency of randomly initialized GD for the aforementioned problems.

- *Other iterative optimization methods.* Apart from gradient descent, other iterative procedures have been applied to solve the phase retrieval problem. Partial examples include *alternating minimization*, *Kaczmarz algorithm,* and *truncated gradient descent (Truncated Wirtinger flow)*. In conjunction with random initialization, whether the iterative algorithms mentioned above enjoy fast global convergence is an interesting open problem. For example, it has been shown that truncated WF together with truncated spectral initialization achieves optimal sample complexity (i.e. $m \asymp n$) and computational complexity simultaneously [CC17]. Does truncated Wirtinger flow still enjoy optimal sample complexity when initialized randomly?

- *Beyond Gaussian sampling vectors.* In this work, we consider the Gaussian phase retrieval problem where the sampling vectors are i.i.d. Gaussian vectors. We expect our results to generalize to other sampling vectors. Experimentally, we can verify that random initialization also converges fast under a Rademacher sampling model; see Figure 8.

- *Applications of leave-one-out tricks.* In this paper, we heavily deploy the *leave-one-out* trick to demonstrate the "near-independence" between the iterates $\boldsymbol{x}^t$ and the sampling vectors $\{\boldsymbol{a}_i\}$. The basic idea is to construct an auxiliary sequence that is (i) independent w.r.t. certain components of the design vectors, and (ii) extremely close to the original sequence. These two properties allow us to propagate the desired independence properties to $\boldsymbol{x}^t$. As mentioned in Section 3, the leave-one-out trick has served as a very powerful hammer for decoupling the dependency between random vectors in several high-dimensional estimation problems. We expect this powerful trick to be useful in broader settings.

Acknowledgements

Y. Chen is supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ARO grant W911NF-18-1-0303, by the ONR grant N00014-19-1-2120, and by the Princeton SEAS innovation award. Y. Chi is supported in part by AFOSR under the grant FA9550-15-1-0205, by ONR under the grant N00014-18-1-2142, by ARO under the grant W911NF-18-1-0303, and by NSF under the grants CAREER ECCS-1818571 and CCF-1806154. J. Fan is supported in part by the NSF grants DMS-1662139 and DMS-1712591, the ONR grant N00014-19-1-2120, and the NIH grant 2R01-GM072611-13.

References

[AAZB$^+$16] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

[AFWZ17] E. Abbe, J. Fan, K. Wang, and Y. Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

[AZ17] Z. Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. *arXiv preprint arXiv:1708.08694*, 2017.

[BCMN14] A. S. Bandeira, J. Cahill, D. G. Mixon, and A. A. Nelson. Saving phase: Injectivity and stability for phase retrieval. *Applied and Computational Harmonic Analysis*, 37(1):106–125, 2014.

[BEB17] T. Bendory, Y. C. Eldar, and N. Boumal. Non-convex phase retrieval from STFT measurements. *IEEE Transactions on Information Theory*, 2017.

[CC17] Y. Chen and E. J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm. Pure Appl. Math.*, 70(5):822–883, 2017.

[CC18] Y. Chen and E. Candès. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Communications on Pure and Applied Mathematics*, 71(8):1648–1714, 2018.

[CCF18] Y. Chen, C. Cheng, and J. Fan. Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *arXiv preprint arXiv:1811.12804*, 2018.

[CCG15] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[CESV13] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[CFL15] P. Chen, A. Fannjiang, and G.-R. Liu. Phase retrieval with one or two diffraction patterns by alternating projections with the null initialization. *Journal of Fourier Analysis and Applications*, pages 1–40, 2015.

[CFMW17] Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized MLE are both optimal for top-K ranking. *arXiv preprint arXiv:1707.09971*, 2017.

[CL14] E. J. Candès and X. Li. Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

[CL16] Y. Chi and Y. M. Lu. Kaczmarz method for solving quadratic equations. *IEEE Signal Processing Letters*, 23(9):1183–1187, 2016.

[CL17] J. Chen and X. Li. Memory-efficient kernel PCA via partial matrix sampling and nonconvex optimization: a model-free analysis of local minima. *arXiv preprint arXiv:1711.01742*, 2017.

[CLC18] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.

[CLM16] T. T. Cai, X. Li, and Z. Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

[CLS15] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, April 2015.

[CLW17] J.-F. Cai, H. Liu, and Y. Wang. Fast rank one alternating minimization algorithm for phase retrieval. *arXiv preprint arXiv:1708.08751*, 2017.

[CSV13] E. J. Candès, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1017–1026, 2013.

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

[CWZG17] J. Chen, L. Wang, X. Zhang, and Q. Gu. Robust wirtinger flow for phase retrieval with arbitrary corruption. *arXiv preprint arXiv:1704.06256*, 2017.

[CYC14] Y. Chen, X. Yi, and C. Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conference on Learning Theory*, pages 560–604, 2014.

[CZ15] T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

[DH14] L. Demanet and P. Hand. Stable optimizationless recovery from phaseless linear measurements. *Journal of Fourier Analysis and Applications*, 20(1):199–221, 2014.

[DJL+17] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017.

[DR17] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *arXiv preprint arXiv:1705.02356*, 2017.

[EK15] N. El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, pages 1–81, 2015.

[EKBB+13] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[FCL18] H. Fu, Y. Chi, and Y. Liang. Local geometry of one-hidden-layer neural networks for logistic regression. *arXiv preprint arXiv:1802.06463*, 2018.

[GHJY15] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[GX16] B. Gao and Z. Xu. Phase retrieval using Gauss-Newton method. *arXiv preprint arXiv:1606.08135*, 2016.

[HH17] W. Huang and P. Hand. Blind deconvolution by a steepest descent algorithm on a quotient manifold. *arXiv preprint arXiv:1710.03309*, 2017.

[HZC18] B. Hao, A. Zhang, and G. Cheng. Sparse and low-rank tensor estimation via cubic sketchings. *arXiv preprint arXiv:1801.09326*, 2018.

[JGN+17] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[JNJ17] C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017.

[KMO10] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980 –2998, June 2010.

[KRT17] R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.

[Lan93] S. Lang. Real and functional analysis. *Springer-Verlag, New York,*, 10:11–13, 1993.

[LGL15] G. Li, Y. Gu, and Y. M. Lu. Phase retrieval using iterative projections: Dynamics in the large systems limit. In *Allerton Conference on Communication, Control, and Computing*, pages 1114–1118. IEEE, 2015.

[LL17] Y. M. Lu and G. Li. Phase transitions of spectral initialization for high-dimensional nonconvex estimation. *arXiv preprint arXiv:1702.06435*, 2017.

[LLSW18] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and Computational Harmonic Analysis*, 2018.

[LMCC18] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex matrix factorization from rank-one measurements. *arXiv preprint arXiv:1802.06286*, 2018.

[LMZ17] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix recovery. *arXiv preprint arXiv:1712.09203*, 2017.

[LSJR16] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

[MM17] M. Mondelli and A. Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *arXiv preprint arXiv:1708.05932*, 2017.

[MSK17] R. Murray, B. Swenson, and S. Kar. Revisiting normalized gradient descent: Evasion of saddle points. *arXiv preprint arXiv:1711.05224*, 2017.

[MWCC17] C. Ma, K. Wang, Y. Chi, and Y. 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.

[MXM18] J. Ma, J. Xu, and A. Maleki. Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2-regularization. *arXiv preprint arXiv:1801.01170*, 2018.

[NJS13] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[NP06] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[QZEW17] Q. Qing, Y. Zhang, Y. Eldar, and J. Wright. Convolutional phase retrieval via gradient descent. *Neural Information Processing Systems*, 2017.

[SCC18] P. Sur, Y. Chen, and E. J. Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *accepted to Probability Theory and Related Fields*, 2018.

[SEC$^+$15] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.

[SJL17] M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *arXiv preprint arXiv:1707.04926*, 2017.

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

[Sol14] M. Soltanolkotabi. *Algorithms and Theory for Clustering and Nonconvex Quadratic Programming*. PhD thesis, Stanford University, 2014.

[Sol17] M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *arXiv preprint arXiv:1702.06175*, 2017.

[SQW16] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2379–2383. IEEE, 2016.

[SS12] W. Schudy and M. Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 437–446. ACM, New York, 2012.

[TBS+16] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 964–973. JMLR. org, 2016.

[TV17] Y. S. Tan and R. Vershynin. Phase retrieval via randomized kaczmarz: Theoretical guarantees. *arXiv preprint arXiv:1706.09993*, 2017.

[Ver12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications*, pages 210 – 268, 2012.

[Wei15] K. Wei. Solving systems of phaseless equations via Kaczmarz methods: A proof of concept study. *Inverse Problems*, 31(12):125008, 2015.

[WGE17] G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 2017.

[WGSC17] G. Wang, G. B. Giannakis, Y. Saad, and J. Chen. Solving almost all systems of random quadratic equations. *arXiv preprint arXiv:1705.10407*, 2017.

[YYF+17] Z. Yang, L. F. Yang, E. X. Fang, T. Zhao, Z. Wang, and M. Neykov. Misspecified nonconvex statistical optimization for phase retrieval. *arXiv preprint arXiv:1712.06245*, 2017.

[ZB17] Y. Zhong and N. Boumal. Near-optimal bounds for phase synchronization. *arXiv preprint arXiv:1703.06605*, 2017.

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

[Zha17] T. Zhang. Phase retrieval using alternating minimization in a batch setting. *arXiv preprint arXiv:1706.08167*, 2017.

[ZL16] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[ZWGC17] L. Zhang, G. Wang, G. B. Giannakis, and J. Chen. Compressive phase retrieval via reweighted amplitude flow. *arXiv preprint arXiv:1712.02426*, 2017.

[ZWL15] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

[ZZLC17] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi. A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 2017.

A Preliminaries

We first gather two standard concentration inequalities used throughout the appendix. The first lemma is the multiplicative form of the Chernoff bound, while the second lemma is a user-friendly version of the Bernstein inequality.

Lemma 10. *Suppose $X_1, \cdots, X_m$ are independent random variables taking values in $\{0,1\}$. Denote $X = \sum_{i=1}^m X_i$ and $\mu = \mathbb{E}[X]$. Then for any $\delta \geq 1$, one has*

$$\mathbb{P}\left(X \geq (1+\delta)\mu\right) \leq e^{-\delta\mu/3}.$$

Lemma 11. *Consider m independent random variables z_l $(1 \leq l \leq m)$, each satisfying $|z_l| \leq B$. For any $a \geq 2$, one has*

$$\left|\sum_{l=1}^m z_l - \sum_{l=1}^m \mathbb{E}[z_l]\right| \leq \sqrt{2a\log m \sum_{l=1}^m \mathbb{E}[z_l^2]} + \frac{2a}{3}B\log m$$

with probability at least $1 - 2m^{-a}$.

Next, we list a few simple facts. The gradient and the Hessian of the nonconvex loss function (2) are given respectively by

$$\nabla f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m \left[\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{x}; \tag{54}$$

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m \left[3\left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_i\boldsymbol{a}_i^\top. \tag{55}$$

In addition, recall that $\boldsymbol{x}^\natural$ is assumed to be $\boldsymbol{x}^\natural = \boldsymbol{e}_1$ throughout the proof. For each $1 \leq i \leq m$, we have the decomposition $\boldsymbol{a}_i = \begin{bmatrix} a_{i,1} \\ \boldsymbol{a}_{i,\perp} \end{bmatrix}$, where $\boldsymbol{a}_{i,\perp}$ contains the 2nd through the nth entries of $\boldsymbol{a}_i$. The standard concentration inequality reveals that

$$\max_{1\leq i\leq m} \left|\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right| = \max_{1\leq i\leq m} |a_{i,1}| \leq 5\sqrt{\log m} \tag{56}$$

with probability $1 - O\left(m^{-10}\right)$. Additionally, apply the standard concentration inequality to see that

$$\max_{1\leq i\leq m} \|\boldsymbol{a}_i\|_2 \leq \sqrt{6n} \tag{57}$$

with probability $1 - O\left(me^{-1.5n}\right)$.

The next lemma provides concentration bounds regarding polynomial functions of $\{\boldsymbol{a}_i\}$.

Lemma 12. *Consider any $\epsilon > 3/n$. Suppose that $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for $1 \leq i \leq m$. Let*

$$\mathcal{S} := \left\{\boldsymbol{z} \in \mathbb{R}^{n-1} \,\middle|\, \max_{1\leq i\leq m} \left|\boldsymbol{a}_{i,\perp}^\top \boldsymbol{z}\right| \leq \beta\|\boldsymbol{z}\|_2\right\},$$

where β is any value obeying $\beta \geq c_1\sqrt{\log m}$ for some sufficiently large constant $c_1 > 0$. Then with probability exceeding $1 - O\left(m^{-10}\right)$, one has

1. $\left|\frac{1}{m}\sum_{i=1}^m a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{z}\right| \leq \epsilon\|\boldsymbol{z}\|_2$ *for all $\boldsymbol{z} \in \mathcal{S}$, provided that $m \geq c_0 \max\left\{\frac{1}{\epsilon^2} n\log n,\ \frac{1}{\epsilon}\beta n \log^{\frac{5}{2}} m\right\}$;*

2. $\left|\frac{1}{m}\sum_{i=1}^m a_{i,1}\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{z}\right)^3\right| \leq \epsilon\|\boldsymbol{z}\|_2^3$ *for all $\boldsymbol{z} \in \mathcal{S}$, provided that $m \geq c_0 \max\left\{\frac{1}{\epsilon^2} n\log n,\ \frac{1}{\epsilon}\beta^3 n \log^{\frac{3}{2}} m\right\}$;*

3. $\left|\frac{1}{m}\sum_{i=1}^m a_{i,1}^2\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{z}\right)^2 - \|\boldsymbol{z}\|_2^2\right| \leq \epsilon\|\boldsymbol{z}\|_2^2$ *for all $\boldsymbol{z} \in \mathcal{S}$, provided that $m \geq c_0 \max\left\{\frac{1}{\epsilon^2} n\log n,\ \frac{1}{\epsilon}\beta^2 n \log^2 m\right\}$;*

4. $\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^{6}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right)^{2}-15\left\|\boldsymbol{z}\right\|_{2}^{2}\right|\leq\epsilon\left\|\boldsymbol{z}\right\|_{2}^{2}$ *for all* $\boldsymbol{z}\in\mathcal{S}$, *provided that* $m\geq c_{0}\max\left\{\frac{1}{\epsilon^{2}}n\log n,\ \frac{1}{\epsilon}\beta^{2}n\log^{4}m\right\}$;

5. $\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^{2}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right)^{6}-15\left\|\boldsymbol{z}\right\|_{2}^{6}\right|\leq\epsilon\left\|\boldsymbol{z}\right\|_{2}^{6}$ *for all* $\boldsymbol{z}\in\mathcal{S}$, *provided that* $m\geq c_{0}\max\left\{\frac{1}{\epsilon^{2}}n\log n,\ \frac{1}{\epsilon}\beta^{6}n\log^{2}m\right\}$;

6. $\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^{2}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right)^{4}-3\left\|\boldsymbol{z}\right\|_{2}^{4}\right|\leq\epsilon\left\|\boldsymbol{z}\right\|_{2}^{4}$ *for all* $\boldsymbol{z}\in\mathcal{S}$, *provided that* $m\geq c_{0}\max\left\{\frac{1}{\epsilon^{2}}n\log n,\ \frac{1}{\epsilon}\beta^{4}n\log^{2}m\right\}$.

Here, $c_0>0$ *is some sufficiently large constant.*

Proof. See Appendix J. □

The next lemmas provide the (uniform) matrix concentration inequalities about $\left\{\boldsymbol{a}_i\boldsymbol{a}_i^{\top}\right\}$.

Lemma 13 ([Ver12, Corollary 5.35])**.** *Suppose that* $\boldsymbol{a}_i\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(\boldsymbol{0},\boldsymbol{I}_n\right)$ *for* $1\leq i\leq m$. *With probability at least* $1-ce^{-\tilde{c}m}$, *one has*

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{a}_i\boldsymbol{a}_i^{\top}\right\|\leq 2,$$

as long as $m\geq c_0 n$ *for some sufficiently large constant* $c_0>0$. *Here,* $c,\tilde{c}>0$ *are some absolute constants.*

Lemma 14. *Fix some* $\boldsymbol{x}^{\natural}\in\mathbb{R}^n$. *Suppose that* $\boldsymbol{a}_i\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(\boldsymbol{0},\boldsymbol{I}_n\right)$, $1\leq i\leq m$. *With probability at least* $1-O\left(m^{-10}\right)$, *one has*

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{\natural}\right)^{2}\boldsymbol{a}_i\boldsymbol{a}_i^{\top}-\left\|\boldsymbol{x}^{\natural}\right\|_2^2\boldsymbol{I}_n-2\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top}\right\|\leq c_0\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}^{\natural}\right\|_2^2,\tag{58}$$

provided that $m>c_1 n\log^3 m$. *Here,* c_0,c_1 *are some universal positive constants. Furthermore, fix any* $c_2>1$ *and suppose that* $m>c_1 n\log^3 m$ *for some sufficiently large constant* $c_1>0$. *Then with probability exceeding* $1-O\left(m^{-10}\right)$,

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\top}\boldsymbol{z}\right)^{2}\boldsymbol{a}_i\boldsymbol{a}_i^{\top}-\left\|\boldsymbol{z}\right\|_2^2\boldsymbol{I}_n-2\boldsymbol{z}\boldsymbol{z}^{\top}\right\|\leq c_0\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{z}\right\|_2^2\tag{59}$$

holds simultaneously for all $\boldsymbol{z}\in\mathbb{R}^n$ *obeying* $\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\boldsymbol{z}\right|\leq c_2\sqrt{\log m}\left\|\boldsymbol{z}\right\|_2$. *On this event, we have*

$$\left\|\frac{1}{m}\sum_{i=1}^{m}|a_{i,1}|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}\right\|\leq\left\|\frac{1}{m}\sum_{i=1}^{m}|a_{i,1}|^2\boldsymbol{a}_i\boldsymbol{a}_i^{\top}\right\|\leq 4.\tag{60}$$

Proof. See Appendix K. □

The following lemma provides the concentration results regarding the Hessian matrix $\nabla^2 f\left(\boldsymbol{x}\right)$.

Lemma 15. *Fix any constant* $c_0>1$. *Suppose that* $m>c_1 n\log^3 m$ *for some sufficiently large constant* $c_1>0$. *Then with probability exceeding* $1-O\left(m^{-10}\right)$,

$$\left\|\left(\boldsymbol{I}_n-\eta\nabla^2 f\left(\boldsymbol{z}\right)\right)-\left\{\left(1-3\eta\left\|\boldsymbol{z}\right\|_2^2+\eta\right)\boldsymbol{I}_n+2\eta\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top}-6\eta\boldsymbol{z}\boldsymbol{z}^{\top}\right\}\right\|\lesssim\sqrt{\frac{n\log^3 m}{m}}\max\left\{\left\|\boldsymbol{z}\right\|_2^2,1\right\}$$

$$\text{and}\qquad\left\|\nabla^2 f\left(\boldsymbol{z}\right)\right\|\leq 10\|\boldsymbol{z}\|_2^2+4$$

hold simultaneously for all $\boldsymbol{z}$ *obeying* $\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\boldsymbol{z}\right|\leq c_0\sqrt{\log m}\left\|\boldsymbol{z}\right\|_2$, *provided that* $0<\eta<\frac{c_2}{\max\{\|\boldsymbol{z}\|_2^2,1\}}$ *for some sufficiently small constant* $c_2>0$.

Proof. See Appendix L. □

Finally, we note that there are a few immediate consequences of the induction hypotheses (40), which we summarize below. These conditions are useful in the subsequent analysis. Note that Lemma 3 is incorporated here.

Lemma 16. *Suppose that $m \geq Cn\log^6 m$ for some sufficiently large constant $C > 0$. Then under the hypotheses (40) for $t \lesssim \log n$, with probability at least $1 - O(me^{-1.5n}) - O(m^{-10})$ one has*

$$c_5/2 \leq \left\|\boldsymbol{x}_\perp^{t,(l)}\right\|_2 \leq \left\|\boldsymbol{x}^{t,(l)}\right\|_2 \leq 2C_5; \tag{61a}$$

$$c_5/2 \leq \left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right\|_2 \leq \left\|\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 \leq 2C_5; \tag{61b}$$

$$c_5/2 \leq \left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2 \leq \left\|\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2 \leq 2C_5; \tag{61c}$$

$$\max_{1\leq l\leq m} \left|\boldsymbol{a}_l^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}^t\right\|_2; \tag{62a}$$

$$\max_{1\leq l\leq m} \left|\boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^t\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}_\perp^t\right\|_2; \tag{62b}$$

$$\max_{1\leq l\leq m} \left|\boldsymbol{a}_l^\top \boldsymbol{x}^{t,\mathrm{sgn}}\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2; \tag{62c}$$

$$\max_{1\leq l\leq m} \left|\boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right\|_2; \tag{62d}$$

$$\max_{1\leq l\leq m} \left|\boldsymbol{a}_l^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn}}\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2; \tag{62e}$$

$$\max_{1\leq l\leq m} \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \ll \frac{1}{\log m}; \tag{63a}$$

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 \ll \frac{1}{\log m}; \tag{63b}$$

$$\max_{1\leq l\leq m} \left|x_\parallel^{t,(l)}\right| \leq 2\alpha_t. \tag{63c}$$

Proof. See Appendix M. □

B Proof of Lemma 1

We focus on the case when

$$\frac{1}{\sqrt{n\log n}} \leq \alpha_0 \leq \frac{\log n}{\sqrt{n}} \qquad \text{and} \qquad 1 - \frac{1}{\log n} \leq \beta_0 \leq 1 + \frac{1}{\log n}$$

The other cases can be proved using very similar arguments as below, and hence omitted.

Let $\eta > 0$ and $c_4 > 0$ be some sufficiently small constants independent of n. In the sequel, we divide Stage 1 (iterations up to T_γ) into several substages. See Figure 9 for an illustration.

- **Stage 1.1:** consider the period when α_t is sufficiently small, which consists of all iterations $0 \leq t \leq T_1$ with T_1 given in (26). We claim that, throughout this substage,

$$\alpha_t > \frac{1}{2\sqrt{n\log n}}, \tag{64a}$$

$$\sqrt{0.5} < \beta_t < \sqrt{1.5}. \tag{64b}$$

  If this claim holds, then we would have $\alpha_t^2 + \beta_t^2 < c_4^2 + 1.5 < 2$ as long as c_4 is small enough. This immediately reveals that $1 + \eta\left(1 - 3\alpha_t^2 - 3\beta_t^2\right) \geq 1 - 6\eta$, which further gives

$$\begin{aligned}
\beta_{t+1} &\geq \left\{1 + \eta\left(1 - 3\alpha_t^2 - 3\beta_t^2\right) + \eta\rho_t\right\}\beta_t \\
&\geq \left(1 - 6\eta - \frac{c_3\eta}{\log n}\right)\beta_t \\
&\geq (1 - 7\eta)\beta_t.
\end{aligned} \tag{65}$$

  In what follows, we further divide this stage into multiple sub-phases.

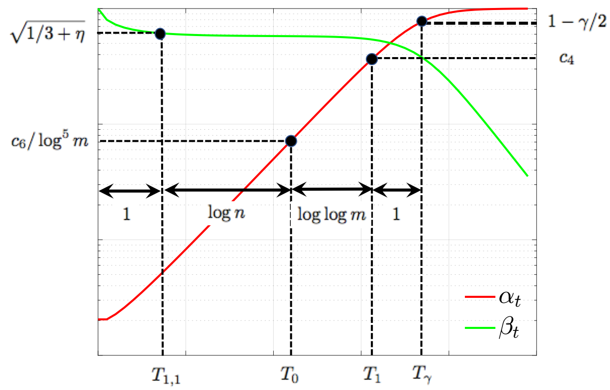


Figure 9: Illustration of the substages for the proof of Lemma 1.

– **Stage 1.1.1**: consider the iterations $0 \le t \le T_{1,1}$ with

$$T_{1,1} = \min\left\{t \mid \beta_{t+1} \le \sqrt{1/3+\eta}\right\}. \tag{66}$$

Fact 1. *For any sufficiently small $\eta > 0$, one has*

$$\begin{aligned}
\beta_{t+1} &\le (1-2\eta^2)\beta_t, \qquad 0 \le t \le T_{1,1}; &(67)\\
\alpha_{t+1} &\le (1+4\eta)\alpha_t, \qquad 0 \le t \le T_{1,1};\\
\alpha_{t+1} &\ge (1+2\eta^3)\alpha_t, \qquad 1 \le t \le T_{1,1}; &(68)\\
\alpha_1 &\ge \alpha_0/2;\\
\beta_{T_{1,1}+1} &\ge \frac{1-7\eta}{\sqrt{3}};\\
T_{1,1} &\lesssim \frac{1}{\eta^2}. &(69)
\end{aligned}$$

Moreover, $\alpha_{T_{1,1}} \ll c_4$ and hence $T_{1,1} < T_1$.

From Fact 1, we see that in this substage, α_t keeps increasing (at least for $t \ge 1$) with

$$c_4 > \alpha_t \ge \frac{\alpha_0}{2} \ge \frac{1}{2\sqrt{n\log n}}, \qquad 0 \le t \le T_{1,1},$$

and β_t is strictly decreasing with

$$1.5 > \beta_0 \ge \beta_t \ge \beta_{T_{1,1}+1} \ge \frac{1-7\eta}{\sqrt{3}}, \qquad 0 \le t \le T_{1,1},$$

which justifies (64). In addition, combining (67) with (68), we arrive at the growth rate of α_t/β_t as

$$\frac{\alpha_{t+1}/\alpha_t}{\beta_{t+1}/\beta_t} \ge \frac{1+2\eta^3}{1-2\eta^2} = 1 + O(\eta^2).$$

These demonstrate (24) for this substage.

– **Stage 1.1.2:** this substage contains all iterations obeying $T_{1,1} < t \le T_1$. We claim the following result.

Fact 2. *Suppose that $\eta > 0$ is sufficiently small. Then for any $T_{1,1} < t \le T_1$,*

$$\beta_t \in \left[\frac{(1-7\eta)^2}{\sqrt{3}}, \frac{1+30\eta}{\sqrt{3}}\right]; \tag{70}$$

$$\beta_{t+1} \le (1+30\eta^2)\beta_t. \tag{71}$$

Furthermore, since

$$\alpha_t^2 + \beta_t^2 \le c_4^2 + \frac{(1+30\eta)^2}{3} < \frac{1}{2},$$

we have, for sufficiently small c_3, that

$$\begin{aligned}
\alpha_{t+1} &\ge \left\{1 + 3\eta\left(1 - \alpha_t^2 - \beta_t^2\right) - \eta|\zeta_t|\right\}\alpha_t \\
&\ge \left(1 + 1.5\eta - \frac{c_3\eta}{\log n}\right)\alpha_t \\
&\ge (1 + 1.4\eta)\alpha_t, \qquad (72)
\end{aligned}$$

and hence α_t keeps increasing. This means $\alpha_t \ge \alpha_1 \ge \frac{1}{2\sqrt{n\log n}}$, which justifies the claim (64) together with (70) for this substage. As a consequence,

$$\begin{aligned}
T_1 - T_{1,1} &\lesssim \frac{\log\frac{c_4}{\alpha_0}}{\log(1+1.4\eta)} \lesssim \frac{\log n}{\eta}; \\
T_1 - T_0 &\lesssim \frac{\log\frac{c_4}{\frac{c_6}{\log^5 m}}}{\log\left(1+1.4\eta\right)} \lesssim \frac{\log\log m}{\eta}.
\end{aligned}$$

Moreover, combining (72) with (71) yields the growth rate of α_t/β_t as

$$\frac{\alpha_{t+1}/\alpha_t}{\beta_{t+1}/\beta_t} \ge \frac{1+1.4\eta}{1+30\eta^2} \ge 1+\eta$$

for $\eta > 0$ sufficiently small.

– Taken collectively, the preceding bounds imply that

$$T_1 = T_{1,1} + (T_1 - T_{1,1}) \lesssim \frac{1}{\eta^2} + \frac{\log n}{\eta} \lesssim \frac{\log n}{\eta^2}.$$

- **Stage 1.2:** in this stage, we consider all iterations $T_1 < t \le T_2$, where

$$T_2 := \min\left\{t \mid \frac{\alpha_{t+1}}{\beta_{t+1}} > \frac{2}{\gamma}\right\}.$$

From the preceding analysis, it is seen that, for η sufficiently small,

$$\frac{\alpha_{T_{1,1}}}{\beta_{T_{1,1}}} \le \frac{c_4}{\frac{(1-7\eta)^2}{\sqrt{3}}} \le \frac{\sqrt{3}c_4}{1-15\eta}.$$

In addition, we have:

Fact 3. *Suppose $\eta > 0$ is sufficiently small. Then for any $T_1 < t \le T_2$, one has*

$$\begin{aligned}
\alpha_t^2 + \beta_t^2 &\le 2; \qquad (73)\\
\frac{\alpha_{t+1}/\beta_{t+1}}{\alpha_t/\beta_t} &\ge 1+\eta; \qquad (74)\\
\alpha_{t+1} &\ge \left\{1 - 3.1\eta\right\}\alpha_t; \qquad (75)\\
\beta_{t+1} &\ge \left\{1 - 5.1\eta\right\}\beta_t. \qquad (76)
\end{aligned}$$

In addition,

$$T_2 - T_1 \lesssim \frac{1}{\eta}.$$

With this fact in place, one has

$$\alpha_t \geq (1-3.1\eta)^{t-T_1}\alpha_{T_1} \gtrsim 1, \qquad T_1 < t \leq T_2.$$

and hence

$$\beta_t \geq (1-5.1\eta)^{t-T_1}\beta_{T_1} \gtrsim 1, \qquad T_1 < t \leq T_2.$$

These taken collectively demonstrate (24) for any $T_1 < t \leq T_2$. Finally, if $T_2 \geq T_\gamma$, then we complete the proof as

$$T_\gamma \leq T_2 = T_1 + (T_2 - T_1) \lesssim \frac{\log n}{\eta^2}.$$

Otherwise we move to the next stage.

- **Stage 1.3:** this stage is composed of all iterations $T_2 < t \leq T_\gamma$. We break the discussion into two cases.
  - If $\alpha_{T_2+1} > 1+\gamma$, then $\alpha_{T_2+1}^2 + \beta_{T_2+1}^2 \geq \alpha_{T_2+1}^2 > 1+2\gamma$. This means that

    $$\begin{aligned}\alpha_{T_2+2} &\leq \left\{1 + 3\eta\left(1-\alpha_{T_2+1}^2-\beta_{T_2+1}^2\right) + \eta|\zeta_{T_2+1}|\right\}\alpha_{T_2+1} \\ &\leq \left\{1 - 6\eta\gamma - \frac{\eta c_3}{\log n}\right\}\alpha_{T_2+1} \\ &\leq \{1-5\eta\gamma\}\,\alpha_{T_2+1}\end{aligned}$$

    when $c_3 > 0$ is sufficiently small. Similarly, one also gets $\beta_{T_2+2} \leq (1-5\eta\gamma)\beta_{T_2+1}$. As a result, both α_t and β_t will decrease. Repeating this argument reveals that

    $$\begin{aligned}\alpha_{t+1} &\leq (1-5\eta\gamma)\alpha_t, \\ \beta_{t+1} &\leq (1-5\eta\gamma)\beta_t\end{aligned}$$

    until $\alpha_t \leq 1+\gamma$. In addition, applying the same argument as for Stage 1.2 yields

    $$\frac{\alpha_{t+1}/\alpha_t}{\beta_{t+1}/\beta_t} \geq 1 + c_{10}\eta$$

    for some constant $c_{10} > 0$. Therefore, when α_t drops below $1+\gamma$, one has

    $$\alpha_t \geq (1-3\eta)(1+\gamma) \geq 1-\gamma$$

    and

    $$\beta_t \leq \frac{\gamma}{2}\alpha_t \leq \gamma.$$

    This justifies that

    $$T_\gamma - T_2 \lesssim \frac{\log\frac{2}{1-\gamma}}{-\log(1-5\eta\gamma)} \lesssim \frac{1}{\eta}.$$

  - If $c_4 \leq \alpha_{T_2+1} < 1-\gamma$, take very similar arguments as in Stage 1.2 to reach that

    $$\frac{\alpha_{t+1}/\alpha_t}{\beta_{t+1}/\beta_t} \geq 1 + c_{10}\eta, \qquad T_\gamma - T_2 \lesssim \frac{1}{\eta}$$

    $$\text{and} \qquad \alpha_t \gtrsim 1, \quad \beta_t \gtrsim 1 \qquad T_2 \leq t \leq T_\gamma$$

    for some constant $c_{10} > 0$. We omit the details for brevity.

In either case, we see that α_t is always bounded away from 0. We can also repeat the argument for Stage 1.2 to show that $\beta_t \gtrsim 1$.

In conclusion, we have established that

$$T_\gamma = T_1 + (T_2 - T_1) + (T_\gamma - T_2) \lesssim \frac{\log n}{\eta^2}, \qquad 0 \le t < T_\gamma$$

$$\text{and} \qquad \frac{\alpha_{t+1}/\alpha_t}{\beta_{t+1}/\beta_t} \ge 1 + c_{10}\eta^2, \qquad c_5 \le \beta_t \le 1.5, \qquad \frac{1}{2\sqrt{n \log n}} \le \alpha_t \le 2, \qquad 0 \le t < T_\gamma$$

for some constants $c_5, c_{10} > 0$.

Proof of Fact 1. The proof proceeds as follows.

- First of all, for any $0 \le t \le T_{1,1}$, one has $\beta_t \ge \sqrt{1/3 + \eta}$ and $\alpha_t^2 + \beta_t^2 \ge 1/3 + \eta$ and, as a result,

$$\begin{aligned} \beta_{t+1} &\le \left\{1 + \eta\left(1 - 3\alpha_t^2 - 3\beta_t^2\right) + \eta|\rho_t|\right\}\beta_t \\ &\le \left(1 - 3\eta^2 + \frac{c_3\eta}{\log n}\right)\beta_t \\ &\le (1 - 2\eta^2)\beta_t \end{aligned} \tag{77}$$

  as long as c_3 and η are both constants. In other words, β_t is strictly decreasing before $T_{1,1}$, which also justifies the claim (64b) for this substage.

- Moreover, given that the contraction factor of β_t is at least $1 - 2\eta^2$, we have

$$T_{1,1} \lesssim \frac{\log \frac{\beta_0}{\sqrt{1/3+\eta}}}{-\log\left(1 - 2\eta^2\right)} \asymp \frac{1}{\eta^2}.$$

  This upper bound also allows us to conclude that β_t will cross the threshold $\sqrt{1/3 + \eta}$ before α_t exceeds c_4, namely, $T_{1,1} < T_1$. To see this, we note that the growth rate of $\{\alpha_t\}$ within this substage is upper bounded by

$$\begin{aligned} \alpha_{t+1} &\le \left\{1 + 3\eta\left(1 - \alpha_t^2 - \beta_t^2\right) + \eta|\zeta_t|\right\}\alpha_t \\ &\le \left(1 + 3\eta + \frac{c_3\eta}{\log n}\right)\alpha_t \\ &\le (1 + 4\eta)\alpha_t. \end{aligned} \tag{78}$$

  This leads to an upper bound

$$|\alpha_{T_{1,1}}| \le (1 + 4\eta)^{T_{1,1}}|\alpha_0| \le (1 + 4\eta)^{O(\eta^{-2})}\frac{\log n}{\sqrt{n}} \ll c_4. \tag{79}$$

- Furthermore, we can also lower bound α_t. First of all,

$$\begin{aligned} \alpha_1 &\ge \left\{1 + 3\eta\left(1 - \alpha_0^2 - \beta_0^2\right) - \eta|\zeta_t|\right\}\alpha_0 \\ &\ge \left(1 - 3\eta - \frac{c_3\eta}{\log n}\right)\alpha_0 \\ &\ge (1 - 4\eta)\alpha_0 \ge \frac{1}{2}\alpha_0 \end{aligned}$$

  for η sufficiently small. For all $1 \le t \le T_{1,1}$, using (78) we have

$$\alpha_t^2 + \beta_t^2 \le (1 + 4\eta)^{T_{1,1}}\alpha_0^2 + \beta_1^2 \le o(1) + (1 - 2\eta^2)\beta_0 \le 1 - \eta^2,$$

  allowing one to deduce that

$$\alpha_{t+1} \ge \left\{1 + 3\eta\left(1 - \alpha_t^2 - \beta_t^2\right) - \eta|\zeta_t|\right\}\alpha_t$$

$$
\begin{aligned}
&\geq \left(1+3\eta^3-\frac{c_3\eta}{\log n}\right)\alpha_t \\
&\geq (1+2\eta^3)\alpha_t.
\end{aligned}
$$

In other words, α_t keeps increasing throughout all $1 \leq t \leq T_{1,1}$. This verifies the condition (64a) for this substage.

- Finally, we make note of one useful lower bound

$$
\beta_{T_{1,1}+1} \geq (1-7\eta)\beta_{T_{1,1}} \geq \frac{1-7\eta}{\sqrt{3}}, \tag{80}
$$

which follows by combining (65) and the condition $\beta_{T_{1,1}} \geq \sqrt{1/3+\eta}$.

□

Proof of Fact 2. Clearly, $\beta_{T_{1,1}+1}$ falls within this range according to (66) and (80). We now divide into several cases.

- If $\frac{1+\eta}{\sqrt{3}} \leq \beta_t < \frac{1+30\eta}{\sqrt{3}}$, then $\alpha_t^2+\beta_t^2 \geq \beta_t^2 \geq (1+\eta)^2/3$, and hence the next iteration obeys

$$
\begin{aligned}
\beta_{t+1} &\leq \left\{1+\eta\left(1-3\beta_t^2\right)+\eta|\rho_t|\right\}\beta_t \\
&\leq \left(1+\eta\left(1-(1+\eta)^2\right)+\frac{c_3\eta}{\log n}\right)\beta_t \\
&\leq (1-\eta^2)\beta_t
\end{aligned} \tag{81}
$$

and, in view of (65), $\beta_{t+1} \geq (1-7\eta)\beta_t \geq \frac{1-7\eta}{\sqrt{3}}$. In summary, in this case one has $\beta_{t+1} \in \left[\frac{1-7\eta}{\sqrt{3}}, \frac{1+30\eta}{\sqrt{3}}\right]$, which still resides within the range (70).

- If $\frac{(1-7\eta)^2}{\sqrt{3}} \leq \beta_t \leq \frac{1-7\eta}{\sqrt{3}}$, then $\alpha_t^2+\beta_t^2 < c_4^2+(1-7\eta)^2/3 < (1-7\eta)/3$ for c_4 sufficiently small. Consequently, for a small enough c_3 one has

$$
\begin{aligned}
\beta_{t+1} &\geq \left\{1+\eta\left(1-3\alpha_t^2-3\beta_t^2\right)-\eta|\rho_t|\right\}\beta_t \\
&\geq \left(1+7\eta^2-\frac{c_3\eta}{\log n}\right)\beta_t \\
&\geq (1+6\eta^2)\beta_t.
\end{aligned}
$$

In other words, β_{t+1} is strictly larger than β_t. Moreover, recognizing that $\alpha_t^2+\beta_t^2 > (1-7\eta)^4/3 > (1-29\eta)/3$, one has

$$
\begin{aligned}
\beta_{t+1} &\leq \left\{1+\eta\left(1-3\alpha_t^2-3\beta_t^2\right)+\eta|\rho_t|\right\}\beta_t \\
&\leq \left(1+29\eta^2+\frac{c_3\eta}{\log n}\right)\beta_t \leq (1+30\eta^2)\beta_t \\
&< \frac{1+30\eta^2}{\sqrt{3}}.
\end{aligned} \tag{82}
$$

Therefore, we have shown that $\beta_{t+1} \in \left[\frac{(1-7\eta)^2}{\sqrt{3}}, \frac{1+30\eta}{\sqrt{3}}\right]$, which continues to lie within the range (70).

- Finally, if $\frac{1-7\eta}{\sqrt{3}} < \beta_t < \frac{1+\eta}{\sqrt{3}}$, we have $\alpha_t^2+\beta_t^2 \geq \frac{(1-7\eta)^2}{3} \geq \frac{1-15\eta}{3}$ for η sufficiently small, which implies

$$
\begin{aligned}
\beta_{t+1} &\leq \left\{1+15\eta^2+\eta|\rho_t|\right\}\beta_t \leq (1+16\eta^2)\beta_t \\
&\leq \frac{(1+16\eta^2)(1+\eta)}{\sqrt{3}} \leq \frac{1+2\eta}{\sqrt{3}}
\end{aligned} \tag{83}
$$

for small $\eta > 0$. In addition, it comes from (80) that $\beta_{t+1} \geq (1-7\eta)\beta_t \geq \frac{(1-7\eta)^2}{\sqrt{3}}$. This justifies that β_{t+1} falls within the range (70).

Combining all of the preceding cases establishes the claim (70) for all $T_{1,1} < t \le T_1$. $\square$

Proof of Fact 3. We first demonstrate that

$$\alpha_t^2 + \beta_t^2 \le 2 \tag{84}$$

throughout this substage. In fact, if $\alpha_t^2 + \beta_t^2 \le 1.5$, then

$$\alpha_{t+1} \le \left\{1 + 3\eta\left(1 - \alpha_t^2 - \beta_t^2\right) + \eta|\zeta_t|\right\}\alpha_t \le (1+4\eta)\,\alpha_t$$

and, similarly, $\beta_{t+1} \le (1+4\eta)\beta_t$. These taken together imply that

$$\alpha_{t+1}^2 + \beta_{t+1}^2 \le (1+4\eta)^2\left(\alpha_t^2 + \beta_t^2\right) \le 1.5(1+9\eta) < 2.$$

Additionally, if $1.5 < \alpha_t^2 + \beta_t^2 \le 2$, then

$$\begin{aligned}
\alpha_{t+1} &\le \left\{1 + 3\eta\left(1 - \alpha_t^2 - \beta_t^2\right) + \eta|\zeta_t|\right\}\alpha_t \\
&\le \left(1 - 1.5\eta + \frac{c_3\eta}{\log n}\right)\alpha_t \\
&\le (1-\eta)\alpha_t
\end{aligned}$$

and, similarly, $\beta_{t+1} \le (1-\eta)\beta_t$. These reveal that

$$\alpha_{t+1}^2 + \beta_{t+1}^2 \le \alpha_t^2 + \beta_t^2.$$

Put together the above argument to establish the claim (84).

With the claim (84) in place, we can deduce that

$$\begin{aligned}
\alpha_{t+1} &\ge \left\{1 + 3\eta\left(1 - \alpha_t^2 - \beta_t^2\right) - \eta|\zeta_t|\right\}\alpha_t \\
&\ge \left\{1 + 3\eta\left(1 - \alpha_t^2 - \beta_t^2\right) - 0.1\eta\right\}\alpha_t
\end{aligned} \tag{85}$$

and

$$\begin{aligned}
\beta_{t+1} &\le \left\{1 + \eta\left(1 - 3\alpha_t^2 - 3\beta_t^2\right) + \eta|\rho_t|\right\}\beta_t \\
&\le \left\{1 + \eta\left(1 - 3\alpha_t^2 - 3\beta_t^2\right) + 0.1\eta\right\}\beta_t.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\frac{\alpha_{t+1}/\beta_{t+1}}{\alpha_t/\beta_t} = \frac{\alpha_{t+1}/\alpha_t}{\beta_{t+1}/\beta_t} &\ge \frac{1 + 3\eta\left(1 - \alpha_t^2 - \beta_t^2\right) - 0.1\eta}{1 + \eta\left(1 - 3\alpha_t^2 - 3\beta_t^2\right) + 0.1\eta} \\
&= 1 + \frac{1.8\eta}{1 + \eta\left(1 - 3\alpha_t^2 - 3\beta_t^2\right) + 0.1\eta} \\
&\ge 1 + \frac{1.8\eta}{1+2\eta} \ge 1 + \eta
\end{aligned}$$

for $\eta > 0$ sufficiently small. This immediately implies that

$$T_2 - T_1 \lesssim \frac{\log\left(\frac{2/\gamma}{\alpha_{T_1}/\beta_{T_1}}\right)}{\log(1+\eta)} \asymp \frac{1}{\eta}.$$

Moreover, combine (84) and (85) to arrive at

$$\alpha_{t+1} \ge \{1 - 3.1\eta\}\,\alpha_t, \tag{86}$$

Similarly, one can show that $\beta_{t+1} \ge \{1 - 5.1\eta\}\,\beta_t$. $\square$

C Proof of Lemma 2

C.1 Proof of (41a)

In view of the gradient update rule (3), we can express the signal component $x_{\|}^{t+1}$ as follows

$$x_{\|}^{t+1} = x_{\|}^{t} - \frac{\eta}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^t\right)^3 - a_{i,1}^2\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^t\right)\right]a_{i,1}.$$

Expanding this expression using $\boldsymbol{a}_i^{\top}\boldsymbol{x}^t = x_{\|}^t a_{i,1} + \boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^t$ and rearranging terms, we are left with

$$\begin{aligned}x_{\|}^{t+1} = x_{\|}^{t} &+ \underbrace{\eta\left[1-\left(x_{\|}^t\right)^2\right]x_{\|}^t\cdot\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^4}_{:=J_1} + \underbrace{\eta\left[1-3\left(x_{\|}^t\right)^2\right]\cdot\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^3\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^t}_{:=J_2}\\ &\underbrace{-3\eta x_{\|}^t\cdot\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^t\right)^2a_{i,1}^2}_{:=J_3}\underbrace{-\eta\cdot\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^t\right)^3a_{i,1}}_{:=J_4}.\end{aligned}$$

In the sequel, we control the above four terms J_1, J_2, J_3 and J_4 separately.

- With regard to the first term J_1, it follows from the standard concentration inequality for Gaussian polynomials [SS12, Theorem 1.9] that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^4-3\right|\geq\tau\right)\leq e^2e^{-c_1m^{1/4}\tau^{1/2}}$$

  for some absolute constant $c_1>0$. Taking $\tau\asymp\frac{\log^3 m}{\sqrt{m}}$ reveals that with probability exceeding $1-O\left(m^{-10}\right)$,

$$\begin{aligned}J_1 &= 3\eta\left[1-\left(x_{\|}^t\right)^2\right]x_{\|}^t + \left(\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^4-3\right)\eta\left[1-\left(x_{\|}^t\right)^2\right]x_{\|}^t\\ &= 3\eta\left[1-\left(x_{\|}^t\right)^2\right]x_{\|}^t + r_1,\end{aligned}\tag{87}$$

  where the remainder term r_1 obeys

$$|r_1| = O\left(\frac{\eta\log^3 m}{\sqrt{m}}\left|x_{\|}^t\right|\right).$$

  Here, the last line also uses the fact that

$$\left|1-\left(x_{\|}^t\right)^2\right|\leq 1+\left\|\boldsymbol{x}^t\right\|_2^2\lesssim 1,\tag{88}$$

  with the last relation coming from the induction hypothesis (40e).

- For the third term J_3, it is easy to see that

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^t\right)^2a_{i,1}^2 - \left\|\boldsymbol{x}_{\perp}^t\right\|_2^2 = \boldsymbol{x}_{\perp}^{t\top}\Bigg[\underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{\natural}\right)^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}}_{:=\boldsymbol{U}}-\boldsymbol{I}_{n-1}\Bigg]\boldsymbol{x}_{\perp}^t,\tag{89}$$

  where $\boldsymbol{U}-\boldsymbol{I}_{n-1}$ is a submatrix of the following matrix (obtained by removing its first row and column)

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{\natural}\right)^2\boldsymbol{a}_i\boldsymbol{a}_i^{\top} - \left(\boldsymbol{I}_n + 2\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top}\right).\tag{90}$$

This fact combined with Lemma 14 reveals that

$$\|\boldsymbol{U} - \boldsymbol{I}_{n-1}\| \le \left\| \frac{1}{m}\sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2 \boldsymbol{a}_i\boldsymbol{a}_i^\top - \left(\boldsymbol{I}_n + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top}\right)\right\| \lesssim \sqrt{\frac{n\log^3 m}{m}}$$

with probability at least $1 - O\left(m^{-10}\right)$, provided that $m \gg n\log^3 m$. This further implies

$$J_3 = 3\eta \left\|\boldsymbol{x}_\perp^t\right\|_2^2 x_\parallel^t + r_2, \tag{91}$$

where the size of the remaining term r_2 satisfies

$$|r_2| \lesssim \eta\sqrt{\frac{n\log^3 m}{m}}|x_\parallel^t|\left\|\boldsymbol{x}_\perp^t\right\|_2^2 \lesssim \eta\sqrt{\frac{n\log^3 m}{m}}|x_\parallel^t|.$$

Here, the last inequality holds under the hypothesis (40e) that $\|\boldsymbol{x}_\perp^t\|_2^2 \le \|\boldsymbol{x}^t\|_2^2 \lesssim 1$.

- When it comes to J_2, our analysis relies on the random-sign sequence $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$. Specifically, one can decompose

$$\frac{1}{m}\sum_{i=1}^{m} a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t = \frac{1}{m}\sum_{i=1}^{m} a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \frac{1}{m}\sum_{i=1}^{m} a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right). \tag{92}$$

For the first term on the right-hand side of (92), note that $|a_{i,1}|^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}$ is statistically independent of $\xi_i = \mathrm{sgn}\left(a_{i,1}\right)$. Therefore we can treat $\frac{1}{m}\sum_{i=1}^m a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}$ as a weighted sum of the ξ_i's and apply the Bernstein inequality (see Lemma 11) to arrive at

$$\left|\frac{1}{m}\sum_{i=1}^{m} a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right| = \left|\frac{1}{m}\sum_{i=1}^{m} \xi_i \left(|a_{i,1}|^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right)\right| \lesssim \frac{1}{m}\left(\sqrt{V_1 \log m} + B_1 \log m\right) \tag{93}$$

with probability exceeding $1 - O\left(m^{-10}\right)$, where

$$V_1 := \sum_{i=1}^{m} |a_{i,1}|^6 \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right)^2 \qquad \text{and} \qquad B_1 := \max_{1\le i\le m} |a_{i,1}|^3 \left|\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right|.$$

Make use of Lemma 12 and the incoherence condition (62d) to deduce that with probability at least $1 - O\left(m^{-10}\right)$,

$$\frac{1}{m}V_1 = \frac{1}{m}\sum_{i=1}^{m} |a_{i,1}|^6 \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right)^2 \lesssim \left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right\|_2^2$$

with the proviso that $m \gg n\log^5 m$. Furthermore, the incoherence condition (62d) together with the fact (56) implies that

$$B_1 \lesssim \log^2 m \left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right\|_2.$$

Substitute the bounds on V_1 and B_1 back to (93) to obtain

$$\left|\frac{1}{m}\sum_{i=1}^{m} a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right| \lesssim \sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right\|_2 + \frac{\log^3 m}{m}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right\|_2 \asymp \sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right\|_2 \tag{94}$$

as long as $m \gtrsim \log^5 m$. Additionally, regarding the second term on the right-hand side of (92), one sees that

$$\frac{1}{m}\sum_{i=1}^{m} a_{i,1}^3 \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right) = \underbrace{\frac{1}{m}\sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2 a_{i,1}\boldsymbol{a}_{i,\perp}^\top}_{:=\boldsymbol{u}^\top} \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}}\right), \tag{95}$$

where $\boldsymbol{u}$ is the first column of (90) without the first entry. Hence we have

$$\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^{3}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)\right|\leq\|\boldsymbol{u}\|_{2}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}\lesssim\sqrt{\frac{n\log^{3}m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2},\tag{96}$$

with probability exceeding $1-O\left(m^{-10}\right)$, with the proviso that $m\gg n\log^{3}m$. Substituting the above two bounds (94) and (96) back into (92) gives

$$\begin{aligned}\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^{3}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right| &\leq\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^{3}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right|+\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^{3}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)\right|\\ &\lesssim\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}+\sqrt{\frac{n\log^{3}m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}.\end{aligned}$$

As a result, we arrive at the following bound on J_2:

$$\begin{aligned}|J_{2}| &\lesssim\eta\left|1-3\left(x_{\|}^{t}\right)^{2}\right|\left(\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}+\sqrt{\frac{n\log^{3}m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}\right)\\ &\overset{\text{(i)}}{\lesssim}\eta\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}+\eta\sqrt{\frac{n\log^{3}m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}\\ &\overset{\text{(ii)}}{\lesssim}\eta\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_{2}+\eta\sqrt{\frac{n\log^{3}m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2},\end{aligned}$$

where (i) uses (88) again and (ii) comes from the triangle inequality $\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}\leq\|\boldsymbol{x}_{\perp}^{t}\|_{2}+\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}$ and the fact that $\sqrt{\frac{\log m}{m}}\leq\sqrt{\frac{n\log^{3}m}{m}}$.

- It remains to control J_4, towards which we resort to the random-sign sequence $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ once again. Write

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right)^{3}a_{i,1}=\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^{3}a_{i,1}+\frac{1}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right)^{3}-\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^{3}\right]a_{i,1}.\tag{97}$$

For the first term in (97), since $\xi_i=\mathrm{sgn}\left(a_{i,1}\right)$ is statistically independent of $\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^{3}|a_{i,1}|$, we can upper bound the first term using the Bernstein inequality (see Lemma 11) as

$$\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^{3}|a_{i,1}|\,\xi_{i}\right|\lesssim\frac{1}{m}\left(\sqrt{V_{2}\log m}+B_{2}\log m\right),$$

where the quantities V_2 and B_2 obey

$$V_{2}:=\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^{6}|a_{i,1}|^{2}\qquad\text{and}\qquad B_{2}:=\max_{1\leq i\leq m}\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right|^{3}|a_{i,1}|.$$

Using similar arguments as in bounding (93) yields

$$V_{2}\lesssim m\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}^{6}\qquad\text{and}\qquad B_{2}\lesssim\log^{2}m\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}^{3}$$

with the proviso that $m\gg n\log^{5}m$ and

$$\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^{3}|a_{i,1}|\,\xi_{i}\right|\lesssim\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}^{3}+\frac{\log^{3}m}{m}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}^{3}\asymp\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}^{3},\tag{98}$$

with probability exceeding $1 - O(m^{-10})$ as soon as $m \gtrsim \log^5 m$. Regarding the second term in (97),

$$
\begin{aligned}
&\left| \frac{1}{m} \sum_{i=1}^{m} \left[\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t \right)^3 - \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right)^3 \right] a_{i,1} \right| \\
&\overset{\text{(i)}}{=} \frac{1}{m} \sum_{i=1}^{m} \left| \left\{ \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right) \left[\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t \right)^2 + \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right)^2 + \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t \right) \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right) \right] \right\} a_{i,1} \right| \\
&\overset{\text{(ii)}}{\leq} \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left[\boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right) \right]^2} \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left[5 \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t \right)^4 + 5 \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right)^4 \right] a_{i,1}^2}. \qquad (99)
\end{aligned}
$$

Here, the first equality (i) utilizes the elementary identity $a^3 - b^3 = (a - b)\left(a^2 + b^2 + ab\right)$, and (ii) follows from the Cauchy-Schwarz inequality as well as the inequality

$$(a^2 + b^2 + ab)^2 \leq (1.5a^2 + 1.5b^2)^2 \leq 5a^4 + 5b^4.$$

Use Lemma 13 to reach

$$\sqrt{\frac{1}{m} \sum_{i=1}^{m} \left[\boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right) \right]^2} = \sqrt{\left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right)^\top \left(\frac{1}{m} \sum_{i=1}^{m} \boldsymbol{a}_{i,\perp} \boldsymbol{a}_{i,\perp}^\top \right) \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right)} \lesssim \left\| \boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2.$$

Additionally, combining Lemma 12 and the incoherence conditions (62b) and (62d), we can obtain

$$\sqrt{\frac{1}{m} \sum_{i=1}^{m} \left[5 \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t \right)^4 + 5 \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right)^4 \right] a_{i,1}^2} \lesssim \left\| \boldsymbol{x}_\perp^t \right\|_2^2 + \left\| \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2^2 \lesssim 1,$$

as long as $m \gg n \log^6 m$. Here, the last relation comes from the norm conditions (40e) and (61b). These in turn imply

$$\left| \frac{1}{m} \sum_{i=1}^{m} \left[\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t \right)^3 - \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right)^3 \right] a_{i,1} \right| \lesssim \left\| \boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2. \qquad (100)$$

Combining the above bounds (98) and (100), we get

$$
\begin{aligned}
|J_4| &\leq \eta \left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right)^3 a_{i,1} \right| + \eta \left| \frac{1}{m} \sum_{i=1}^{m} \left[\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t \right)^3 - \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right)^3 \right] a_{i,1} \right| \\
&\lesssim \eta \sqrt{\frac{\log m}{m}} \left\| \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2^3 + \eta \left\| \boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2 \\
&\lesssim \eta \sqrt{\frac{\log m}{m}} \left\| \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2 + \eta \left\| \boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2 \\
&\lesssim \eta \sqrt{\frac{\log m}{m}} \left\| \boldsymbol{x}_\perp^t \right\|_2 + \eta \left\| \boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2,
\end{aligned}
$$

where the penultimate inequality arises from the norm condition (61b) and the last one comes from the triangle inequality $\left\| \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2 \leq \left\| \boldsymbol{x}_\perp^t \right\|_2 + \left\| \boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} \right\|_2$ and the fact that $\sqrt{\frac{\log m}{m}} \leq 1$.

- Putting together the above estimates for J_1, J_2, J_3 and J_4, we reach

$$
\begin{aligned}
x_\parallel^{t+1} &= x_\parallel^t + J_1 - J_3 + J_2 - J_4 \\
&= x_\parallel^t + 3\eta \left[1 - \left(x_\parallel^t \right)^2 \right] x_\parallel^t - 3\eta \left\| \boldsymbol{x}_\perp^t \right\|_2^2 x_\parallel^t + R_1 \\
&= \left\{ 1 + 3\eta \left(1 - \left\| \boldsymbol{x}^t \right\|_2^2 \right) \right\} x_\parallel^t + R_1, \qquad (101)
\end{aligned}
$$

where R_1 is the residual term obeying

$$|R_1| \lesssim \eta\sqrt{\frac{n\log^3 m}{m}}\left|x_{\|}^t\right| + \eta\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^t\right\|_2 + \eta\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2.$$

Substituting the hypotheses (40) into (101) and recalling that $\alpha_t = \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural\rangle$ lead us to conclude that

$$\begin{aligned}
\alpha_{t+1} &= \left\{1 + 3\eta\left(1 - \left\|\boldsymbol{x}^t\right\|_2^2\right)\right\}\alpha_t + O\left(\eta\sqrt{\frac{n\log^3 m}{m}}\alpha_t\right) + O\left(\eta\sqrt{\frac{\log m}{m}}\beta_t\right) \\
&\quad + O\left(\eta\alpha_t\left(1 + \frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}}\right) \\
&= \left\{1 + 3\eta\left(1 - \left\|\boldsymbol{x}^t\right\|_2^2\right) + \eta\zeta_t\right\}\alpha_t, \qquad (102)
\end{aligned}$$

for some $|\zeta_t| \ll \frac{1}{\log m}$, provided that

$$\sqrt{\frac{n\log^3 m}{m}} \ll \frac{1}{\log m} \qquad (103\text{a})$$

$$\sqrt{\frac{\log m}{m}}\beta_t \ll \frac{1}{\log m}\alpha_t \qquad (103\text{b})$$

$$\left(1 + \frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}} \ll \frac{1}{\log m}. \qquad (103\text{c})$$

Here, the first condition (103a) naturally holds under the sample complexity $m \gg n\log^5 m$, whereas the second condition (103b) is true since $\beta_t \le \|\boldsymbol{x}^t\|_2 \lesssim \alpha_t\sqrt{n\log m}$ (cf. the induction hypothesis (40f)) and $m \gg n\log^4 m$. For the last condition (103c), observe that for $t \le T_0 = O(\log n)$,

$$\left(1 + \frac{1}{\log m}\right)^t = O(1),$$

which further implies

$$\left(1 + \frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}} \lesssim C_3\sqrt{\frac{n\log^5 m}{m}} \ll \frac{1}{\log m}$$

as long as the number of samples obeys $m \gg n\log^7 m$. This concludes the proof.

C.2 Proof of (41b)

Given the gradient update rule (3), the orthogonal component $\boldsymbol{x}_{\perp}^{t+1}$ can be decomposed as

$$\begin{aligned}
\boldsymbol{x}_{\perp}^{t+1} &= \boldsymbol{x}_{\perp}^t - \frac{\eta}{m}\sum_{i=1}^m\left[\left(\boldsymbol{a}_i^\top\boldsymbol{x}^t\right)^2 - \left(\boldsymbol{a}_i^\top\boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_{i,\perp}\boldsymbol{a}_i^\top\boldsymbol{x}^t \\
&= \boldsymbol{x}_{\perp}^t + \underbrace{\frac{\eta}{m}\sum_{i=1}^m\left(\boldsymbol{a}_i^\top\boldsymbol{x}^\natural\right)^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_i^\top\boldsymbol{x}^t}_{:=\boldsymbol{v}_1} - \underbrace{\frac{\eta}{m}\sum_{i=1}^m\left(\boldsymbol{a}_i^\top\boldsymbol{x}^t\right)^3\boldsymbol{a}_{i,\perp}}_{:=\boldsymbol{v}_2}. \qquad (104)
\end{aligned}$$

In what follows, we bound $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ in turn.

- We begin with $\boldsymbol{v}_1$. Using the identity $\boldsymbol{a}_i^\top \boldsymbol{x}^t = a_{i,1} x_\parallel^t + \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t$, one can further decompose $\boldsymbol{v}_1$ into the following two terms:

$$\begin{aligned}
\frac{1}{\eta}\boldsymbol{v}_1 &= x_\parallel^t \cdot \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2 a_{i,1}\boldsymbol{a}_{i,\perp} + \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2 \boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t \\
&= x_\parallel^t \boldsymbol{u} + \boldsymbol{U}\boldsymbol{x}_\perp^t,
\end{aligned}$$

where $\boldsymbol{U}$, $\boldsymbol{u}$ are as defined, respectively, in (89) and (95). Recall that we have shown that

$$\|\boldsymbol{u}\|_2 \lesssim \sqrt{\frac{n\log^3 m}{m}} \qquad \text{and} \qquad \|\boldsymbol{U} - \boldsymbol{I}_{n-1}\| \lesssim \sqrt{\frac{n\log^3 m}{m}}$$

hold with probability exceeding $1 - O\left(m^{-10}\right)$. Consequently, one has

$$\boldsymbol{v}_1 = \eta \boldsymbol{x}_\perp^t + \boldsymbol{r}_1, \tag{105}$$

where the residual term $\boldsymbol{r}_1$ obeys

$$\|\boldsymbol{r}_1\|_2 \lesssim \eta\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_\perp^t\right\|_2 + \eta\sqrt{\frac{n\log^3 m}{m}}\left|x_\parallel^t\right|. \tag{106}$$

- It remains to bound $\boldsymbol{v}_2$ in (104). To this end, we make note of the following fact

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{x}^t\right)^3 \boldsymbol{a}_{i,\perp} &= \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^3 \boldsymbol{a}_{i,\perp} + \left(x_\parallel^t\right)^3 \frac{1}{m}\sum_{i=1}^m a_{i,1}^3 \boldsymbol{a}_{i,\perp} \\
&\quad + \frac{3x_\parallel^t}{m}\sum_{i=1}^m a_{i,1}\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 \boldsymbol{a}_{i,\perp} + 3\left(x_\parallel^t\right)^2 \frac{1}{m}\sum_{i=1}^m a_{i,1}^2 \boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t \\
&= \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^3 \boldsymbol{a}_{i,\perp} + \frac{3x_\parallel^t}{m}\sum_{i=1}^m a_{i,1}\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 \boldsymbol{a}_{i,\perp} + \left(x_\parallel^t\right)^3 \boldsymbol{u} + 3\left(x_\parallel^t\right)^2 \boldsymbol{U}\boldsymbol{x}_\perp^t.
\end{aligned} \tag{107}$$

Applying Lemma 14 and using the incoherence condition (62b), we get

$$\left\|\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 \boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top - \left\|\boldsymbol{x}_\perp^t\right\|_2^2 \boldsymbol{I}_{n-1} - 2\boldsymbol{x}_\perp^t \boldsymbol{x}_\perp^{t\top}\right\| \lesssim \sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_\perp^t\right\|_2^2,$$

$$\left\|\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \begin{bmatrix} 0 \\ \boldsymbol{x}_\perp^t \end{bmatrix}\right)^2 \boldsymbol{a}_i\boldsymbol{a}_i^\top - \left\|\boldsymbol{x}_\perp^t\right\|_2^2 \boldsymbol{I}_n - 2\begin{bmatrix} 0 \\ \boldsymbol{x}_\perp^t \end{bmatrix}\begin{bmatrix} 0 \\ \boldsymbol{x}_\perp^t \end{bmatrix}^\top\right\| \lesssim \sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_\perp^t\right\|_2^2,$$

as long as $m \gg n\log^3 m$. These two together allow us to derive

$$\begin{aligned}
\left\|\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^3 \boldsymbol{a}_{i,\perp} - 3\left\|\boldsymbol{x}_\perp^t\right\|_2^2 \boldsymbol{x}_\perp^t\right\|_2 &= \left\|\left\{\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 \boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top - \left\|\boldsymbol{x}_\perp^t\right\|_2^2 \boldsymbol{I}_{n-1} - 2\boldsymbol{x}_\perp^t \boldsymbol{x}_\perp^{t\top}\right\}\boldsymbol{x}_\perp^t\right\|_2 \\
&\le \left\|\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 \boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top - \left\|\boldsymbol{x}_\perp^t\right\|_2^2 \boldsymbol{I}_{n-1} - 2\boldsymbol{x}_\perp^t \boldsymbol{x}_\perp^{t\top}\right\| \left\|\boldsymbol{x}_\perp^t\right\|_2 \\
&\lesssim \sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_\perp^t\right\|_2^3;
\end{aligned}$$

and

$$\left\|\frac{1}{m}\sum_{i=1}^m a_{i,1}\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 \boldsymbol{a}_{i,\perp}\right\|_2 \le \Bigg\|\underbrace{\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \begin{bmatrix} 0 \\ \boldsymbol{x}_\perp^t \end{bmatrix}\right)^2 \boldsymbol{a}_i\boldsymbol{a}_i^\top - \left\|\boldsymbol{x}_\perp^t\right\|_2^2 \boldsymbol{I}_n - 2\begin{bmatrix} 0 \\ \boldsymbol{x}_\perp^t \end{bmatrix}\begin{bmatrix} 0 \\ \boldsymbol{x}_\perp^t \end{bmatrix}^\top}_{:=\boldsymbol{A}}\Bigg\|$$

$$\lesssim \sqrt{\frac{n\log^3 m}{m}} \left\| \boldsymbol{x}_\perp^t \right\|_2^2,$$

where the second one follows since $\frac{1}{m}\sum_{i=1}^m a_{i,1}\big(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\big)^2 \boldsymbol{a}_{i,\perp}$ is the first column of $\boldsymbol{A}$ except for the first entry. Substitute the preceding bounds into (107) to arrive at

$$\begin{aligned}
&\left\| \frac{1}{m}\sum_{i=1}^m \big(\boldsymbol{a}_i^\top \boldsymbol{x}^t\big)^3 \boldsymbol{a}_{i,\perp} - 3\left\|\boldsymbol{x}_\perp^t\right\|_2^2 \boldsymbol{x}_\perp^t - 3\big(x_\parallel^t\big)^2 \boldsymbol{x}_\perp^t \right\|_2 \\
&\quad \le \left\| \frac{1}{m}\sum_{i=1}^m \big(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\big)^3 \boldsymbol{a}_{i,\perp} - 3\left\|\boldsymbol{x}_\perp^t\right\|_2^2 \boldsymbol{x}_\perp^t \right\|_2 + 3\left|x_\parallel^t\right| \left\| \frac{1}{m}\sum_{i=1}^m a_{i,1}\left(\boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^t\right)^2 \boldsymbol{a}_{i,\perp} \right\|_2 \\
&\qquad + \left\| \big(x_\parallel^t\big)^3 \boldsymbol{u} \right\|_2 + 3\big(x_\parallel^t\big)^2 \left\| \left(\boldsymbol{U} - \boldsymbol{I}_{n-1}\right) \boldsymbol{x}_\perp^t \right\|_2 \\
&\quad \lesssim \sqrt{\frac{n\log^3 m}{m}} \left(\left\|\boldsymbol{x}_\perp^t\right\|_2^3 + \left|x_\parallel^t\right| \left\|\boldsymbol{x}_\perp^t\right\|_2^2 + \left|x_\parallel^t\right|^3 + \left|x_\parallel^t\right|^2 \left\|\boldsymbol{x}_\perp^t\right\|_2 \right) \lesssim \sqrt{\frac{n\log^3 m}{m}} \left\|\boldsymbol{x}^t\right\|_2
\end{aligned}$$

with probability at least $1 - O(m^{-10})$. Here, the last relation holds owing to the norm condition (40e) and the fact that

$$\left\|\boldsymbol{x}_\perp^t\right\|_2^3 + \left|x_\parallel^t\right| \left\|\boldsymbol{x}_\perp^t\right\|_2^2 + \left|x_\parallel^t\right|^3 + \left|x_\parallel^t\right|^2 \left\|\boldsymbol{x}_\perp^t\right\|_2 \asymp \left\|\boldsymbol{x}^t\right\|_2^3 \lesssim \left\|\boldsymbol{x}^t\right\|_2.$$

This in turn tells us that

$$\boldsymbol{v}_2 = \frac{\eta}{m}\sum_{i=1}^m \big(\boldsymbol{a}_i^\top \boldsymbol{x}^t\big)^3 \boldsymbol{a}_{i,\perp} = 3\eta \left\|\boldsymbol{x}_\perp^t\right\|_2^2 \boldsymbol{x}_\perp^t + 3\eta\big(x_\parallel^t\big)^2 \boldsymbol{x}_\perp^t + \boldsymbol{r}_2 = 3\eta \left\|\boldsymbol{x}^t\right\|_2^2 \boldsymbol{x}_\perp^t + \boldsymbol{r}_2,$$

where the residual term $\boldsymbol{r}_2$ is bounded by

$$\left\|\boldsymbol{r}_2\right\|_2 \lesssim \eta\sqrt{\frac{n\log^3 m}{m}} \left\|\boldsymbol{x}^t\right\|_2.$$

- Putting the above estimates on $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ together, we conclude that

$$\boldsymbol{x}_\perp^{t+1} = \boldsymbol{x}_\perp^t + \boldsymbol{v}_1 - \boldsymbol{v}_2 = \left\{1 + \eta\left(1 - 3\left\|\boldsymbol{x}^t\right\|_2^2\right)\right\} \boldsymbol{x}_\perp^t + \boldsymbol{r}_3,$$

where $\boldsymbol{r}_3 = \boldsymbol{r}_1 - \boldsymbol{r}_2$ satisfies

$$\left\|\boldsymbol{r}_3\right\|_2 \lesssim \eta\sqrt{\frac{n\log^3 m}{m}} \left\|\boldsymbol{x}^t\right\|_2.$$

Plug in the definitions of α_t and β_t to realize that

$$\begin{aligned}
\beta_{t+1} &= \left\{1 + \eta\left(1 - 3\left\|\boldsymbol{x}^t\right\|_2^2\right)\right\} \beta_t + O\left(\eta\sqrt{\frac{n\log^3 m}{m}} \left(\alpha_t + \beta_t\right)\right) \\
&= \left\{1 + \eta\left(1 - 3\left\|\boldsymbol{x}^t\right\|_2^2\right) + \eta\rho_t\right\} \beta_t,
\end{aligned}$$

for some $|\rho_t| \ll \frac{1}{\log m}$, with the proviso that $m \gg n\log^5 m$ and

$$\sqrt{\frac{n\log^3 m}{m}}\alpha_t \ll \frac{1}{\log m}\beta_t. \tag{108}$$

The last condition holds true since

$$\sqrt{\frac{n\log^3 m}{m}}\alpha_t \lesssim \sqrt{\frac{n\log^3 m}{m}} \frac{1}{\log^5 m} \ll \frac{1}{\log m} \ll \frac{1}{\log m}\beta_t,$$

where we have used the assumption $\alpha_t \lesssim \frac{1}{\log^5 m}$ (see definition of T_0), the sample size condition $m \gg n\log^{11} m$ and the induction hypothesis $\beta_t \ge c_5$ (see (40e)). This finishes the proof.

D Proof of Lemma 4

It follows from the gradient update rules (3) and (29) that

$$\begin{aligned}
\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)} &= \boldsymbol{x}^{t}-\eta\nabla f\left(\boldsymbol{x}^{t}\right)-\left(\boldsymbol{x}^{t,(l)}-\eta\nabla f^{(l)}\big(\boldsymbol{x}^{t,(l)}\big)\right)\\
&= \boldsymbol{x}^{t}-\eta\nabla f\left(\boldsymbol{x}^{t}\right)-\left(\boldsymbol{x}^{t,(l)}-\eta\nabla f\big(\boldsymbol{x}^{t,(l)}\big)\right)+\eta\nabla f^{(l)}\big(\boldsymbol{x}^{t,(l)}\big)-\eta\nabla f\big(\boldsymbol{x}^{t,(l)}\big)\\
&= \left[\boldsymbol{I}_n-\eta\int_0^1\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right)\mathrm{d}\tau\right]\left(\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}\right)-\frac{\eta}{m}\left[\left(\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t,(l)}\right)^2-\left(\boldsymbol{a}_l^{\top}\boldsymbol{x}^{\natural}\right)^2\right]\boldsymbol{a}_l\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t,(l)}, \quad (109)
\end{aligned}$$

where we denote $\boldsymbol{x}\left(\tau\right):=\boldsymbol{x}^{t}+\tau\left(\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t}\right)$. Here, the last identity is due to the fundamental theorem of calculus [Lan93, Chapter XIII, Theorem 4.2].

- Controlling the first term in (109) requires exploring the properties of the Hessian $\nabla^2 f\left(\boldsymbol{x}\right)$. Since $\boldsymbol{x}\left(\tau\right)$ lies between $\boldsymbol{x}^{t}$ and $\boldsymbol{x}^{t,(l)}$ for any $0\leq\tau\leq1$, we have the following two consequences

$$\left\|\boldsymbol{x}_{\perp}\left(\tau\right)\right\|_2\leq\left\|\boldsymbol{x}\left(\tau\right)\right\|_2\leq 2C_5 \qquad\text{and}\qquad \max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right|\lesssim\sqrt{\log m}\lesssim\sqrt{\log m}\left\|\boldsymbol{x}\left(\tau\right)\right\|_2. \quad (110)$$

To see the left statement in (110), one has

$$\left\|\boldsymbol{x}\left(\tau\right)\right\|_2\leq\max\{\|\boldsymbol{x}^{t}\|_2,\|\boldsymbol{x}^{t,(l)}\|_2\}\leq 2C_5,$$

where the last inequality follows from (40e) and (61a). Moreover, for the right statement in (110), one can see

$$\begin{aligned}
\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right| &= \max_{1\leq i\leq m}\left|\left(1-\tau\right)\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t}+\tau\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,(l)}\right|\\
&\leq \max_{1\leq i\leq m}\left|\left(1-\tau\right)\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t}+\tau\boldsymbol{a}_i^{\top}\left(\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,(i)}\right)+\tau\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,(i)}\right|\\
&\leq \left(1-\tau\right)\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t}\right|+\tau\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\left(\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,(i)}\right)\right|+\tau\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,(i)}\right|.
\end{aligned}$$

In view of (62a), we have

$$\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t}\right|\lesssim\log m.$$

Furthermore, due to the independence between $\boldsymbol{a}_i$ and $\boldsymbol{x}^{t,(i)}$, one can apply standard Gaussian concentration inequalities to show that with high probability

$$\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,(i)}\right|\lesssim\sqrt{\log m}.$$

We are left with the middle term, which can be controlled using Cauchy-Schwarz as follows:

$$\begin{aligned}
\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^{\top}\left(\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,(i)}\right)\right| &\leq \max_i\|\boldsymbol{a}_i\|_2\max_{1\leq i\leq m}\left\|\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,(i)}\right\|_2\\
&\overset{\text{(i)}}{\lesssim}\sqrt{n}\cdot\max_{1\leq i\leq m}\left(\left\|\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t}\right\|_2+\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(i)}\right\|_2\right)\\
&\lesssim\sqrt{n}\cdot\max_{1\leq i\leq m}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(i)}\right\|_2\\
&\overset{\text{(ii)}}{\lesssim}\sqrt{n}\cdot\beta_t\left(1+\frac{1}{\log m}\right)^{t}C_1\eta\frac{\sqrt{n\log^5 m}}{m}\\
&\overset{\text{(iii)}}{\lesssim}\sqrt{\log m}.
\end{aligned}$$

Here, the inequality (i) arises from the concentration of norm of Gaussian vectors and the triangle inequality; the relation (ii) holds because of the induction hypothesis (40a) and the last inequality (iii) holds true under the sample size condition $m\gg n\log^2 m$.

In addition, combining (40e) and (63) leads to

$$\left\|\boldsymbol{x}_{\perp}\left(\tau\right)\right\|_{2} \geq \left\|\boldsymbol{x}_{\perp}^{t}\right\|_{2} - \left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_{2} \geq c_{5} - \log^{-1} m \geq c_{5}/4. \tag{111}$$

Armed with these bounds, we can readily apply Lemma 15 to obtain

$$\begin{aligned}
&\left\|\boldsymbol{I}_{n} - \eta\nabla^{2} f\left(\boldsymbol{x}\left(\tau\right)\right) - \left\{\left(1 - 3\eta\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2} + \eta\right)\boldsymbol{I}_{n} + 2\eta\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top} - 6\eta\boldsymbol{x}\left(\tau\right)\boldsymbol{x}\left(\tau\right)^{\top}\right\}\right\| \\
&\quad \lesssim \eta\sqrt{\frac{n\log^{3} m}{m}}\max\left\{\left\|\boldsymbol{x}(\tau)\right\|_{2}^{2}, 1\right\} \lesssim \eta\sqrt{\frac{n\log^{3} m}{m}}.
\end{aligned}$$

This further allows one to derive

$$\begin{aligned}
&\left\|\left\{\boldsymbol{I}_{n} - \eta\nabla^{2} f\left(\boldsymbol{x}\left(\tau\right)\right)\right\}\left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right)\right\|_{2} \\
&\leq \left\|\left\{\left(1 - 3\eta\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2} + \eta\right)\boldsymbol{I}_{n} + 2\eta\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top} - 6\eta\boldsymbol{x}\left(\tau\right)\boldsymbol{x}\left(\tau\right)^{\top}\right\}\left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right)\right\|_{2} + O\left(\eta\sqrt{\frac{n\log^{3} m}{m}}\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_{2}\right).
\end{aligned}$$

Moreover, we can apply the triangle inequality to get

$$\begin{aligned}
&\left\|\left\{\left(1 - 3\eta\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2} + \eta\right)\boldsymbol{I}_{n} + 2\eta\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top} - 6\eta\boldsymbol{x}\left(\tau\right)\boldsymbol{x}\left(\tau\right)^{\top}\right\}\left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right)\right\|_{2} \\
&\quad \leq \left\|\left\{\left(1 - 3\eta\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2} + \eta\right)\boldsymbol{I}_{n} - 6\eta\boldsymbol{x}\left(\tau\right)\boldsymbol{x}\left(\tau\right)^{\top}\right\}\left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right)\right\|_{2} + \left\|2\eta\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top}\left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right)\right\|_{2} \\
&\quad \overset{\text{(i)}}{=} \left\|\left\{\left(1 - 3\eta\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2} + \eta\right)\boldsymbol{I}_{n} - 6\eta\boldsymbol{x}\left(\tau\right)\boldsymbol{x}\left(\tau\right)^{\top}\right\}\left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right)\right\|_{2} + 2\eta\left|x_{\|}^{t} - x_{\|}^{t,(l)}\right| \\
&\quad \overset{\text{(ii)}}{\leq} \left(1 - 3\eta\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2} + \eta\right)\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_{2} + 2\eta\left|x_{\|}^{t} - x_{\|}^{t,(l)}\right|,
\end{aligned}$$

where (i) holds since $\boldsymbol{x}^{\natural\top}\left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right) = x_{\|}^{t} - x_{\|}^{t,(l)}$ (recall that $\boldsymbol{x}^{\natural} = \boldsymbol{e}_{1}$) and (ii) follows from the fact that

$$\left(1 - 3\eta\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2} + \eta\right)\boldsymbol{I}_{n} - 6\eta\boldsymbol{x}\left(\tau\right)\boldsymbol{x}\left(\tau\right)^{\top} \succeq \boldsymbol{0},$$

as long as $\eta \leq 1/\left(18C_{5}\right)$. This further reveals

$$\begin{aligned}
&\left\|\left\{\boldsymbol{I}_{n} - \eta\nabla^{2} f\left(\boldsymbol{x}\left(\tau\right)\right)\right\}\left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right)\right\|_{2} \\
&\quad \leq \left\{1 + \eta\left(1 - 3\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2}\right) + O\left(\eta\sqrt{\frac{n\log^{3} m}{m}}\right)\right\}\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_{2} + 2\eta\left|x_{\|}^{t} - x_{\|}^{t,(l)}\right| \\
&\quad \overset{\text{(i)}}{\leq} \left\{1 + \eta\left(1 - 3\left\|\boldsymbol{x}^{t}\right\|_{2}^{2}\right) + O\left(\eta\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_{2}\right) + O\left(\eta\sqrt{\frac{n\log^{3} m}{m}}\right)\right\}\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_{2} + 2\eta\left|x_{\|}^{t} - x_{\|}^{t,(l)}\right| \\
&\quad \overset{\text{(ii)}}{\leq} \left\{1 + \eta\left(1 - 3\left\|\boldsymbol{x}^{t}\right\|_{2}^{2}\right) + \eta\phi_{1}\right\}\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_{2} + 2\eta\left|x_{\|}^{t} - x_{\|}^{t,(l)}\right|,
\end{aligned} \tag{112}$$

for some $|\phi_{1}| \ll \frac{1}{\log m}$, where (i) holds since for every $0 \leq \tau \leq 1$

$$\begin{aligned}
\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2} &\geq \left\|\boldsymbol{x}^{t}\right\|_{2}^{2} - \left|\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2}^{2} - \left\|\boldsymbol{x}^{t}\right\|_{2}^{2}\right| \\
&\geq \left\|\boldsymbol{x}^{t}\right\|_{2}^{2} - \left\|\boldsymbol{x}\left(\tau\right) - \boldsymbol{x}^{t}\right\|_{2}\left(\left\|\boldsymbol{x}\left(\tau\right)\right\|_{2} + \left\|\boldsymbol{x}^{t}\right\|_{2}\right) \\
&\geq \left\|\boldsymbol{x}^{t}\right\|_{2}^{2} - O\left(\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_{2}\right),
\end{aligned} \tag{113}$$

and (ii) comes from the fact (63a) and the sample complexity assumption $m \gg n\log^{5} m$.

- We then move on to the second term of (109). Observing that $\boldsymbol{x}^{t,(l)}$ is statistically independent of $\boldsymbol{a}_l$, we have

$$\begin{aligned}\left\|\frac{1}{m}\left[\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2-\left(\boldsymbol{a}_l^\top\boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right\|_2 &\le \frac{1}{m}\left[\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2+\left(\boldsymbol{a}_l^\top\boldsymbol{x}^\natural\right)^2\right]\left|\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right|\|\boldsymbol{a}_l\|_2\\ &\lesssim \frac{1}{m}\cdot\log m\cdot\sqrt{\log m}\left\|\boldsymbol{x}^{t,(l)}\right\|_2\cdot\sqrt{n}\\ &\asymp \frac{\sqrt{n\log^3 m}}{m}\left\|\boldsymbol{x}^{t,(l)}\right\|_2, \end{aligned}\tag{114}$$

where the second inequality makes use of the facts (56), (57) and the standard concentration results

$$\left|\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right|\lesssim\sqrt{\log m}\|\boldsymbol{x}^{t,(l)}\|_2\lesssim\sqrt{\log m}.$$

- Combine the previous two bounds (112) and (114) to reach

$$\begin{aligned}&\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\|_2\\ &\le\left\|\left\{\boldsymbol{I}-\eta\int_0^1\nabla^2 f(\boldsymbol{x}(\tau))\,\mathrm{d}\tau\right\}\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right)\right\|_2+\eta\left\|\frac{1}{m}\left[\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2-\left(\boldsymbol{a}_l^\top\boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right\|_2\\ &\le\left\{1+\eta\left(1-3\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_1\right\}\|\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\|_2+2\eta\left|x_\|^t-x_\|^{t,(l)}\right|+O\left(\frac{\eta\sqrt{n\log^3 m}}{m}\left\|\boldsymbol{x}^{t,(l)}\right\|_2\right)\\ &\le\left\{1+\eta\left(1-3\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_1\right\}\|\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\|_2+O\left(\frac{\eta\sqrt{n\log^3 m}}{m}\right)\|\boldsymbol{x}^t\|_2+2\eta|x_\|^t-x_\|^{t,(l)}|.\end{aligned}$$

Here the last relation holds because of the triangle inequality

$$\|\boldsymbol{x}^{t,(l)}\|_2\le\|\boldsymbol{x}^t\|_2+\|\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\|_2$$

and the fact that $\frac{\sqrt{n\log^3 m}}{m}\ll\frac{1}{\log m}$.

In view of the inductive hypotheses (40), one has

$$\begin{aligned}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\|_2 &\overset{\text{(i)}}{\le}\left\{1+\eta\left(1-3\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_1\right\}\beta_t\left(1+\frac{1}{\log m}\right)^t C_1\frac{\sqrt{n\log^5 m}}{m}\\ &\quad+O\left(\frac{\eta\sqrt{n\log^3 m}}{m}\right)(\alpha_t+\beta_t)+2\eta\alpha_t\left(1+\frac{1}{\log m}\right)^t C_2\frac{\sqrt{n\log^{12} m}}{m}\\ &\overset{\text{(ii)}}{\le}\left\{1+\eta\left(1-3\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_2\right\}\beta_t\left(1+\frac{1}{\log m}\right)^t C_1\frac{\sqrt{n\log^5 m}}{m}\\ &\overset{\text{(iii)}}{\le}\beta_{t+1}\left(1+\frac{1}{\log m}\right)^{t+1}C_1\frac{\sqrt{n\log^5 m}}{m},\end{aligned}$$

for some $|\phi_2|\ll\frac{1}{\log m}$, where the inequality (i) uses $\|\boldsymbol{x}^t\|_2\le|x_\|^t|+\|\boldsymbol{x}_\perp^t\|_2=\alpha_t+\beta_t$, the inequality (ii) holds true as long as

$$\frac{\sqrt{n\log^3 m}}{m}(\alpha_t+\beta_t)\ll\frac{1}{\log m}\beta_t\left(1+\frac{1}{\log m}\right)^t C_1\frac{\sqrt{n\log^5 m}}{m},\tag{115a}$$

$$\alpha_t C_2\frac{\sqrt{n\log^{12} m}}{m}\ll\frac{1}{\log m}\beta_t C_1\frac{\sqrt{n\log^5 m}}{m}.\tag{115b}$$

Here, the first condition (115a) comes from the fact that for $t<T_0$,

$$\frac{\sqrt{n\log^3 m}}{m}(\alpha_t+\beta_t)\asymp\frac{\sqrt{n\log^3 m}}{m}\beta_t\ll C_1\beta_t\frac{\sqrt{n\log^3 m}}{m},$$

as long as $C_1 > 0$ is sufficiently large. The other one (115b) is valid owing to the assumption of Phase I $\alpha_t \ll 1/\log^5 m$. Regarding the inequality (iii) above, it is easy to check that for some $|\phi_3| \ll \frac{1}{\log m}$,

$$\begin{aligned}\left\{1+\eta\left(1-3\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_2\right\}\beta_t &= \left\{\frac{\beta_{t+1}}{\beta_t}+\eta\phi_3\right\}\beta_t \\ &= \left\{\frac{\beta_{t+1}}{\beta_t}+\eta O\left(\frac{\beta_{t+1}}{\beta_t}\phi_3\right)\right\}\beta_t \\ &\le \beta_{t+1}\left(1+\frac{1}{\log m}\right), \end{aligned} \tag{116}$$

where the second equality holds since $\frac{\beta_{t+1}}{\beta_t} \asymp 1$ in Phase I.

The proof is completed by applying the union bound over all $1 \le l \le m$.

E Proof of Lemma 5

Use (109) once again to deduce

$$\begin{aligned} x_{\|}^{t+1} - x_{\|}^{t+1,(l)} &= \boldsymbol{e}_1^\top\left(\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\right) \\ &= \boldsymbol{e}_1^\top\left\{\boldsymbol{I}_n - \eta\int_0^1 \nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right)\mathrm{d}\tau\right\}\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right) - \frac{\eta}{m}\left[\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2-\left(\boldsymbol{a}_l^\top\boldsymbol{x}^\natural\right)^2\right]\boldsymbol{e}_1^\top\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)} \\ &= \left[x_{\|}^t - x_{\|}^{t,(l)} - \eta\int_0^1 \boldsymbol{e}_1^\top\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right)\mathrm{d}\tau\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right)\right] - \frac{\eta}{m}\left[\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2-\left(\boldsymbol{a}_l^\top\boldsymbol{x}^\natural\right)^2\right]a_{l,1}\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}, \end{aligned} \tag{117}$$

where we recall that $\boldsymbol{x}\left(\tau\right) := \boldsymbol{x}^t + \tau\left(\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^t\right)$.

We begin by controlling the second term of (117). Applying similar arguments as in (114) yields

$$\left|\frac{1}{m}\left[\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2-\left(\boldsymbol{a}_l^\top\boldsymbol{x}^\natural\right)^2\right]a_{l,1}\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right| \lesssim \frac{\log^2 m}{m}\left\|\boldsymbol{x}^{t,(l)}\right\|_2$$

with probability at least $1-O\left(m^{-10}\right)$.

Regarding the first term in (117), one can use the decomposition

$$\boldsymbol{a}_i^\top\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right) = a_{i,1}\left(x_{\|}^t - x_{\|}^{t,(l)}\right) + \boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)}\right)$$

to obtain that

$$\begin{aligned}\boldsymbol{e}_1^\top\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right)\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right) &= \frac{1}{m}\sum_{i=1}^m\left[3\left(\boldsymbol{a}_i^\top\boldsymbol{x}\left(\tau\right)\right)^2-\left(\boldsymbol{a}_i^\top\boldsymbol{x}^\natural\right)^2\right]a_{i,1}\boldsymbol{a}_i^\top\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right) \\ &= \underbrace{\frac{1}{m}\sum_{i=1}^m\left[3\left(\boldsymbol{a}_i^\top\boldsymbol{x}\left(\tau\right)\right)^2-\left(\boldsymbol{a}_i^\top\boldsymbol{x}^\natural\right)^2\right]a_{i,1}^2\left(x_{\|}^t - x_{\|}^{t,(l)}\right)}_{:=\omega_1(\tau)} \\ &\quad + \underbrace{\frac{1}{m}\sum_{i=1}^m\left[3\left(\boldsymbol{a}_i^\top\boldsymbol{x}\left(\tau\right)\right)^2-\left(\boldsymbol{a}_i^\top\boldsymbol{x}^\natural\right)^2\right]a_{i,1}\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)}\right)}_{:=\omega_2(\tau)}. \end{aligned}$$

In the sequel, we shall bound $\omega_1\left(\tau\right)$ and $\omega_2\left(\tau\right)$ separately.

- For $\omega_1\left(\tau\right)$, Lemma 14 together with the facts (110) tells us that

$$\left|\frac{1}{m}\sum_{i=1}^m\left[3\left(\boldsymbol{a}_i^\top\boldsymbol{x}\left(\tau\right)\right)^2-\left(\boldsymbol{a}_i^\top\boldsymbol{x}^\natural\right)^2\right]a_{i,1}^2 - \left[3\left\|\boldsymbol{x}\left(\tau\right)\right\|_2^2+6\left|x_{\|}\left(\tau\right)\right|^2-3\right]\right|$$

$$\lesssim \sqrt{\frac{n\log^3 m}{m}} \max\left\{\|\boldsymbol{x}(\tau)\|_2^2, 1\right\} \lesssim \sqrt{\frac{n\log^3 m}{m}},$$

which further implies that

$$\omega_1(\tau) = \left(3\|\boldsymbol{x}(\tau)\|_2^2 + 6\left|x_\parallel(\tau)\right|^2 - 3\right)\left(x_\parallel^t - x_\parallel^{t,(l)}\right) + r_1$$

with the residual term r_1 obeying

$$|r_1| = O\left(\sqrt{\frac{n\log^3 m}{m}}\left|x_\parallel^t - x_\parallel^{t,(l)}\right|\right).$$

- We proceed to bound $\omega_2(\tau)$. Decompose $w_2(\tau)$ into the following:

$$\omega_2(\tau) = \underbrace{\frac{3}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau)\right)^2 a_{i,1}\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)}\right)}_{:=\omega_3(\tau)} - \underbrace{\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2 a_{i,1}\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)}\right)}_{:=\omega_4}.$$

  – The term ω_4 is relatively simple to control. Recognizing $\left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2 = a_{i,1}^2$ and $a_{i,1} = \xi_i|a_{i,1}|$, one has

$$\omega_4 = \frac{1}{m}\sum_{i=1}^m \xi_i|a_{i,1}|^3\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right) + \frac{1}{m}\sum_{i=1}^m \xi_i|a_{i,1}|^3\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right).$$

  In view of the independence between ξ_i and $|a_{i,1}|^3\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)$, one can thus invoke the Bernstein inequality (see Lemma 11) to obtain

$$\left|\frac{1}{m}\sum_{i=1}^m \xi_i|a_{i,1}|^3\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right| \lesssim \frac{1}{m}\left(\sqrt{V_1\log m} + B_1\log m\right) \tag{118}$$

  with probability at least $1 - O\left(m^{-10}\right)$, where

$$V_1 := \sum_{i=1}^m |a_{i,1}|^6\left|\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right|^2 \qquad \text{and} \qquad B_1 := \max_{1\le i\le m}|a_{i,1}|^3\left|\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right|.$$

  Regarding V_1, one can combine the fact (56) and Lemma 14 to reach

$$\begin{aligned}\frac{1}{m}V_1 &\lesssim \log^2 m\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)^\top\left(\frac{1}{m}\sum_{i=1}^m |a_{i,1}|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\right)\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\\ &\lesssim \log^2 m\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2^2.\end{aligned}$$

  For B_1, it is easy to check from (56) and (57) that

$$B_1 \lesssim \sqrt{n\log^3 m}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2.$$

  The previous two bounds taken collectively yield

$$\begin{aligned}\left|\frac{1}{m}\sum_{i=1}^m \xi_i|a_{i,1}|^3\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right| &\lesssim \sqrt{\frac{\log^3 m}{m}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2^2} + \frac{\sqrt{n\log^5 m}}{m}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2\\ &\lesssim \sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2,\end{aligned} \tag{119}$$

as long as $m \gtrsim n \log^2 m$. The second term in ω_4 can be simply controlled by the Cauchy-Schwarz inequality and Lemma 14. Specifically, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i\left|a_{i,1}\right|^3\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)\right| \\
&\leq\left\|\frac{1}{m}\sum_{i=1}^{m}\xi_i\left|a_{i,1}\right|^3\boldsymbol{a}_{i,\perp}^{\top}\right\|_2\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\lesssim\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2, \qquad (120)
\end{aligned}$$

where the second relation holds due to Lemma 14. Take the preceding two bounds (119) and (120) collectively to conclude that

$$\begin{aligned}
\left|\omega_4\right| &\lesssim\sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2+\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\lesssim\sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2+\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2,
\end{aligned}$$

where the second line follows from the triangle inequality

$$\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2\leq\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2+\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2$$

and the fact that $\sqrt{\frac{\log^3 m}{m}}\leq\sqrt{\frac{n\log^3 m}{m}}$.

– It remains to bound $\omega_3(\tau)$. To this end, one can decompose

$$\begin{aligned}
\omega_3(\tau)=&\underbrace{\frac{3}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}(\tau)\right)^2-\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}(\tau)\right)^2\right]a_{i,1}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right)}_{:=\theta_1(\tau)} \\
&+\underbrace{\frac{3}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}(\tau)\right)^2-\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{\mathrm{sgn}}(\tau)\right)^2\right]a_{i,1}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right)}_{:=\theta_2(\tau)} \\
&+\underbrace{\frac{3}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{\mathrm{sgn}}(\tau)\right)^2a_{i,1}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)}_{:=\theta_3(\tau)} \\
&+\underbrace{\frac{3}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{\mathrm{sgn}}(\tau)\right)^2a_{i,1}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)}_{:=\theta_4(\tau)},
\end{aligned}$$

where we denote $\boldsymbol{x}^{\mathrm{sgn}}(\tau)=\boldsymbol{x}^{t,\mathrm{sgn}}+\tau\left(\boldsymbol{x}^{t,\mathrm{sgn},(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}\right)$. A direct consequence of (61) and (62) is that

$$\left|\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{\mathrm{sgn}}(\tau)\right|\lesssim\sqrt{\log m}. \qquad (121)$$

Recalling that $\xi_i=\mathrm{sgn}\left(a_{i,1}\right)$ and $\xi_i^{\mathrm{sgn}}=\mathrm{sgn}\left(a_{i,1}^{\mathrm{sgn}}\right)$, one has

$$\begin{aligned}
\boldsymbol{a}_i^{\top}\boldsymbol{x}(\tau)-\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}(\tau)&=\left(\xi_i-\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|x_{\|}(\tau), \\
\boldsymbol{a}_i^{\top}\boldsymbol{x}(\tau)+\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}(\tau)&=\left(\xi_i+\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|x_{\|}(\tau)+2\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}(\tau),
\end{aligned}$$

which implies that

$$\begin{aligned}\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2-\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}\left(\tau\right)\right)^2&=\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)-\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}\left(\tau\right)\right)\cdot\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)+\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}\left(\tau\right)\right)\\&=\left(\xi_i-\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|x_{\parallel}\left(\tau\right)\left\{\left(\xi_i+\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|x_{\parallel}\left(\tau\right)+2\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}(\tau)\right\}\\&=2\left(\xi_i-\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|x_{\parallel}\left(\tau\right)\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}\left(\tau\right)\end{aligned}\tag{122}$$

owing to the identity $\left(\xi_i-\xi_i^{\mathrm{sgn}}\right)\left(\xi_i+\xi_i^{\mathrm{sgn}}\right)=\xi_i^2-\left(\xi_i^{\mathrm{sgn}}\right)^2=0$. In light of (122), we have

$$\begin{aligned}\theta_1\left(\tau\right)&=\frac{6}{m}\sum_{i=1}^m\left(\xi_i-\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|x_{\parallel}\left(\tau\right)\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}\left(\tau\right)a_{i,1}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right)\\&=6x_{\parallel}\left(\tau\right)\cdot\boldsymbol{x}_{\perp}^{\top}\left(\tau\right)\left[\frac{1}{m}\sum_{i=1}^m\left(1-\xi_i\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}\right]\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right).\end{aligned}$$

First note that

$$\left\|\frac{1}{m}\sum_{i=1}^m\left(1-\xi_i\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}\right\|\leq2\left\|\frac{1}{m}\sum_{i=1}^m\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}\right\|\lesssim1,\tag{123}$$

where the last relation holds due to Lemma 14. This results in the following upper bound on $\theta_1\left(\tau\right)$

$$\left|\theta_1\left(\tau\right)\right|\lesssim\left|x_{\parallel}\left(\tau\right)\right|\left\|\boldsymbol{x}_{\perp}\left(\tau\right)\right\|_2\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2\lesssim\left|x_{\parallel}\left(\tau\right)\right|\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2,$$

where we have used the fact that $\left\|\boldsymbol{x}_{\perp}\left(\tau\right)\right\|_2\lesssim1$ (see (110)). Regarding $\theta_2\left(\tau\right)$, one obtains

$$\theta_2\left(\tau\right)=\frac{3}{m}\sum_{i=1}^m\left[\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}\left(\tau\right)-\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right)\right]\left[\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}\left(\tau\right)+\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right)\right]a_{i,1}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right).$$

Apply the Cauchy-Schwarz inequality to reach

$$\begin{aligned}\left|\theta_2\left(\tau\right)\right|&\lesssim\sqrt{\frac{1}{m}\sum_{i=1}^m\left[\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}\left(\tau\right)-\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right)\right]^2\left[\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}\left(\tau\right)+\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right)\right]^2}\sqrt{\frac{1}{m}\sum_{i=1}^m\left|a_{i,1}\right|^2\left[\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right)\right]^2}\\&\lesssim\sqrt{\frac{1}{m}\sum_{i=1}^m\left[\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}\left(\tau\right)-\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right)\right]^2\log m}\cdot\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2\\&\lesssim\sqrt{\log m}\left\|\boldsymbol{x}\left(\tau\right)-\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right\|_2\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2.\end{aligned}$$

Here the second relation comes from Lemma 14 and the fact that

$$\left|\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}\left(\tau\right)+\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right)\right|\lesssim\sqrt{\log m}.$$

When it comes to $\theta_3\left(\tau\right)$, we need to exploit the independence between

$$\{\xi_i\}\quad\text{and}\quad\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right)^2\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right).$$

Similar to (118), one can obtain

$$\left|\theta_3\left(\tau\right)\right|\lesssim\frac{1}{m}\left(\sqrt{V_2\log m}+B_2\log m\right)$$

with probability at least $1-O\left(m^{-10}\right)$, where

$$V_2:=\sum_{i=1}^m\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right)^4\left|a_{i,1}\right|^2\left|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)\right|^2$$

$$B_2 := \max_{1\le i\le m} \left(\boldsymbol{a}_i^{\mathrm{sgn}\top} \boldsymbol{x}^{\mathrm{sgn}}(\tau)\right)^2 |a_{i,1}| \left|\boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right|.$$

It is easy to see from Lemma 14, (121), (56) and (57) that

$$V_2 \lesssim m\log^2 m \left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2^2 \qquad \text{and} \qquad B_2 \lesssim \sqrt{n\log^3 m}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2,$$

which implies

$$|\theta_3(\tau)| \lesssim \left(\sqrt{\frac{\log^3 m}{m}} + \frac{\sqrt{n\log^5 m}}{m}\right)\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2 \asymp \sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2$$

with the proviso that $m \gtrsim n\log^2 m$. We are left with $\theta_4(\tau)$. Invoking Cauchy-Schwarz inequality,

$$\begin{aligned}
|\theta_4(\tau)| &\lesssim \sqrt{\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{\mathrm{sgn}}(\tau)\right)^4}\sqrt{\frac{1}{m}\sum_{i=1}^m |a_{i,1}|^2\left[\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right]^2} \\
&\lesssim \sqrt{\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{\mathrm{sgn}}(\tau)\right)^2 \log m} \cdot \left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\lesssim \sqrt{\log m}\left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2,
\end{aligned}$$

where we have used the fact that $\left|\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{\mathrm{sgn}}(\tau)\right| \lesssim \sqrt{\log m}$. In summary, we have obtained

$$\begin{aligned}
|\omega_3(\tau)| &\lesssim \left\{\left|x_\parallel(\tau)\right| + \sqrt{\log m}\left\|\boldsymbol{x}(\tau) - \boldsymbol{x}^{\mathrm{sgn}}(\tau)\right\|_2\right\}\left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)}\right\|_2 \\
&\quad + \sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2 + \sqrt{\log m}\left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\lesssim \left\{\left|x_\parallel(\tau)\right| + \sqrt{\log m}\left\|\boldsymbol{x}(\tau) - \boldsymbol{x}^{\mathrm{sgn}}(\tau)\right\|_2 + \sqrt{\frac{\log^3 m}{m}}\right\}\left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)}\right\|_2 \\
&\quad + \sqrt{\log m}\left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2,
\end{aligned}$$

where the last inequality utilizes the triangle inequality

$$\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2 \le \left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)}\right\|_2 + \left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2$$

and the fact that $\sqrt{\frac{\log^3 m}{m}} \le \sqrt{\log m}$. This together with the bound for $\omega_4(\tau)$ gives

$$\begin{aligned}
|\omega_2(\tau)| &\le |\omega_3(\tau)| + |\omega_4(\tau)| \\
&\lesssim \left\{\left|x_\parallel(\tau)\right| + \sqrt{\log m}\left\|\boldsymbol{x}(\tau) - \boldsymbol{x}^{\mathrm{sgn}}(\tau)\right\|_2 + \sqrt{\frac{\log^3 m}{m}}\right\}\left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)}\right\|_2 \\
&\quad + \sqrt{\log m}\left\|\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2,
\end{aligned}$$

as long as $m \gg n\log^2 m$.

- Combine the bounds to arrive at

$$x_\parallel^{t+1} - x_\parallel^{t+1,(l)} = \left\{1 + 3\eta\left(1 - \int_0^1 \|\boldsymbol{x}(\tau)\|_2^2 \,\mathrm{d}\tau\right) + \eta\cdot O\left(\left|x_\parallel(\tau)\right|^2 + \sqrt{\frac{n\log^3 m}{m}}\right)\right\}\left(x_\parallel^t - x_\parallel^{t,(l)}\right)$$

$$+O\left(\eta\frac{\log^2 m}{m}\left\|\boldsymbol{x}^{t,(l)}\right\|_2\right)+O\left(\eta\sqrt{\log m}\left\|\boldsymbol{x}_\perp^t-\boldsymbol{x}_\perp^{t,(l)}-\boldsymbol{x}_\perp^{t,\mathrm{sgn}}+\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2\right)$$
$$+O\left(\eta\sup_{0\le\tau\le1}\left\{\left|x_\|(\tau)\right|+\sqrt{\log m}\left\|\boldsymbol{x}(\tau)-\boldsymbol{x}^{\mathrm{sgn}}(\tau)\right\|_2+\sqrt{\frac{\log^3 m}{m}}\right\}\left\|\boldsymbol{x}_\perp^t-\boldsymbol{x}_\perp^{t,(l)}\right\|_2\right).$$

To simplify the above bound, notice that for the last term, for any $t<T_0\lesssim\log n$ and $0\le\tau\le1$, one has

$$\left|x_\|(\tau)\right|\le\left|x_\|^t\right|+\left|x_\|^{t,(l)}-x_\|^t\right|\le\alpha_t+\alpha_t\left(1+\frac{1}{\log m}\right)^t C_2\frac{\sqrt{n\log^{12}m}}{m}\lesssim\alpha_t,$$

as long as $m\gg\sqrt{n\log^{12}m}$. Similarly, one can show that

$$\sqrt{\log m}\left\|\boldsymbol{x}(\tau)-\boldsymbol{x}^{\mathrm{sgn}}(\tau)\right\|_2\le\sqrt{\log m}\left(\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2+\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\right)$$
$$\lesssim\alpha_t\sqrt{\log m}\left(\sqrt{\frac{n\log^5 m}{m}}+\frac{\sqrt{n\log^9 m}}{m}\right)\lesssim\alpha_t,$$

with the proviso that $m\gg n\log^6 m$. Therefore, we can further obtain

$$\left|x_\|^{t+1}-x_\|^{t+1,(l)}\right|\le\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\cdot O\left(\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right\|_2+\left|x_\|^t\right|^2+\sqrt{\frac{n\log^3 m}{m}}\right)\right\}\left|x_\|^t-x_\|^{t,(l)}\right|$$
$$+O\left(\eta\frac{\log^2 m}{m}\left\|\boldsymbol{x}^t\right\|_2\right)+O\left(\eta\sqrt{\log m}\left\|\boldsymbol{x}_\perp^t-\boldsymbol{x}_\perp^{t,(l)}-\boldsymbol{x}_\perp^{t,\mathrm{sgn}}+\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2\right)$$
$$+O\left(\eta\alpha_t\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right\|_2\right)$$
$$\le\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_1\right\}\left|x_\|^t-x_\|^{t,(l)}\right|+O\left(\eta\alpha_t\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right\|_2\right)$$
$$+O\left(\eta\frac{\log^2 m}{m}\left\|\boldsymbol{x}^t\right\|_2\right)+O\left(\eta\sqrt{\log m}\left\|\boldsymbol{x}_\perp^t-\boldsymbol{x}_\perp^{t,(l)}-\boldsymbol{x}_\perp^{t,\mathrm{sgn}}+\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2\right)$$

for some $|\phi_1|\ll\frac{1}{\log m}$. Here the last inequality comes from the sample complexity $m\gg n\log^5 m$, the assumption $\alpha_t\ll\frac{1}{\log^5 m}$ and the fact (63a). Given the inductive hypotheses (40), we can conclude

$$\left|x_\|^{t+1}-x_\|^{t+1,(l)}\right|\le\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_1\right\}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_2\frac{\sqrt{n\log^{12}m}}{m}$$
$$+O\left(\frac{\eta\log^2 m}{m}(\alpha_t+\beta_t)\right)+O\left(\eta\sqrt{\log m}\cdot\alpha_t\left(1+\frac{1}{\log m}\right)^t C_4\frac{\sqrt{n\log^9 m}}{m}\right)$$
$$+O\left(\eta\alpha_t\beta_t\left(1+\frac{1}{\log m}\right)^t C_1\frac{\sqrt{n\log^5 m}}{m}\right)$$
$$\overset{\text{(i)}}{\le}\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_2\right\}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_2\frac{\sqrt{n\log^{12}m}}{m}$$
$$\overset{\text{(ii)}}{\le}\alpha_{t+1}\left(1+\frac{1}{\log m}\right)^{t+1} C_2\frac{\sqrt{n\log^{12}m}}{m}$$

for some $|\phi_2|\ll\frac{1}{\log m}$. Here, the inequality (i) holds true as long as

$$\frac{\log^2 m}{m}(\alpha_t+\beta_t)\ll\frac{1}{\log m}\alpha_t C_2\frac{\sqrt{n\log^{12}m}}{m}\tag{124a}$$

$$\sqrt{\log m}C_4\frac{\sqrt{n\log^9 m}}{m} \ll \frac{1}{\log m}C_2\frac{\sqrt{n\log^{12} m}}{m} \tag{124b}$$

$$\beta_t C_1\frac{\sqrt{n\log^5 m}}{m} \ll \frac{1}{\log m}C_2\frac{\sqrt{n\log^{12} m}}{m}, \tag{124c}$$

where the first condition (124a) is satisfied since (according to Lemma 1)

$$\alpha_t+\beta_t \lesssim \beta_t \lesssim \alpha_t\sqrt{n\log m}.$$

The second condition (124b) holds as long as $C_2 \gg C_4$. The third one (124c) holds trivially. Moreover, the second inequality (ii) follows from the same reasoning as in (116). Specifically, we have for some $|\phi_3| \ll \frac{1}{\log m}$,

$$\begin{aligned}\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_2\right\}\alpha_t &= \left\{\frac{\alpha_{t+1}}{\alpha_t}+\eta\phi_3\right\}\alpha_t \\ &\le \left\{\frac{\alpha_{t+1}}{\alpha_t}+\eta O\left(\frac{\alpha_{t+1}}{\alpha_t}\phi_3\right)\right\}\alpha_t \\ &\le \alpha_{t+1}\left(1+\frac{1}{\log m}\right),\end{aligned}$$

as long as $\frac{\alpha_{t+1}}{\alpha_t} \asymp 1$.

The proof is completed by applying the union bound over all $1 \le l \le m$.

F Proof of Lemma 6

By similar calculations as in (109), we get the identity

$$\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,\mathrm{sgn}} = \left\{\boldsymbol{I}-\eta\int_0^1\nabla^2 f\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right)\mathrm{d}\tau\right\}\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right)+\eta\left(\nabla f^{\mathrm{sgn}}\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)-\nabla f\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)\right), \tag{125}$$

where $\tilde{\boldsymbol{x}}\left(\tau\right) := \boldsymbol{x}^t+\tau\left(\boldsymbol{x}^{t,\mathrm{sgn}}-\boldsymbol{x}^t\right)$. The first term satisfies

$$\begin{aligned}&\left\|\left\{\boldsymbol{I}-\eta\int_0^1\nabla^2 f\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right)\mathrm{d}\tau\right\}\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right)\right\| \\ &\quad\le \left\|\boldsymbol{I}-\eta\int_0^1\nabla^2 f\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right)\mathrm{d}\tau\right\|\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 \\ &\quad\le \left\{1+3\eta\left(1-\int_0^1\left\|\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_2^2\mathrm{d}\tau\right)+O\left(\eta\sqrt{\frac{n\log^3 m}{m}}\right)\right\}\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2,\end{aligned} \tag{126}$$

where we have invoked Lemma 15. Furthermore, one has for all $0 \le \tau \le 1$

$$\begin{aligned}\left\|\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_2^2 &\ge \left\|\boldsymbol{x}^t\right\|_2^2-\left|\left\|\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_2^2-\left\|\boldsymbol{x}^t\right\|_2^2\right| \\ &\ge \left\|\boldsymbol{x}^t\right\|_2^2-\left\|\tilde{\boldsymbol{x}}\left(\tau\right)-\boldsymbol{x}^t\right\|_2\left(\left\|\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_2+\left\|\boldsymbol{x}^t\right\|_2\right) \\ &\ge \left\|\boldsymbol{x}^t\right\|_2^2-\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2\left(\left\|\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_2+\left\|\boldsymbol{x}^t\right\|_2\right).\end{aligned}$$

This combined with the norm conditions $\|\boldsymbol{x}^t\|_2 \lesssim 1$, $\|\tilde{\boldsymbol{x}}\left(\tau\right)\|_2 \lesssim 1$ reveals that

$$\min_{0\le\tau\le1}\left\|\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_2^2 \ge \left\|\boldsymbol{x}^t\right\|_2^2+O\left(\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2\right),$$

and hence we can further upper bound (126) as

$$\left\|\left\{\boldsymbol{I}-\eta\int_0^1\nabla^2 f\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right)\mathrm{d}\tau\right\}\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right)\right\|$$

$$\le \left\{1+3\eta\left(1-\left\|\boldsymbol{x}^{t}\right\|_{2}^{2}\right)+\eta\cdot O\left(\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_{2}+\sqrt{\frac{n\log^{3}m}{m}}\right)\right\}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_{2}$$

$$\le \left\{1+3\eta\left(1-\left\|\boldsymbol{x}^{t}\right\|_{2}^{2}\right)+\eta\phi_{1}\right\}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_{2},$$

for some $|\phi_1| \ll \frac{1}{\log m}$, where the last line follows from $m \gg n\log^5 m$ and the fact (63b).

The remainder of this subsection is largely devoted to controlling the gradient difference $\nabla f^{\mathrm{sgn}}\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)-\nabla f\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)$ in (125). By the definition of $f^{\mathrm{sgn}}\left(\cdot\right)$, one has

$$\begin{aligned}
&\nabla f^{\mathrm{sgn}}\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)-\nabla f\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)\\
&=\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\boldsymbol{a}_{i}^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^{3}\boldsymbol{a}_{i}^{\mathrm{sgn}}-\left(\boldsymbol{a}_{i}^{\mathrm{sgn}\top}\boldsymbol{x}^{\natural}\right)^{2}\left(\boldsymbol{a}_{i}^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\boldsymbol{a}_{i}^{\mathrm{sgn}}-\left(\boldsymbol{a}_{i}^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^{3}\boldsymbol{a}_{i}+\left(\boldsymbol{a}_{i}^{\top}\boldsymbol{x}^{\natural}\right)^{2}\left(\boldsymbol{a}_{i}^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\boldsymbol{a}_{i}\right\}\\
&=\underbrace{\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\boldsymbol{a}_{i}^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^{3}\boldsymbol{a}_{i}^{\mathrm{sgn}}-\left(\boldsymbol{a}_{i}^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^{3}\boldsymbol{a}_{i}\right\}}_{:=\boldsymbol{r}_{1}}-\underbrace{\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^{2}\left(\boldsymbol{a}_{i}^{\mathrm{sgn}}\boldsymbol{a}_{i}^{\mathrm{sgn}\top}-\boldsymbol{a}_{i}\boldsymbol{a}_{i}^{\top}\right)\boldsymbol{x}^{t,\mathrm{sgn}}}_{:=\boldsymbol{r}_{2}}.
\end{aligned}$$

Here, the last identity holds because of $\left(\boldsymbol{a}_{i}^{\top}\boldsymbol{x}^{\natural}\right)^{2}=\left(\boldsymbol{a}_{i}^{\mathrm{sgn}\top}\boldsymbol{x}^{\natural}\right)^{2}=a_{i,1}^{2}$ (see (37)).

- We begin with the second term $\boldsymbol{r}_2$. By construction, one has $\boldsymbol{a}_{i,\perp}^{\mathrm{sgn}}=\boldsymbol{a}_{i,\perp}$, $a_{i,1}^{\mathrm{sgn}}=\xi_{i}^{\mathrm{sgn}}\left|a_{i,1}\right|$ and $a_{i,1}=\xi_{1}\left|a_{i,1}\right|$. These taken together yield

$$\boldsymbol{a}_{i}^{\mathrm{sgn}}\boldsymbol{a}_{i}^{\mathrm{sgn}\top}-\boldsymbol{a}_{i}\boldsymbol{a}_{i}^{\top}=\left(\xi_{i}^{\mathrm{sgn}}-\xi_{i}\right)\left|a_{i,1}\right|\left[\begin{array}{cc}0 & \boldsymbol{a}_{i,\perp}^{\top}\\ \boldsymbol{a}_{i,\perp} & \boldsymbol{0}\end{array}\right], \tag{127}$$

and hence $\boldsymbol{r}_2$ can be rewritten as

$$\boldsymbol{r}_{2}=\left[\begin{array}{c}\frac{1}{m}\sum_{i=1}^{m}\left(\xi_{i}^{\mathrm{sgn}}-\xi_{i}\right)\left|a_{i,1}\right|^{3}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\\ x_{\|}^{t,\mathrm{sgn}}\cdot\frac{1}{m}\sum_{i=1}^{m}\left(\xi_{i}^{\mathrm{sgn}}-\xi_{i}\right)\left|a_{i,1}\right|^{3}\boldsymbol{a}_{i,\perp}\end{array}\right]. \tag{128}$$

For the first entry of $\boldsymbol{r}_2$, the triangle inequality gives

$$\begin{aligned}
\left|\frac{1}{m}\sum_{i=1}^{m}\left(\xi_{i}^{\mathrm{sgn}}-\xi_{i}\right)\left|a_{i,1}\right|^{3}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right| &\le \underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\left|a_{i,1}\right|^{3}\xi_{i}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right|}_{:=\phi_{1}}+\underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\left|a_{i,1}\right|^{3}\xi_{i}^{\mathrm{sgn}}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right|}_{:=\phi_{2}}\\
&\quad+\underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\left|a_{i,1}\right|^{3}\xi_{i}^{\mathrm{sgn}}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t}\right)\right|}_{:=\phi_{3}}.
\end{aligned}$$

Regarding ϕ_1, we make use of the independence between ξ_i and $\left|a_{i,1}\right|^{3}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}$ and invoke the Bernstein inequality (see Lemma 11) to reach that with probability at least $1-O\left(m^{-10}\right)$,

$$\phi_{1}\lesssim\frac{1}{m}\left(\sqrt{V_{1}\log m}+B_{1}\log m\right),$$

where V_1 and B_1 are defined to be

$$V_{1}:=\sum_{i=1}^{m}\left|a_{i,1}\right|^{6}\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right|^{2}\qquad\text{and}\qquad B_{1}:=\max_{1\le i\le m}\left\{\left|a_{i,1}\right|^{3}\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right|\right\}.$$

It is easy to see from Lemma 12 and the incoherence condition (62d) that with probability exceeding $1-O\left(m^{-10}\right)$, $V_{1}\lesssim m\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}^{2}$ and $B_{1}\lesssim\log^{2}m\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}$, which implies

$$\phi_{1}\lesssim\left(\sqrt{\frac{\log m}{m}}+\frac{\log^{3}m}{m}\right)\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}\asymp\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2},$$

as long as $m \gg \log^5 m$. Similarly, one can obtain

$$\phi_2 \lesssim \sqrt{\frac{\log m}{m}} \left\| \boldsymbol{x}_\perp^t \right\|_2.$$

The last term ϕ_3 can be bounded through the Cauchy-Schwarz inequality. Specifically, one has

$$\phi_3 \leq \left\| \frac{1}{m} \sum_{i=1}^m \left| a_{i,1} \right|^3 \xi_i^{\text{sgn}} \boldsymbol{a}_{i,\perp} \right\|_2 \left\| \boldsymbol{x}_\perp^{t,\text{sgn}} - \boldsymbol{x}_\perp^t \right\|_2 \lesssim \sqrt{\frac{n \log^3 m}{m}} \left\| \boldsymbol{x}_\perp^{t,\text{sgn}} - \boldsymbol{x}_\perp^t \right\|_2,$$

where the second relation arises from Lemma 14. The previous three bounds taken collectively yield

$$\begin{aligned} \left| \frac{1}{m} \sum_{i=1}^m \left(\xi_i^{\text{sgn}} - \xi_i \right) \left| a_{i,1} \right|^3 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{x}_\perp^{t,\text{sgn}} \right| &\lesssim \sqrt{\frac{\log m}{m}} \left(\left\| \boldsymbol{x}_\perp^{t,\text{sgn}} \right\|_2 + \left\| \boldsymbol{x}_\perp^t \right\|_2 \right) + \sqrt{\frac{n \log^3 m}{m}} \left\| \boldsymbol{x}_\perp^{t,\text{sgn}} - \boldsymbol{x}_\perp^t \right\|_2 \\ &\lesssim \sqrt{\frac{\log m}{m}} \left\| \boldsymbol{x}_\perp^t \right\|_2 + \sqrt{\frac{n \log^3 m}{m}} \left\| \boldsymbol{x}_\perp^{t,\text{sgn}} - \boldsymbol{x}_\perp^t \right\|_2. \end{aligned} \tag{129}$$

Here the second inequality results from the triangle inequality $\left\| \boldsymbol{x}_\perp^{t,\text{sgn}} \right\|_2 \leq \left\| \boldsymbol{x}_\perp^t \right\|_2 + \left\| \boldsymbol{x}_\perp^{t,\text{sgn}} - \boldsymbol{x}_\perp^t \right\|_2$ and the fact that $\sqrt{\frac{\log m}{m}} \leq \sqrt{\frac{n \log^3 m}{m}}$. In addition, for the second through the nth entries of $\boldsymbol{r}_2$, one can again invoke Lemma 14 to obtain

$$\begin{aligned} \left\| \frac{1}{m} \sum_{i=1}^m \left| a_{i,1} \right|^3 \left(\xi_i^{\text{sgn}} - \xi_i \right) \boldsymbol{a}_{i,\perp} \right\|_2 &\leq \left\| \frac{1}{m} \sum_{i=1}^m \left| a_{i,1} \right|^3 \xi_i^{\text{sgn}} \boldsymbol{a}_{i,\perp} \right\|_2 + \left\| \frac{1}{m} \sum_{i=1}^m \left| a_{i,1} \right|^3 \xi_i \boldsymbol{a}_{i,\perp} \right\|_2 \\ &\lesssim \sqrt{\frac{n \log^3 m}{m}}. \end{aligned} \tag{130}$$

This combined with (128) and (129) yields

$$\left\| \boldsymbol{r}_2 \right\|_2 \lesssim \sqrt{\frac{\log m}{m}} \left\| \boldsymbol{x}_\perp^t \right\|_2 + \sqrt{\frac{n \log^3 m}{m}} \left\| \boldsymbol{x}_\perp^{t,\text{sgn}} - \boldsymbol{x}_\perp^t \right\|_2 + \left| x_\parallel^{t,\text{sgn}} \right| \sqrt{\frac{n \log^3 m}{m}}.$$

- Moving on to the term $\boldsymbol{r}_1$, we can also decompose

$$\boldsymbol{r}_1 = \begin{bmatrix} \frac{1}{m} \sum_{i=1}^m \left\{ \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^3 a_{i,1}^{\text{sgn}} - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^{t,\text{sgn}} \right)^3 a_{i,1} \right\} \\ \frac{1}{m} \sum_{i=1}^m \left\{ \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^3 \boldsymbol{a}_{i,\perp}^{\text{sgn}} - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^{t,\text{sgn}} \right)^3 \boldsymbol{a}_{i,\perp} \right\} \end{bmatrix}.$$

For the second through the nth entries, we see that

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^m \left\{ \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^3 \boldsymbol{a}_{i,\perp}^{\text{sgn}} - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^{t,\text{sgn}} \right)^3 \boldsymbol{a}_{i,\perp} \right\} &\overset{\text{(i)}}{=} \frac{1}{m} \sum_{i=1}^m \left\{ \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^3 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^{t,\text{sgn}} \right)^3 \right\} \boldsymbol{a}_{i,\perp} \\ &\overset{\text{(ii)}}{=} \frac{1}{m} \sum_{i=1}^m \left\{ \left(\xi_i^{\text{sgn}} - \xi_i \right) \left| a_{i,1} \right| x_\parallel^{t,\text{sgn}} \left[\left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^2 + \left(\boldsymbol{a}_i^\top \boldsymbol{x}^{t,\text{sgn}} \right)^2 + \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right) \left(\boldsymbol{a}_i^\top \boldsymbol{x}^{t,\text{sgn}} \right) \right] \right\} \boldsymbol{a}_{i,\perp} \\ &= \frac{x_\parallel^{t,\text{sgn}}}{m} \sum_{i=1}^m \left\{ \left(\xi_i^{\text{sgn}} - \xi_i \right) \left| a_{i,1} \right| \left[\left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^2 + \left(\boldsymbol{a}_i^\top \boldsymbol{x}^{t,\text{sgn}} \right)^2 + \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right) \left(\boldsymbol{a}_i^\top \boldsymbol{x}^{t,\text{sgn}} \right) \right] \right\} \boldsymbol{a}_{i,\perp}, \end{aligned}$$

where (i) follows from $\boldsymbol{a}_{i,\perp}^{\text{sgn}} = \boldsymbol{a}_{i,\perp}$ and (ii) relies on the elementary identity $a^3 - b^3 = (a-b)\left(a^2 + b^2 + ab\right)$. Treating $\frac{1}{m} \sum_{i=1}^m \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^2 a_{i,1}^{\text{sgn}} \boldsymbol{a}_{i,\perp}$ as the first column (except its first entry) of $\frac{1}{m} \sum_{i=1}^m \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^2 \boldsymbol{a}_i^{\text{sgn}} \boldsymbol{a}_i^{\text{sgn}\top}$, by Lemma 14 and the incoherence condition (62e), we have

$$\frac{1}{m} \sum_{i=1}^m \xi_i^{\text{sgn}} \left| a_{i,1} \right| \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^2 \boldsymbol{a}_{i,\perp} = \frac{1}{m} \sum_{i=1}^m \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}} \right)^2 a_{i,1}^{\text{sgn}} \boldsymbol{a}_{i,\perp} = 2 x_\parallel^{t,\text{sgn}} \boldsymbol{x}_\perp^{t,\text{sgn}} + \boldsymbol{v}_1,$$

where $\|\boldsymbol{v}_1\|_2 \lesssim \sqrt{\frac{n\log^3 m}{m}}$. Similarly,

$$-\frac{1}{m}\sum_{i=1}^{m}\xi_i\left|a_{i,1}\right|\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2\boldsymbol{a}_{i,\perp} = -2x_{\|}^{t,\mathrm{sgn}}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} + \boldsymbol{v}_2,$$

where $\|\boldsymbol{v}_2\|_2 \lesssim \sqrt{\frac{n\log^3 m}{m}}$. Moreover, we have

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2\boldsymbol{a}_{i,\perp}\\
&\quad= \frac{1}{m}\sum_{i=1}^{m}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2\boldsymbol{a}_{i,\perp} + \frac{1}{m}\sum_{i=1}^{m}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\left[\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2 - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2\right]\boldsymbol{a}_{i,\perp}\\
&\quad= 2x_{\|}^{t,\mathrm{sgn}}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} + \boldsymbol{v}_1 + \boldsymbol{v}_3,
\end{aligned}$$

where $\boldsymbol{v}_3$ is defined as

$$\begin{aligned}
\boldsymbol{v}_3 &= \frac{1}{m}\sum_{i=1}^{m}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\left[\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2 - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2\right]\boldsymbol{a}_{i,\perp}\\
&= 2x_{\|}^{t,\mathrm{sgn}}\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i - \xi_i^{\mathrm{sgn}}\right)\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp}\\
&= 2x_{\|}^{t,\mathrm{sgn}}\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i\xi_i^{\mathrm{sgn}} - 1\right)\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}.
\end{aligned} \tag{131}$$

Here the second equality comes from the identity (122). Similarly one can get

$$-\frac{1}{m}\sum_{i=1}^{m}\xi_i\left|a_{i,1}\right|\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2\boldsymbol{a}_{i,\perp} = -2x_{\|}^{t,\mathrm{sgn}}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{v}_2 - \boldsymbol{v}_4,$$

where $\boldsymbol{v}_4$ obeys

$$\begin{aligned}
\boldsymbol{v}_4 &= \frac{1}{m}\sum_{i=1}^{m}\xi_i\left|a_{i,1}\right|\left[\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2 - \left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2\right]\boldsymbol{a}_{i,\perp}\\
&= 2x_{\|}^{t,\mathrm{sgn}}\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i\xi_i^{\mathrm{sgn}} - 1\right)\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}.
\end{aligned}$$

It remains to bound $\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i^{\mathrm{sgn}} - \xi_i\right)\left|a_{i,1}\right|\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\boldsymbol{a}_{i,\perp}$. To this end, we have

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\boldsymbol{a}_{i,\perp}\\
&\quad= \frac{1}{m}\sum_{i=1}^{m}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2\boldsymbol{a}_{i,\perp} + \frac{1}{m}\sum_{i=1}^{m}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\left[\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right) - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\right]\boldsymbol{a}_{i,\perp}\\
&\quad= 2x_{\|}^{t,\mathrm{sgn}}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} + \boldsymbol{v}_1 + \boldsymbol{v}_5,
\end{aligned}$$

where

$$\boldsymbol{v}_5 = x_{\|}^{t,\mathrm{sgn}}\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i\xi_i^{\mathrm{sgn}} - 1\right)\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}.$$

The same argument yields

$$-\frac{1}{m}\sum_{i=1}^{m}\xi_i\left|a_{i,1}\right|\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)\boldsymbol{a}_{i,\perp} = -2x_{\|}^{t,\mathrm{sgn}}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{v}_2 - \boldsymbol{v}_6,$$

where

$$\boldsymbol{v}_6 = x_{\|}^{t,\mathrm{sgn}} \frac{1}{m} \sum_{i=1}^{m} \left(\xi_i \xi_i^{\mathrm{sgn}} - 1\right) \left|a_{i,1}\right|^2 \boldsymbol{a}_{i,\perp} \boldsymbol{a}_i^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn}}.$$

Combining all of the previous bounds and recognizing that $\boldsymbol{v}_3 = \boldsymbol{v}_4$ and $\boldsymbol{v}_5 = \boldsymbol{v}_6$, we arrive at

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \left\{ \left(\boldsymbol{a}_i^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn}}\right)^3 \boldsymbol{a}_{i,\perp}^{\mathrm{sgn}} - \left(\boldsymbol{a}_i^{\top} \boldsymbol{x}^{t,\mathrm{sgn}}\right)^3 \boldsymbol{a}_{i,\perp} \right\} \right\|_2 \lesssim \|\boldsymbol{v}_1\|_2 + \|\boldsymbol{v}_2\|_2 \lesssim \sqrt{\frac{n \log^3 m}{m}} \left| x_{\|}^{t,\mathrm{sgn}} \right|.$$

Regarding the first entry of $\boldsymbol{r}_1$, one has

$$\begin{aligned}
&\left| \frac{1}{m} \sum_{i=1}^{m} \left\{ \left(\boldsymbol{a}_i^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn}}\right)^3 a_{i,1}^{\mathrm{sgn}} - \left(\boldsymbol{a}_i^{\top} \boldsymbol{x}^{t,\mathrm{sgn}}\right)^3 a_{i,1} \right\} \right| \\
&\quad = \left| \frac{1}{m} \sum_{i=1}^{m} \left\{ \left(\xi_i^{\mathrm{sgn}} \left|a_{i,1}\right| x_{\|}^{t,\mathrm{sgn}} + \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} \right)^3 \xi_i^{\mathrm{sgn}} \left|a_{i,1}\right| - \left(\xi_i \left|a_{i,1}\right| x_{\|}^{t,\mathrm{sgn}} + \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} \right)^3 \xi_i \left|a_{i,1}\right| \right\} \right| \\
&\quad = \left| \frac{1}{m} \sum_{i=1}^{m} \left(\xi_i^{\mathrm{sgn}} - \xi_i\right) \left|a_{i,1}\right| \left\{ 3 \left|a_{i,1}\right|^2 \left| x_{\|}^{t,\mathrm{sgn}} \right|^2 \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} + \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} \right)^3 \right\} \right|.
\end{aligned}$$

In view of the independence between ξ_i and $\left|a_{i,1}\right| \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^3$, from the Bernstein's inequality (see Lemma 11), we have that

$$\left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \left|a_{i,1}\right| \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^3 \right| \lesssim \frac{1}{m} \left(\sqrt{V_2 \log m} + B_2 \log m \right)$$

holds with probability exceeding $1 - O\left(m^{-10}\right)$, where

$$V_2 := \sum_{i=1}^{m} \left|a_{i,1}\right|^2 \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^6 \qquad \text{and} \qquad B_2 := \max_{1 \le i \le m} \left|a_{i,1}\right| \left| \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} \right|^3.$$

It is straightforward to check that $V_2 \lesssim m \left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2^6$ and $B_2 \lesssim \log^2 m \left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2^3$, which further implies

$$\left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \left|a_{i,1}\right| \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^3 \right| \lesssim \sqrt{\frac{\log m}{m}} \left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2^3 + \frac{\log^3 m}{m} \left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2^3 \asymp \sqrt{\frac{\log m}{m}} \left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2^3,$$

as long as $m \gg \log^5 m$. For the term involving ξ_i^{sgn}, we have

$$\frac{1}{m} \sum_{i=1}^{m} \xi_i^{\mathrm{sgn}} \left|a_{i,1}\right| \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^3 = \underbrace{\frac{1}{m} \sum_{i=1}^{m} \xi_i^{\mathrm{sgn}} \left|a_{i,1}\right| \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right)^3}_{:=\theta_1} + \underbrace{\frac{1}{m} \sum_{i=1}^{m} \xi_i^{\mathrm{sgn}} \left|a_{i,1}\right| \left[\left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right)^3 - \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^3 \right]}_{:=\theta_2}.$$

Similarly one can obtain

$$\left|\theta_1\right| \lesssim \sqrt{\frac{\log m}{m}} \left\|\boldsymbol{x}_{\perp}^{t}\right\|_2^3.$$

Expand θ_2 using the elementary identity $a^3 - b^3 = (a-b)\left(a^2 + ab + b^2\right)$ to get

$$\begin{aligned}
\theta_2 &= \frac{1}{m} \sum_{i=1}^{m} \xi_i^{\mathrm{sgn}} \left|a_{i,1}\right| \boldsymbol{a}_{i,\perp}^{\top} \left(\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right) \left[\left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right)^2 + \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^2 + \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right) \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right) \right] \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right)^2 \xi_i^{\mathrm{sgn}} \left|a_{i,1}\right| \boldsymbol{a}_{i,\perp}^{\top} \left(\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right) \\
&\quad + \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^2 \xi_i^{\mathrm{sgn}} \left|a_{i,1}\right| \boldsymbol{a}_{i,\perp}^{\top} \left(\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)
\end{aligned}$$

$$+\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right)\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t}\right)\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right).$$

Once more, we can apply Lemma 14 with the incoherence conditions (62b) and (62d) to obtain

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right)^2\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\right\|_2\lesssim\sqrt{\frac{n\log^3 m}{m}};$$

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)^2\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\right\|_2\lesssim\sqrt{\frac{n\log^3 m}{m}}.$$

In addition, one can use the Cauchy-Schwarz inequality to deduce that

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right)\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t}\right)\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)\right|\\
&\quad\leq\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right)^2\left[\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t}\right)\right]^2}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left|a_{i,1}\right|^2\left[\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)\right]^2}\\
&\quad\leq\sqrt{\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right)^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}\right\|\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t}\right\|_2^2}\sqrt{\left\|\frac{1}{m}\sum_{i=1}^{m}\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}\right\|\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t}\right\|_2^2}\\
&\quad\lesssim\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{\mathrm{sgn}}\right\|_2^2,
\end{aligned}$$

where the last inequality comes from Lemma 14. Combine the preceding bounds to reach

$$|\theta_2|\lesssim\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{\mathrm{sgn}}\right\|_2+\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{\mathrm{sgn}}\right\|_2^2.$$

Applying the similar arguments as above we get

$$\begin{aligned}
&\left|\left|x_{\parallel}^{t,\mathrm{sgn}}\right|^2\frac{3}{m}\sum_{i=1}^{m}\left(\xi_i^{\mathrm{sgn}}-\xi_i\right)\left|a_{i,1}\right|^3\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right|\\
&\quad\lesssim\left|x_{\parallel}^{t,\mathrm{sgn}}\right|^2\left(\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2+\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2+\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2\right)\\
&\quad\lesssim\left|x_{\parallel}^{t,\mathrm{sgn}}\right|^2\left(\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2+\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2\right),
\end{aligned}$$

where the last line follows from the triangle inequality $\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2\leq\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2+\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2$ and the fact that $\sqrt{\frac{\log m}{m}}\leq\sqrt{\frac{n\log^3 m}{m}}$. Putting the above results together yields

$$\begin{aligned}
\left\|\boldsymbol{r}_1\right\|_2\lesssim&\sqrt{\frac{n\log^3 m}{m}}\left|x_{\parallel}^{t,\mathrm{sgn}}\right|+\sqrt{\frac{\log m}{m}}\left(\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2+\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2\right)+\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{\mathrm{sgn}}\right\|_2\\
&+\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{\mathrm{sgn}}\right\|_2^2+\left|x_{\parallel}^{t,\mathrm{sgn}}\right|^2\left(\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2+\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{\mathrm{sgn}}\right\|_2\right),
\end{aligned}$$

which can be further simplified to

$$\left\|\boldsymbol{r}_1\right\|_2\lesssim\sqrt{\frac{n\log^3 m}{m}}\left|x_{\parallel}^{t}\right|+\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2+\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{\mathrm{sgn}}\right\|_2+\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{\mathrm{sgn}}\right\|_2^2.$$

- Combine all of the above estimates to reach

$$\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,\mathrm{sgn}}\right\|_2 \le \left\|\left\{\boldsymbol{I}-\eta\int_0^1 \nabla^2 f\left(\tilde{\boldsymbol{x}}(\tau)\right)\mathrm{d}\tau\right\}\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right)\right\|_2 + \eta\left\|\nabla f^{\mathrm{sgn}}\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)-\nabla f\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)\right\|_2$$
$$\le \left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_2\right\}\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 + O\left(\eta\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_\perp^t\right\|_2\right)+\eta\sqrt{\frac{n\log^3 m}{m}}\left|x_\parallel^t\right|$$

for some $|\phi_2| \ll \frac{1}{\log m}$. Here the second inequality follows from the fact (63b). Substitute the induction hypotheses into this bound to reach

$$\begin{aligned}\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,\mathrm{sgn}}\right\|_2 &\le \left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_2\right\}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}} \\ &\quad + \eta\sqrt{\frac{\log m}{m}}\beta_t + \eta\sqrt{\frac{n\log^3 m}{m}}\alpha_t \\ &\overset{\text{(i)}}{\le} \left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_3\right\}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}} \\ &\overset{\text{(ii)}}{\le} \alpha_{t+1}\left(1+\frac{1}{\log m}\right)^{t+1} C_3\sqrt{\frac{n\log^5 m}{m}},\end{aligned}$$

for some $|\phi_3| \ll \frac{1}{\log m}$, where (ii) follows the same reasoning as in (116) and (i) holds as long as

$$\sqrt{\frac{\log m}{m}}\beta_t \ll \frac{1}{\log m}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}}, \tag{132a}$$
$$\sqrt{\frac{n\log^3 m}{m}}\alpha_t \ll \frac{1}{\log m}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}}. \tag{132b}$$

Here the first condition (132a) results from (see Lemma 1)

$$\beta_t \lesssim \sqrt{n\log m}\cdot\alpha_t,$$

and the second one is trivially true with the proviso that $C_3 > 0$ is sufficiently large.

G Proof of Lemma 7

Consider any l $(1 \le l \le m)$. According to the gradient update rules (3), (29), (30) and (31), we have

$$\begin{aligned}&\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^{t+1,\mathrm{sgn}}+\boldsymbol{x}^{t+1,\mathrm{sgn},(l)} \\ &\quad = \boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}-\eta\left[\nabla f\left(\boldsymbol{x}^t\right)-\nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right)-\nabla f^{\mathrm{sgn}}\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)+\nabla f^{\mathrm{sgn},(l)}\left(\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right)\right].\end{aligned}$$

It then boils down to controlling the gradient difference, i.e. $\nabla f\left(\boldsymbol{x}^t\right)-\nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right)-\nabla f^{\mathrm{sgn}}\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)+\nabla f^{\mathrm{sgn},(l)}\left(\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right)$. To this end, we first see that

$$\begin{aligned}\nabla f\left(\boldsymbol{x}^t\right)-\nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right) &= \nabla f\left(\boldsymbol{x}^t\right)-\nabla f\left(\boldsymbol{x}^{t,(l)}\right)+\nabla f\left(\boldsymbol{x}^{t,(l)}\right)-\nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right) \\ &= \left(\int_0^1 \nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right)+\frac{1}{m}\left[\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2-\left(\boldsymbol{a}_l^\top\boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)},\end{aligned} \tag{133}$$

where we denote $\boldsymbol{x}(\tau) := \boldsymbol{x}^t+\tau\left(\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^t\right)$ and the last identity results from the fundamental theorem of calculus [Lan93, Chapter XIII, Theorem 4.2]. Similar calculations yield

$$\nabla f^{\mathrm{sgn}}\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)-\nabla f^{\mathrm{sgn},(l)}\left(\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right)$$

$$= \left(\int_0^1 \nabla^2 f^{\text{sgn}}\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right) \mathrm{d}\tau\right) \left(\boldsymbol{x}^{t,\text{sgn}} - \boldsymbol{x}^{t,\text{sgn}(l)}\right) + \frac{1}{m}\left[\left(\boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn},(l)}\right)^2 - \left(\boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{\natural}\right)^2\right] \boldsymbol{a}_l^{\text{sgn}} \boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn},(l)} \tag{134}$$

with $\tilde{\boldsymbol{x}}(\tau) := \boldsymbol{x}^{t,\text{sgn}} + \tau\left(\boldsymbol{x}^{t,\text{sgn},(l)} - \boldsymbol{x}^{t,\text{sgn}}\right)$. Combine (133) and (134) to arrive at

$$\begin{aligned}
&\nabla f\left(\boldsymbol{x}^t\right) - \nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right) - \nabla f^{\text{sgn}}\left(\boldsymbol{x}^{t,\text{sgn}}\right) + \nabla f^{\text{sgn},(l)}\left(\boldsymbol{x}^{t,\text{sgn},(l)}\right) \\
&\quad = \underbrace{\left(\int_0^1 \nabla^2 f\left(\boldsymbol{x}(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right) - \left(\int_0^1 \nabla^2 f^{\text{sgn}}\left(\tilde{\boldsymbol{x}}(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}^{t,\text{sgn}} - \boldsymbol{x}^{t,\text{sgn},(l)}\right)}_{:=\boldsymbol{v}_1} \\
&\qquad + \underbrace{\frac{1}{m}\left[\left(\boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)}\right)^2 - \left(\boldsymbol{a}_l^\top \boldsymbol{x}^{\natural}\right)^2\right]\boldsymbol{a}_l \boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)} - \frac{1}{m}\left[\left(\boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn},(l)}\right)^2 - \left(\boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{\natural}\right)^2\right]\boldsymbol{a}_l^{\text{sgn}} \boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn},(l)}}_{:=\boldsymbol{v}_2}.
\end{aligned} \tag{135}$$

In what follows, we shall control $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ separately.

- We start with the simpler term $\boldsymbol{v}_2$. In light of the fact that $\left(\boldsymbol{a}_l^\top \boldsymbol{x}^{\natural}\right)^2 = \left(\boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{\natural}\right)^2 = \left|a_{l,1}\right|^2$ (see (37)), one can decompose $\boldsymbol{v}_2$ as

$$\begin{aligned}
m\boldsymbol{v}_2 &= \underbrace{\left[\left(\boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)}\right)^2 - \left(\boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn},(l)}\right)^2\right]\boldsymbol{a}_l \boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)}}_{:=\boldsymbol{\theta}_1} \\
&\quad + \left[\left(\boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn},(l)}\right)^2 - \left|a_{l,1}\right|^2\right]\underbrace{\left(\boldsymbol{a}_l \boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)} - \boldsymbol{a}_l^{\text{sgn}} \boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn},(l)}\right)}_{:=\boldsymbol{\theta}_2}.
\end{aligned}$$

  First, it is easy to see from (56) and the independence between $\boldsymbol{a}_l^{\text{sgn}}$ and $\boldsymbol{x}^{t,\text{sgn},(l)}$ that

$$\begin{aligned}
\left|\left(\boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn},(l)}\right)^2 - \left|a_{l,1}\right|^2\right| &\leq \left(\boldsymbol{a}_l^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn},(l)}\right)^2 + \left|a_{l,1}\right|^2 \\
&\lesssim \log m \cdot \left\|\boldsymbol{x}^{t,\text{sgn},(l)}\right\|_2^2 + \log m \lesssim \log m
\end{aligned} \tag{136}$$

  with probability at least $1-O\left(m^{-10}\right)$, where the last inequality results from the norm condition $\left\|\boldsymbol{x}^{t,\text{sgn},(l)}\right\|_2 \lesssim 1$ (see (61c)). Regarding the term $\boldsymbol{\theta}_2$, one has

$$\boldsymbol{\theta}_2 = \left(\boldsymbol{a}_l \boldsymbol{a}_l^\top - \boldsymbol{a}_l^{\text{sgn}} \boldsymbol{a}_l^{\text{sgn}\top}\right)\boldsymbol{x}^{t,(l)} + \boldsymbol{a}_l^{\text{sgn}} \boldsymbol{a}_l^{\text{sgn}\top}\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\text{sgn},(l)}\right),$$

  which together with the identity (127) gives

$$\boldsymbol{\theta}_2 = \left(\xi_l - \xi_l^{\text{sgn}}\right)\left|a_{l,1}\right|\begin{bmatrix}\boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^{t,(l)} \\ x_\parallel^{t,(l)} \boldsymbol{a}_{l,\perp}\end{bmatrix} + \boldsymbol{a}_l^{\text{sgn}} \boldsymbol{a}_l^{\text{sgn}\top}\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\text{sgn},(l)}\right).$$

  In view of the independence between $\boldsymbol{a}_l$ and $\boldsymbol{x}^{t,(l)}$, and between $\boldsymbol{a}_l^{\text{sgn}}$ and $\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\text{sgn},(l)}$, one can again apply standard Gaussian concentration results to obtain that

$$\left|\boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^{t,(l)}\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}_\perp^{t,(l)}\right\|_2 \qquad \text{and} \qquad \left|\boldsymbol{a}_l^{\text{sgn}\top}\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\text{sgn},(l)}\right)\right| \lesssim \sqrt{\log m}\left\|\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\text{sgn},(l)}\right\|_2$$

  with probability exceeding $1 - O\left(m^{-10}\right)$. Combining these two with the facts (56) and (57) leads to

$$\begin{aligned}
\left\|\boldsymbol{\theta}_2\right\|_2 &\leq \left|\xi_l - \xi_l^{\text{sgn}}\right|\left|a_{l,1}\right|\left(\left|\boldsymbol{a}_{l,\perp}^\top \boldsymbol{x}_\perp^{t,(l)}\right| + \left|x_\parallel^{t,(l)}\right|\left\|\boldsymbol{a}_{l,\perp}\right\|_2\right) + \left\|\boldsymbol{a}_l^{\text{sgn}}\right\|_2\left|\boldsymbol{a}_l^{\text{sgn}\top}\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\text{sgn},(l)}\right)\right| \\
&\lesssim \sqrt{\log m}\left(\sqrt{\log m}\left\|\boldsymbol{x}_\perp^{t,(l)}\right\|_2 + \sqrt{n}\left|x_\parallel^{t,(l)}\right|\right) + \sqrt{n\log m}\left\|\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\text{sgn},(l)}\right\|_2
\end{aligned}$$

$$\lesssim \log m \left\| \boldsymbol{x}_{\perp}^{t,(l)} \right\|_2 + \sqrt{n \log m} \left(\left| x_{\|}^{t,(l)} \right| + \left\| \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 \right). \tag{137}$$

We now move on to controlling $\boldsymbol{\theta}_1$. Use the elementary identity $a^2 - b^2 = (a-b)(a+b)$ to get

$$\boldsymbol{\theta}_1 = \left(\boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)} - \boldsymbol{a}_l^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right) \left(\boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)} + \boldsymbol{a}_l^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right) \boldsymbol{a}_l \boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)}. \tag{138}$$

The constructions of $\boldsymbol{a}_l^{\mathrm{sgn}}$ requires that

$$\boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)} - \boldsymbol{a}_l^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn},(l)} = \xi_l \left| a_{l,1} \right| x_{\|}^{t,(l)} - \xi_l^{\mathrm{sgn}} \left| a_{l,1} \right| x_{\|}^{t,\mathrm{sgn},(l)} + \boldsymbol{a}_{l,\perp}^{\top} \left(\boldsymbol{x}_{\perp}^{t,(l)} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)} \right).$$

Similarly, in view of the independence between $\boldsymbol{a}_{l,\perp}$ and $\boldsymbol{x}_{\perp}^{t,(l)} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}$, and the fact (56), one can see that with probability at least $1 - O\left(m^{-10}\right)$

$$\begin{aligned}
\left| \boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)} - \boldsymbol{a}_l^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right| &\le \left| \xi_l \right| \left| a_{l,1} \right| \left| x_{\|}^{t,(l)} \right| + \left| \xi_l^{\mathrm{sgn}} \right| \left| a_{l,1} \right| \left| x_{\|}^{t,\mathrm{sgn},(l)} \right| + \left| \boldsymbol{a}_{l,\perp}^{\top} \left(\boldsymbol{x}_{\perp}^{t,(l)} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)} \right) \right| \\
&\lesssim \sqrt{\log m} \left(\left| x_{\|}^{t,(l)} \right| + \left| x_{\|}^{t,\mathrm{sgn},(l)} \right| + \left\| \boldsymbol{x}_{\perp}^{t,(l)} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)} \right\|_2 \right) \\
&\lesssim \sqrt{\log m} \left(\left| x_{\|}^{t,(l)} \right| + \left\| \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 \right),
\end{aligned} \tag{139}$$

where the last inequality results from the triangle inequality $\left| x_{\|}^{t,\mathrm{sgn},(l)} \right| \le \left| x_{\|}^{t,(l)} \right| + \left\| \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2$. Substituting (139) into (138) results in

$$\begin{aligned}
\left\| \boldsymbol{\theta}_1 \right\|_2 &= \left| \boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)} - \boldsymbol{a}_l^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right| \left| \boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)} + \boldsymbol{a}_l^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right| \left\| \boldsymbol{a}_l \right\|_2 \left| \boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)} \right| \\
&\lesssim \sqrt{\log m} \left(\left| x_{\|}^{t,(l)} \right| + \left\| \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 \right) \cdot \sqrt{\log m} \cdot \sqrt{n} \cdot \sqrt{\log m} \\
&\asymp \sqrt{n \log^3 m} \left(\left| x_{\|}^{t,(l)} \right| + \left\| \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 \right),
\end{aligned} \tag{140}$$

where the second line comes from the simple facts (57),

$$\left| \boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)} + \boldsymbol{a}_l^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right| \le \sqrt{\log m} \qquad \text{and} \qquad \left| \boldsymbol{a}_l^{\top} \boldsymbol{x}^{t,(l)} \right| \lesssim \sqrt{\log m}.$$

Taking the bounds (136), (137) and (140) collectively, we can conclude that

$$\begin{aligned}
\left\| \boldsymbol{v}_2 \right\|_2 &\le \frac{1}{m} \left(\left\| \boldsymbol{\theta}_1 \right\|_2 + \left| \left(\boldsymbol{a}_l^{\mathrm{sgn}\top} \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right)^2 - \left| a_{l,1} \right|^2 \right| \left\| \boldsymbol{\theta}_2 \right\|_2 \right) \\
&\lesssim \frac{\log^2 m}{m} \left\| \boldsymbol{x}_{\perp}^{t,(l)} \right\|_2 + \frac{\sqrt{n \log^3 m}}{m} \left(\left| x_{\|}^{t,(l)} \right| + \left\| \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 \right).
\end{aligned}$$

- To bound $\boldsymbol{v}_1$, one first observes that

$$\begin{aligned}
&\nabla^2 f\left(\boldsymbol{x}(\tau)\right) \left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)} \right) - \nabla^2 f^{\mathrm{sgn}}\left(\tilde{\boldsymbol{x}}(\tau)\right) \left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right) \\
&= \underbrace{\nabla^2 f\left(\boldsymbol{x}(\tau)\right) \left(\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right)}_{:=\boldsymbol{w}_1(\tau)} + \underbrace{\left[\nabla^2 f\left(\boldsymbol{x}(\tau)\right) - \nabla^2 f\left(\tilde{\boldsymbol{x}}(\tau)\right) \right] \left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right)}_{:=\boldsymbol{w}_2(\tau)} \\
&\quad + \underbrace{\left[\nabla^2 f\left(\tilde{\boldsymbol{x}}(\tau)\right) - \nabla^2 f^{\mathrm{sgn}}\left(\tilde{\boldsymbol{x}}(\tau)\right) \right] \left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right)}_{:=\boldsymbol{w}_3(\tau)}.
\end{aligned}$$

  - The first term $\boldsymbol{w}_1(\tau)$ satisfies

$$\left\| \boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} - \eta \int_0^1 \boldsymbol{w}_1(\tau) \mathrm{d}\tau \right\|_2$$

$$
\begin{aligned}
&= \left\| \left\{ \boldsymbol{I} - \eta \int_0^1 \nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right) \mathrm{d}\tau \right\} \left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right) \right\|_2 \\
&\leq \left\| \boldsymbol{I} - \eta \int_0^1 \nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right) \mathrm{d}\tau \right\| \cdot \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 \\
&\leq \left\{ 1 + 3\eta \left(1 - \left\| \boldsymbol{x}^t \right\|_2^2 \right) + O\left(\eta \frac{1}{\log m} \right) + \eta \phi_1 \right\} \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 ,
\end{aligned}
$$

for some $|\phi_1| \ll \frac{1}{\log m}$, where the last line follows from the same argument as in (112).

– Regarding the second term $\boldsymbol{w}_2(\tau)$, it is seen that

$$
\begin{aligned}
\left\| \nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right) - \nabla^2 f\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right) \right\| &= \left\| \frac{3}{m} \sum_{i=1}^m \left[\left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 - \left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}(\tau) \right)^2 \right] \boldsymbol{a}_i \boldsymbol{a}_i^\top \right\| \\
&\leq \max_{1 \leq i \leq m} \left| \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 - \left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}(\tau) \right)^2 \right| \left\| \frac{3}{m} \sum_{i=1}^m \boldsymbol{a}_i \boldsymbol{a}_i^\top \right\| \\
&\leq \max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}\left(\tau\right) - \tilde{\boldsymbol{x}}\left(\tau\right) \right) \right| \max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}\left(\tau\right) + \tilde{\boldsymbol{x}}\left(\tau\right) \right) \right| \left\| \frac{3}{m} \sum_{i=1}^m \boldsymbol{a}_i \boldsymbol{a}_i^\top \right\| \\
&\lesssim \max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}\left(\tau\right) - \tilde{\boldsymbol{x}}\left(\tau\right) \right) \right| \sqrt{\log m},
\end{aligned} \tag{141}
$$

where the last line makes use of Lemma 13 as well as the incoherence conditions

$$
\max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}\left(\tau\right) + \tilde{\boldsymbol{x}}\left(\tau\right) \right) \right| \leq \max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \boldsymbol{x}\left(\tau\right) \right| + \max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}\left(\tau\right) \right| \lesssim \sqrt{\log m}. \tag{142}
$$

Note that

$$
\begin{aligned}
\boldsymbol{x}(\tau) - \tilde{\boldsymbol{x}}(\tau) &= \boldsymbol{x}^t + \tau \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^t \right) - \left[\boldsymbol{x}^{t,\mathrm{sgn}} + \tau \left(\boldsymbol{x}^{t,\mathrm{sgn},(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} \right) \right] \\
&= \left(1 - \tau\right) \left(\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} \right) + \tau \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right).
\end{aligned}
$$

This implies for all $0 \leq \tau \leq 1$,

$$
\left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}(\tau) - \tilde{\boldsymbol{x}}(\tau) \right) \right| \leq \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} \right) \right| + \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right) \right|.
$$

Moreover, the triangle inequality together with the Cauchy-Schwarz inequality tells us that

$$
\begin{aligned}
\left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right) \right| &\leq \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} \right) \right| + \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right) \right| \\
&\leq \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} \right) \right| + \left\| \boldsymbol{a}_i \right\|_2 \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2
\end{aligned}
$$

and

$$
\begin{aligned}
\left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} \right) \right| &\leq \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^{t,(i)} - \boldsymbol{x}^{t,\mathrm{sgn},(i)} \right) \right| + \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(i)} + \boldsymbol{x}^{t,\mathrm{sgn},(i)} \right) \right| \\
&\leq \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^{t,(i)} - \boldsymbol{x}^{t,\mathrm{sgn},(i)} \right) \right| + \left\| \boldsymbol{a}_i \right\|_2 \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(i)} + \boldsymbol{x}^{t,\mathrm{sgn},(i)} \right\|_2 .
\end{aligned}
$$

Combine the previous three inequalities to obtain

$$
\begin{aligned}
&\max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}\left(\tau\right) - \tilde{\boldsymbol{x}}\left(\tau\right) \right) \right| \leq \max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} \right) \right| + \max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right) \right| \\
&\quad \leq 2 \max_{1 \leq i \leq m} \left| \boldsymbol{a}_i^\top \left(\boldsymbol{x}^{t,(i)} - \boldsymbol{x}^{t,\mathrm{sgn},(i)} \right) \right| + 3 \max_{1 \leq i \leq m} \left\| \boldsymbol{a}_i \right\|_2 \max_{1 \leq l \leq m} \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 \\
&\quad \lesssim \sqrt{\log m} \max_{1 \leq i \leq m} \left\| \boldsymbol{x}^{t,(i)} - \boldsymbol{x}^{t,\mathrm{sgn},(i)} \right\|_2 + \sqrt{n} \max_{1 \leq l \leq m} \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 ,
\end{aligned}
$$

where the last inequality follows from the independence between $\boldsymbol{a}_i$ and $\boldsymbol{x}^{t,(i)} - \boldsymbol{x}^{t,\mathrm{sgn},(i)}$ and the fact (57). Substituting the above bound into (141) results in

$$\begin{aligned}
&\left\|\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right) - \nabla^2 f\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right)\right\| \\
&\quad\lesssim \log m \max_{1\le i\le m}\left\|\boldsymbol{x}^{t,(i)} - \boldsymbol{x}^{t,\mathrm{sgn},(i)}\right\|_2 + \sqrt{n\log m}\max_{1\le l\le m}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\quad\lesssim \log m\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 + \sqrt{n\log m}\max_{1\le l\le m}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2.
\end{aligned}$$

Here, we use the triangle inequality

$$\left\|\boldsymbol{x}^{t,(i)} - \boldsymbol{x}^{t,\mathrm{sgn},(i)}\right\|_2 \le \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 + \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(i)} + \boldsymbol{x}^{t,\mathrm{sgn},(i)}\right\|_2$$

and the fact $\log m \le \sqrt{n\log m}$. Consequently, we have the following bound for $\boldsymbol{w}_2\left(\tau\right)$:

$$\begin{aligned}
&\left\|\boldsymbol{w}_2(\tau)\right\|_2 \le \left\|\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right) - \nabla^2 f\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right)\right\| \cdot \left\|\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\lesssim \left\{\log m\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 + \sqrt{n\log m}\max_{1\le l\le m}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\right\}\left\|\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2.
\end{aligned}$$

– It remains to control $\boldsymbol{w}_3\left(\tau\right)$. To this end, one has

$$\begin{aligned}
\boldsymbol{w}_3(\tau) &= \frac{1}{m}\sum_{i=1}^m \underbrace{\left[3\left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}\left(\tau\right)\right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2\right]}_{:=\rho_i}\boldsymbol{a}_i\boldsymbol{a}_i^\top\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right) \\
&\quad - \frac{1}{m}\sum_{i=1}^m \underbrace{\left[3\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2 - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^\natural\right)^2\right]}_{:=\rho_i^{\mathrm{sgn}}}\boldsymbol{a}_i^{\mathrm{sgn}}\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right).
\end{aligned}$$

We consider the first entry of $\boldsymbol{w}_3\left(\tau\right)$, i.e. $w_{3,\parallel}\left(\tau\right)$, and the 2nd through the nth entries, $\boldsymbol{w}_{3,\perp}\left(\tau\right)$, separately. For the first entry $w_{3,\parallel}\left(\tau\right)$, we obtain

$$w_{3,\parallel}\left(\tau\right) = \frac{1}{m}\sum_{i=1}^m \rho_i\xi_i\left|a_{i,1}\right|\boldsymbol{a}_i^\top\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right) - \frac{1}{m}\sum_{i=1}^m \rho_i^{\mathrm{sgn}}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right). \tag{143}$$

Use the expansions

$$\begin{aligned}
\boldsymbol{a}_i^\top\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right) &= \xi_i\left|a_{i,1}\right|\left(x_\parallel^{t,\mathrm{sgn}} - x_\parallel^{t,\mathrm{sgn},(l)}\right) + \boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right) \\
\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right) &= \xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\left(x_\parallel^{t,\mathrm{sgn}} - x_\parallel^{t,\mathrm{sgn},(l)}\right) + \boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)
\end{aligned}$$

to further obtain

$$\begin{aligned}
w_{3,\parallel}\left(\tau\right) &= \frac{1}{m}\sum_{i=1}^m\left(\rho_i - \rho_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|^2\left(x_\parallel^{t,\mathrm{sgn}} - x_\parallel^{t,\mathrm{sgn},(l)}\right) + \frac{1}{m}\sum_{i=1}^m\left(\rho_i\xi_i - \rho_i^{\mathrm{sgn}}\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right) \\
&= \underbrace{\frac{1}{m}\sum_{i=1}^m\left(\rho_i - \rho_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|^2\left(x_\parallel^{t,\mathrm{sgn}} - x_\parallel^{t,\mathrm{sgn},(l)}\right)}_{:=\theta_1(\tau)} \\
&\quad + \underbrace{\frac{1}{m}\sum_{i=1}^m\left(\rho_i - \rho_i^{\mathrm{sgn}}\right)\left(\xi_i + \xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)}_{:=\theta_2(\tau)} \\
&\quad + \underbrace{\frac{1}{m}\sum_{i=1}^m \rho_i^{\mathrm{sgn}}\xi_i\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)}_{:=\theta_3(\tau)} - \underbrace{\frac{1}{m}\sum_{i=1}^m \rho_i\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)}_{:=\theta_4(\tau)}
\end{aligned}$$

The identity (122) reveals that

$$\rho_i - \rho_i^{\text{sgn}} = 6\left(\xi_i - \xi_i^{\text{sgn}}\right)|a_{i,1}|\,\tilde{x}_{\|}\left(\tau\right)\boldsymbol{a}_{i,\perp}^{\top}\tilde{\boldsymbol{x}}_{\perp}\left(\tau\right), \tag{144}$$

and hence

$$\theta_1\left(\tau\right) = \tilde{x}_{\|}\left(\tau\right)\cdot\frac{6}{m}\sum_{i=1}^{m}\left(\xi_i - \xi_i^{\text{sgn}}\right)|a_{i,1}|^3\,\boldsymbol{a}_{i,\perp}^{\top}\tilde{\boldsymbol{x}}_{\perp}\left(\tau\right)\left(x_{\|}^{t,\text{sgn}} - x_{\|}^{t,\text{sgn},(l)}\right),$$

which together with (130) implies

$$\begin{aligned}
\left|\theta_1\left(\tau\right)\right| &\leq 6\left|\tilde{x}_{\|}\left(\tau\right)\right|\left|x_{\|}^{t,\text{sgn}} - x_{\|}^{t,\text{sgn},(l)}\right|\left\|\tilde{\boldsymbol{x}}_{\perp}\left(\tau\right)\right\|_2\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i - \xi_i^{\text{sgn}}\right)|a_{i,1}|^3\,\boldsymbol{a}_{i,\perp}^{\top}\right\| \\
&\lesssim \sqrt{\frac{n\log^3 m}{m}}\left|\tilde{x}_{\|}\left(\tau\right)\right|\left|x_{\|}^{t,\text{sgn}} - x_{\|}^{t,\text{sgn},(l)}\right|\left\|\tilde{\boldsymbol{x}}_{\perp}\left(\tau\right)\right\|_2 \\
&\lesssim \sqrt{\frac{n\log^3 m}{m}}\left|\tilde{x}_{\|}\left(\tau\right)\right|\left|x_{\|}^{t,\text{sgn}} - x_{\|}^{t,\text{sgn},(l)}\right|,
\end{aligned}$$

where the penultimate inequality arises from (130) and the last inequality utilizes the fact that

$$\left\|\tilde{\boldsymbol{x}}_{\perp}\left(\tau\right)\right\|_2 \leq \left\|\boldsymbol{x}_{\perp}^{t,\text{sgn}}\right\|_2 + \left\|\boldsymbol{x}_{\perp}^{t,\text{sgn},(l)}\right\|_2 \lesssim 1.$$

Again, we can use (144) and the identity $\left(\xi_i - \xi_i^{\text{sgn}}\right)\left(\xi_i + \xi_i^{\text{sgn}}\right) = 0$ to deduce that

$$\theta_2\left(\tau\right) = 0.$$

When it comes to $\theta_3\left(\tau\right)$, we exploit the independence between ξ_i and $\rho_i^{\text{sgn}}|a_{i,1}|\,\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\text{sgn}} - \boldsymbol{x}_{\perp}^{t,\text{sgn},(l)}\right)$ and apply the Bernstein inequality (see Lemma 11) to obtain that with probability exceeding $1 - O\left(m^{-10}\right)$

$$\left|\theta_3\left(\tau\right)\right| \lesssim \frac{1}{m}\left(\sqrt{V_1\log m} + B_1\log m\right),$$

where

$$V_1 := \sum_{i=1}^{m}\left(\rho_i^{\text{sgn}}\right)^2|a_{i,1}|^2\left|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\text{sgn}} - \boldsymbol{x}_{\perp}^{t,\text{sgn},(l)}\right)\right|^2 \quad \text{and} \quad B_1 := \max_{1\leq i\leq m}\left|\rho_i^{\text{sgn}}\right||a_{i,1}|\left|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\text{sgn}} - \boldsymbol{x}_{\perp}^{t,\text{sgn},(l)}\right)\right|.$$

Combine the fact $\left|\rho_i^{\text{sgn}}\right| \lesssim \log m$ and Lemma 14 to see that

$$V_1 \lesssim \left(m\log^2 m\right)\left\|\boldsymbol{x}_{\perp}^{t,\text{sgn}} - \boldsymbol{x}_{\perp}^{t,\text{sgn},(l)}\right\|_2^2.$$

In addition, the facts $\left|\rho_i^{\text{sgn}}\right| \lesssim \log m$, (56) and (57) tell us that

$$B_1 \lesssim \sqrt{n\log^3 m}\left\|\boldsymbol{x}_{\perp}^{t,\text{sgn}} - \boldsymbol{x}_{\perp}^{t,\text{sgn},(l)}\right\|_2.$$

Continue the derivation to reach

$$\left|\theta_3\left(\tau\right)\right| \lesssim \left(\sqrt{\frac{\log^3 m}{m}} + \frac{\sqrt{n\log^5 m}}{m}\right)\left\|\boldsymbol{x}_{\perp}^{t,\text{sgn}} - \boldsymbol{x}_{\perp}^{t,\text{sgn},(l)}\right\|_2 \lesssim \sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\text{sgn}} - \boldsymbol{x}_{\perp}^{t,\text{sgn},(l)}\right\|_2, \tag{145}$$

provided that $m \gtrsim n\log^2 m$. This further allows us to obtain

$$\left|\theta_4\left(\tau\right)\right| = \left|\frac{1}{m}\sum_{i=1}^{m}\left[3\left(\boldsymbol{a}_i^{\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2 - \left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{\natural}\right)^2\right]\xi_i^{\text{sgn}}|a_{i,1}|\,\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\text{sgn}} - \boldsymbol{x}_{\perp}^{t,\text{sgn},(l)}\right)\right|$$

$$
\begin{aligned}
&\leq \left|\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2-\left|a_{i,1}\right|^2\right\}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right)\right| \\
&\quad+\left|\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^{\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2-3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2\right\}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)\right| \\
&\quad+\left|\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2-\left|a_{i,1}\right|^2\right\}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)\right| \\
&\lesssim \sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2+\sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2\left\|\boldsymbol{x}\left(\tau\right)-\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_2 \\
&\quad+\frac{1}{\log^{3/2} m}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}(l)}\right\|_2. \qquad (146)
\end{aligned}
$$

To justify the last inequality, we first use similar bounds as in (145) to show that with probability exceeding $1-O\left(m^{-10}\right)$,

$$
\left|\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}(\tau)\right)^2-\left|a_{i,1}\right|^2\right\}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right)\right|\lesssim\sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2.
$$

In addition, we can invoke the Cauchy-Schwarz inequality to get

$$
\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^{\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2-3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2\right\}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)\right| \\
&\quad\leq\sqrt{\left(\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^{\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2-3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2\right\}^2\left|a_{i,1}\right|^2\right)\left(\frac{1}{m}\sum_{i=1}^{m}\left|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)\right|^2\right)} \\
&\quad\lesssim\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\boldsymbol{a}_i^{\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2-\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2\right\}^2\left|a_{i,1}\right|^2}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2,
\end{aligned}
$$

where the last line arises from Lemma 13. For the remaining term in the expression above, we have

$$
\begin{aligned}
\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\boldsymbol{a}_i^{\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2-\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2\right\}^2\left|a_{i,1}\right|^2}&=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left|a_{i,1}\right|^2\left[\boldsymbol{a}_i^{\top}\left(\boldsymbol{x}\left(\tau\right)-\tilde{\boldsymbol{x}}\left(\tau\right)\right)\right]^2\left[\boldsymbol{a}_i^{\top}\left(\boldsymbol{x}\left(\tau\right)+\tilde{\boldsymbol{x}}\left(\tau\right)\right)\right]^2} \\
&\overset{\text{(i)}}{\lesssim}\sqrt{\frac{\log m}{m}\sum_{i=1}^{m}\left|a_{i,1}\right|^2\left[\boldsymbol{a}_i^{\top}\left(\boldsymbol{x}\left(\tau\right)-\tilde{\boldsymbol{x}}\left(\tau\right)\right)\right]^2} \\
&\overset{\text{(ii)}}{\lesssim}\sqrt{\log m}\left\|\boldsymbol{x}\left(\tau\right)-\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_2.
\end{aligned}
$$

Here, (i) makes use of the incoherence condition (142), whereas (ii) comes from Lemma 14. Regarding the last line in (146), we have

$$
\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2-\left|a_{i,1}\right|^2\right\}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)\right| \\
&\quad\leq\left\|\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2-\left|a_{i,1}\right|^2\right\}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\right\|_2\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2.
\end{aligned}
$$

Since ξ_i^{sgn} is independent of $\left\{3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2-\left|a_{i,1}\right|^2\right\}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}$, one can apply the Bernstein inequality (see Lemma 11) to deduce that

$$
\left\|\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}\left(\tau\right)\right)^2-\left|a_{i,1}\right|^2\right\}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}^{\top}\right\|_2\lesssim\frac{1}{m}\left(\sqrt{V_2\log m}+B_2\log m\right),
$$

where

$$V_2 := \sum_{i=1}^{m} \left\{3\left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau)\right)^2 - |a_{i,1}|^2\right\}^2 |a_{i,1}|^2 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{a}_{i,\perp} \lesssim mn\log^3 m;$$

$$B_2 := \max_{1\le i\le m} \left|3\left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau)\right)^2 - |a_{i,1}|^2\right| |a_{i,1}| \left\|\boldsymbol{a}_{i,\perp}\right\|_2 \lesssim \sqrt{n}\log^{3/2} m.$$

This further implies

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\left\{3\left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau)\right)^2 - |a_{i,1}|^2\right\}\xi_i^{\mathrm{sgn}} |a_{i,1}| \boldsymbol{a}_{i,\perp}^\top\right\|_2 \lesssim \sqrt{\frac{n\log^4 m}{m}} + \frac{\sqrt{n}\log^{5/2} m}{m} \lesssim \frac{1}{\log^{3/2} m},$$

as long as $m \gg n\log^7 m$. Take the previous bounds on $\theta_1(\tau)$, $\theta_2(\tau)$, $\theta_3(\tau)$ and $\theta_4(\tau)$ collectively to arrive at

$$\begin{aligned}
\left|w_{3,\|}(\tau)\right| &\lesssim \sqrt{\frac{n\log^3 m}{m}}\left|\tilde{x}_{\|}(\tau)\right|\left|x_{\|}^{t,\mathrm{sgn}} - x_{\|}^{t,\mathrm{sgn},(l)}\right| + \sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\quad + \sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,(l)}\right\|_2 + \sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2 \left\|\boldsymbol{x}(\tau) - \tilde{\boldsymbol{x}}(\tau)\right\|_2 \\
&\quad + \frac{1}{\log^{3/2} m}\left\|\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,(l)} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} + \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}(l)}\right\|_2 \\
&\lesssim \sqrt{\frac{n\log^3 m}{m}}\left|\tilde{x}_{\|}(\tau)\right|\left|x_{\|}^{t,\mathrm{sgn}} - x_{\|}^{t,\mathrm{sgn},(l)}\right| \\
&\quad + \sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,(l)}\right\|_2 + \sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2 \left\|\boldsymbol{x}(\tau) - \tilde{\boldsymbol{x}}(\tau)\right\|_2 \\
&\quad + \frac{1}{\log^{3/2} m}\left\|\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,(l)} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} + \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}(l)}\right\|_2,
\end{aligned}$$

where the last inequality follows from the triangle inequality

$$\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2 \le \left\|\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,(l)}\right\|_2 + \left\|\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,(l)} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} + \boldsymbol{x}_{\perp}^{t,\mathrm{sgn}(l)}\right\|_2$$

and the fact that $\sqrt{\frac{\log^3 m}{m}} \le \frac{1}{\log^{3/2} m}$ for m sufficiently large. Similar to (143), we have the following identity for the 2nd through the nth entries of $\boldsymbol{w}_3(\tau)$:

$$\begin{aligned}
\boldsymbol{w}_{3,\perp}(\tau) &= \frac{1}{m}\sum_{i=1}^{m}\rho_i \boldsymbol{a}_{i,\perp}\boldsymbol{a}_i^\top\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right) - \frac{1}{m}\sum_{i=1}^{m}\rho_i^{\mathrm{sgn}}\boldsymbol{a}_{i,\perp}\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right) \\
&= \frac{3}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^\top\tilde{\boldsymbol{x}}(\tau)\right)^2\xi_i - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}(\tau)\right)^2\xi_i^{\mathrm{sgn}}\right]|a_{i,1}|\,\boldsymbol{a}_{i,\perp}\left(x_{\|}^{t,\mathrm{sgn}} - x_{\|}^{t,\mathrm{sgn},(l)}\right) \\
&\quad + \frac{3}{m}\sum_{i=1}^{m}|a_{i,1}|^2\left(\xi_i - \xi_i^{\mathrm{sgn}}\right)|a_{i,1}|\,\boldsymbol{a}_{i,\perp}\left(x_{\|}^{t,\mathrm{sgn}} - x_{\|}^{t,\mathrm{sgn},(l)}\right) \\
&\quad + \frac{3}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^\top\tilde{\boldsymbol{x}}(\tau)\right)^2 - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}(\tau)\right)^2\right]\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right).
\end{aligned}$$

It is easy to check by Lemma 14 and the incoherence conditions $\left|\boldsymbol{a}_i^\top\tilde{\boldsymbol{x}}(\tau)\right| \lesssim \sqrt{\log m}\left\|\tilde{\boldsymbol{x}}(\tau)\right\|_2$ and $\left|\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}(\tau)\right| \lesssim \sqrt{\log m}\left\|\tilde{\boldsymbol{x}}(\tau)\right\|_2$ that

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^\top\tilde{\boldsymbol{x}}(\tau)\right)^2\xi_i\,|a_{i,1}|\,\boldsymbol{a}_{i,\perp} = 2\tilde{x}_1(\tau)\,\tilde{\boldsymbol{x}}_{\perp}(\tau) + O\left(\sqrt{\frac{n\log^3 m}{m}}\right),$$

and

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}(\tau)\right)^2\xi_i^{\mathrm{sgn}}|a_{i,1}|\,\boldsymbol{a}_{i,\perp}=2\tilde{x}_1(\tau)\,\tilde{\boldsymbol{x}}_\perp(\tau)+O\left(\sqrt{\frac{n\log^3 m}{m}}\right).$$

Besides, in view of (130), we have

$$\left\|\frac{3}{m}\sum_{i=1}^{m}|a_{i,1}|^2\left(\xi_i-\xi_i^{\mathrm{sgn}}\right)|a_{i,1}|\,\boldsymbol{a}_{i,\perp}\right\|_2\lesssim\sqrt{\frac{n\log^3 m}{m}}.$$

We are left with controlling $\left\|\frac{3}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^\top\tilde{\boldsymbol{x}}(\tau)\right)^2-\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}(\tau)\right)^2\right]\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right\|_2$. To this end, one can see from (144) that

$$\begin{aligned}
&\left\|\frac{3}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^\top\tilde{\boldsymbol{x}}(\tau)\right)^2-\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}(\tau)\right)^2\right]\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right\|_2\\
&\quad=\left\|\tilde{x}_\|(\tau)\cdot\frac{6}{m}\sum_{i=1}^{m}\left(\xi_i-\xi_i^{\mathrm{sgn}}\right)|a_{i,1}|\,\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\tilde{\boldsymbol{x}}_\perp(\tau)\,\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right\|\\
&\quad\le 12\max_{1\le i\le m}|a_{i,1}|\left|\tilde{x}_\|(\tau)\right|\max_{1\le i\le m}\left|\boldsymbol{a}_{i,\perp}^\top\tilde{\boldsymbol{x}}_\perp(\tau)\right|\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\right\|\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2\\
&\quad\lesssim\log m\left|\tilde{x}_\|(\tau)\right|\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2,
\end{aligned}$$

where the last relation arises from (56), the incoherence condition $\max_{1\le i\le m}\left|\boldsymbol{a}_{i,\perp}^\top\tilde{\boldsymbol{x}}_\perp(\tau)\right|\lesssim\sqrt{\log m}$ and Lemma 13. Hence the 2nd through the nth entries of $\boldsymbol{w}_3(\tau)$ obey

$$\left\|\boldsymbol{w}_{3,\perp}(\tau)\right\|_2\lesssim\sqrt{\frac{n\log^3 m}{m}}\left|x_\|^{t,\mathrm{sgn}}-x_\|^{t,\mathrm{sgn},(l)}\right|+\log m\left|\tilde{x}_\|(\tau)\right|\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2.$$

Combine the above estimates to arrive at

$$\begin{aligned}
\left\|\boldsymbol{w}_3(\tau)\right\|_2&\le\left|w_{3,\|}(\tau)\right|+\left\|\boldsymbol{w}_{3,\perp}(\tau)\right\|_2\\
&\le\log m\left|\tilde{x}_\|(\tau)\right|\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2+\sqrt{\frac{n\log^3 m}{m}}\left|x_\|^{t,\mathrm{sgn}}-x_\|^{t,\mathrm{sgn},(l)}\right|\\
&\quad+\sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_\perp^{t}-\boldsymbol{x}_\perp^{t,(l)}\right\|_2+\sqrt{\log m}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2\left\|\boldsymbol{x}(\tau)-\tilde{\boldsymbol{x}}(\tau)\right\|_2\\
&\quad+\frac{1}{\log^{3/2}m}\left\|\boldsymbol{x}_\perp^{t}-\boldsymbol{x}_\perp^{t,(l)}-\boldsymbol{x}_\perp^{t,\mathrm{sgn}}+\boldsymbol{x}_\perp^{t,\mathrm{sgn}(l)}\right\|_2.
\end{aligned}$$

- Putting together the preceding bounds on $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ ($\boldsymbol{w}_1(\tau)$, $\boldsymbol{w}_2(\tau)$ and $\boldsymbol{w}_3(\tau)$), we can deduce that

$$\begin{aligned}
&\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^{t+1,\mathrm{sgn}}+\boldsymbol{x}^{t+1,\mathrm{sgn},(l)}\right\|_2\\
&\quad=\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}-\eta\left(\int_0^1\boldsymbol{w}_1(\tau)\,\mathrm{d}\tau+\int_0^1\boldsymbol{w}_2(\tau)\,\mathrm{d}\tau+\int_0^1\boldsymbol{w}_3(\tau)\,\mathrm{d}\tau\right)-\eta\boldsymbol{v}_2\right\|_2\\
&\quad\le\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}-\eta\int_0^1\boldsymbol{w}_1(\tau)\,\mathrm{d}\tau\right\|_2+\eta\sup_{0\le\tau\le1}\left\|\boldsymbol{w}(\tau)\right\|_2+\eta\sup_{0\le\tau\le1}\left\|\boldsymbol{w}_3(\tau)\right\|_2+\eta\left\|\boldsymbol{v}_2\right\|_2\\
&\quad\le\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_1\right\}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2
\end{aligned}$$

$$+O\left(\eta\left\{\sqrt{n\log m}\max_{1\le l\le m}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}-\boldsymbol{x}^{t,(l)}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2+\log m\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2\right\}\left\|\boldsymbol{x}^{t,\mathrm{sgn}}-\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\right)$$
$$+O\left(\eta\log m\sup_{0\le\tau\le1}\left|\tilde{x}_{\|}(\tau)\right|\left\|\boldsymbol{x}^{t,\mathrm{sgn}}-\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\right)+O\left(\eta\sqrt{\frac{n\log^3 m}{m}}\left|x_{\|}^{t,\mathrm{sgn}}-x_{\|}^{t,\mathrm{sgn},(l)}\right|\right)$$
$$+O\left(\eta\sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2\right)+O\left(\eta\sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2\sup_{0\le\tau\le1}\left\|\boldsymbol{x}(\tau)-\tilde{\boldsymbol{x}}(\tau)\right\|_2\right).$$
$$+O\left(\eta\frac{\log^2 m}{m}\left\|\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2\right)+O\left(\eta\frac{\sqrt{n\log^3 m}}{m}\left(\left|x_{\|}^{t,(l)}\right|+\left\|\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\right)\right).\tag{147}$$

To simplify the preceding bound, we first make the following claim, whose proof is deferred to the end of this subsection.

Claim 1. For $t\le T_0$, the following inequalities hold:

$$\sqrt{n\log m}\left\|\boldsymbol{x}^{t,\mathrm{sgn}}-\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\ll\frac{1}{\log m};$$
$$\log m\sup_{0\le\tau\le1}\left|\tilde{x}_{\|}(\tau)\right|+\log m\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2+\sqrt{\log m}\sup_{0\le\tau\le1}\left\|\boldsymbol{x}(\tau)-\tilde{\boldsymbol{x}}(\tau)\right\|_2+\frac{\sqrt{n\log^3 m}}{m}\lesssim\alpha_t\log m;$$
$$\alpha_t\log m\ll\frac{1}{\log m}.$$

Armed with Claim 1, one can rearrange terms in (147) to obtain for some $|\phi_2|,|\phi_3|\ll\frac{1}{\log m}$

$$\begin{aligned}
&\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^{t+1,\mathrm{sgn}}+\boldsymbol{x}^{t+1,\mathrm{sgn},(l)}\right\|_2\\
&\quad\le\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^{t}\right\|_2^2\right)+\eta\phi_2\right\}\max_{1\le l\le m}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\\
&\qquad+\eta O\left(\log m\cdot\alpha_t+\sqrt{\frac{\log^3 m}{m}}+\frac{\log^2 m}{m}\right)\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}\right\|_2\\
&\qquad+\eta O\left(\sqrt{\frac{n\log^3 m}{m}}+\frac{\sqrt{n\log^3 m}}{m}\right)\left|x_{\|}^{t}-x_{\|}^{t,(l)}\right|+\eta\frac{\log^2 m}{m}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2\\
&\qquad+\eta O\left(\frac{\sqrt{n\log^3 m}}{m}\right)\left(\left|x_{\|}^{t}\right|+\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2\right)\\
&\quad\le\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^{t}\right\|_2^2\right)+\eta\phi_3\right\}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\\
&\qquad+O\left(\eta\log m\right)\cdot\alpha_t\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}\right\|_2\\
&\qquad+O\left(\eta\sqrt{\frac{n\log^3 m}{m}}\right)\left|x_{\|}^{t}-x_{\|}^{t,(l)}\right|+O\left(\eta\frac{\log^2 m}{m}\right)\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2\\
&\qquad+O\left(\eta\frac{\sqrt{n\log^3 m}}{m}\right)\left(\left|x_{\|}^{t}\right|+\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2\right).
\end{aligned}$$

Substituting in the hypotheses (40), we can arrive at

$$\begin{aligned}
&\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^{t+1,\mathrm{sgn}}+\boldsymbol{x}^{t+1,\mathrm{sgn},(l)}\right\|_2\\
&\quad\le\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^{t}\right\|_2^2\right)+\eta\phi_3\right\}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_4\frac{\sqrt{n\log^9 m}}{m}
\end{aligned}$$

$$
\begin{aligned}
&+ O\left(\eta \log m\right) \alpha_t \beta_t \left(1+\frac{1}{\log m}\right)^t C_1 \frac{\sqrt{n\log^5 m}}{m} \\
&+ O\left(\eta \sqrt{\frac{\log^3 m}{m}}\right) \beta_t \left(1+\frac{1}{\log m}\right)^t C_1 \frac{\sqrt{n\log^5 m}}{m} \\
&+ O\left(\eta \sqrt{\frac{n\log^3 m}{m}}\right) \alpha_t \left(1+\frac{1}{\log m}\right)^t C_2 \frac{\sqrt{n\log^{12} m}}{m} \\
&+ O\left(\eta \frac{\log^2 m}{m}\right) \beta_t + O\left(\eta \frac{\sqrt{n\log^3 m}}{m}\right) \alpha_t \\
&+ O\left(\eta \frac{\sqrt{n\log^3 m}}{m}\right) \alpha_t \left(1+\frac{1}{\log m}\right)^t C_3 \sqrt{\frac{n\log^5 m}{m}} \\
&\overset{\text{(i)}}{\leq} \left\{1 + 3\eta\left(1 - \left\|\boldsymbol{x}^t\right\|_2^2\right) + \eta\phi_4\right\} \alpha_t \left(1+\frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n\log^9 m}}{m} \\
&\overset{\text{(ii)}}{\leq} \alpha_{t+1} \left(1+\frac{1}{\log m}\right)^{t+1} C_4 \frac{\sqrt{n\log^9 m}}{m}
\end{aligned}
$$

for some $|\phi_4| \ll \frac{1}{\log m}$. Here, the last relation (ii) follows the same argument as in (116) and (i) holds true as long as

$$
\left(\log m\right) \alpha_t \beta_t \left(1+\frac{1}{\log m}\right)^t C_1 \frac{\sqrt{n\log^5 m}}{m} \ll \frac{1}{\log m} \alpha_t \left(1+\frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n\log^9 m}}{m}; \tag{148a}
$$

$$
\sqrt{\frac{n\log^3 m}{m}} \alpha_t \left(1+\frac{1}{\log m}\right)^t C_2 \frac{\sqrt{n\log^{12} m}}{m} \ll \frac{1}{\log m} \alpha_t \left(1+\frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n\log^9 m}}{m}; \tag{148b}
$$

$$
\frac{\log^2 m}{m} \beta_t \ll \frac{1}{\log m} \alpha_t \left(1+\frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n\log^9 m}}{m}; \tag{148c}
$$

$$
\frac{\sqrt{n\log^3 m}}{m} \alpha_t \left(1+\frac{1}{\log m}\right)^t C_3 \sqrt{\frac{n\log^5 m}{m}} \ll \frac{1}{\log m} \alpha_t \left(1+\frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n\log^9 m}}{m}; \tag{148d}
$$

$$
\frac{\sqrt{n\log^3 m}}{m} \alpha_t \ll \frac{1}{\log m} \alpha_t \left(1+\frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n\log^9 m}}{m}, \tag{148e}
$$

where we recall that $t \leq T_0 \lesssim \log n$. The first condition (148a) can be checked using $\beta_t \lesssim 1$ and the assumption that $C_4 > 0$ is sufficiently large. The second one is valid if $m \gg n\log^8 m$. In addition, the third condition follows from the relationship (see Lemma 1)

$$
\beta_t \lesssim \alpha_t \sqrt{n\log m}.
$$

It is also easy to see that the last two are both valid.

Proof of Claim 1. For the first claim, it is east to see from the triangle inequality that

$$
\begin{aligned}
&\sqrt{n\log m} \left\| \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 \\
&\leq \sqrt{n\log m} \left(\left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} \right\|_2 + \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} + \boldsymbol{x}^{t,\mathrm{sgn},(l)} \right\|_2 \right) \\
&\leq \sqrt{n\log m} \beta_t \left(1+\frac{1}{\log m}\right)^t C_1 \frac{\sqrt{n\log^5 m}}{m} + \sqrt{n\log m} \alpha_t \left(1+\frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n\log^9 m}}{m} \\
&\lesssim \frac{n\log^3 m}{m} + \frac{n\log^5 m}{m} \ll \frac{1}{\log m},
\end{aligned}
$$

as long as $m \gg n \log^6 m$. Here, we have invoked the upper bounds on α_t and β_t provided in Lemma 1. Regarding the second claim, we have

$$\begin{aligned}
\left|\tilde{x}_{\|}(\tau)\right| &\le \left|x_{\|}^{t,\mathrm{sgn}}\right| + \left|x_{\|}^{t,\mathrm{sgn},(l)}\right| \le 2\left|x_{\|}^{t,\mathrm{sgn}}\right| + \left|x_{\|}^{t,\mathrm{sgn}} - x_{\|}^{t,\mathrm{sgn},(l)}\right| \\
&\le 2\left|x_{\|}^{t}\right| + 2\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 + \left|x_{\|}^{t} - x_{\|}^{t,(l)}\right| + \left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\lesssim \alpha_t\left(1 + \sqrt{\frac{n\log^5 m}{m}} + \frac{\sqrt{n\log^{12} m}}{m} + \frac{\sqrt{n\log^9 m}}{m}\right) \lesssim \alpha_t,
\end{aligned}$$

as long as $m \gg n \log^5 m$. Similar arguments can lead us to conclude that the remaining terms on the left-hand side of the second inequality in the claim are bounded by $O(\alpha_t)$. The third claim is an immediate consequence of the fact $\alpha_t \ll \frac{1}{\log^5 m}$ (see Lemma 1). $\square$

H Proof of Lemma 8

Recall from Appendix C that

$$x_{\|}^{t+1} = \left\{1 + 3\eta\left(1 - \left\|\boldsymbol{x}^t\right\|_2^2\right) + O\left(\eta\sqrt{\frac{n\log^3 m}{m}}\right)\right\} x_{\|}^{t} + J_2 - J_4,$$

where J_2 and J_4 are defined respectively as

$$\begin{aligned}
J_2 &:= \eta\left[1 - 3\left(x_{\|}^{t}\right)^2\right] \cdot \frac{1}{m}\sum_{i=1}^{m} a_{i,1}^3 \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}; \\
J_4 &:= \eta \cdot \frac{1}{m}\sum_{i=1}^{m} \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right)^3 a_{i,1}.
\end{aligned}$$

Instead of resorting to the leave-one-out sequence $\{\boldsymbol{x}^{t,\mathrm{sgn}}\}$ as in Appendix C, we can directly apply Lemma 12 and the incoherence condition (49a) to obtain

$$\begin{aligned}
|J_2| &\le \eta\left|1 - 3\left(x_{\|}^{t}\right)^2\right| \left|\frac{1}{m}\sum_{i=1}^{m} a_{i,1}^3 \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right| \ll \eta\frac{1}{\log^6 m}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2 \ll \eta\frac{1}{\log m}\alpha_t; \\
|J_4| &\le \eta\left|\frac{1}{m}\sum_{i=1}^{m} \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right) a_{i,1}\right| \ll \eta\frac{1}{\log^6 m}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2^3 \ll \eta\frac{1}{\log m}\alpha_t
\end{aligned}$$

with probability at least $1 - O\left(m^{-10}\right)$, as long as $m \gg n \log^{13} m$. Here, the last relations come from the fact that $\alpha_t \ge \frac{c}{\log^5 m}$ (see Lemma 1). Combining the previous estimates gives

$$\alpha_{t+1} = \left\{1 + 3\eta\left(1 - \left\|\boldsymbol{x}^t\right\|_2^2\right) + \eta\zeta_t\right\}\alpha_t,$$

with $|\zeta_t| \ll \frac{1}{\log m}$. This finishes the proof.

I Proof of Lemma 9

In view of Appendix D, one has

$$\left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)}\right\|_2 \le \left\{1 + 3\eta\left(1 - \left\|\boldsymbol{x}^t\right\|_2^2\right) + \eta\phi_1\right\}\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_2 + O\left(\eta\frac{\sqrt{n\log^3 m}}{m}\left\|\boldsymbol{x}^t\right\|_2\right),$$

for some $|\phi_1| \ll \frac{1}{\log m}$, where we use the trivial upper bound

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

Under the hypotheses (48a), we can obtain

$$\begin{aligned}
\left\| \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)} \right\|_2 &\le \left\{ 1 + 3\eta \left(1 - \left\| \boldsymbol{x}^t \right\|_2^2 \right) + \eta\phi_1 \right\} \alpha_t \left(1 + \frac{1}{\log m} \right)^t C_6 \frac{\sqrt{n \log^{15} m}}{m} + O\left(\eta \frac{\sqrt{n \log^3 m}}{m} (\alpha_t + \beta_t) \right) \\
&\le \left\{ 1 + 3\eta \left(1 - \left\| \boldsymbol{x}^t \right\|_2^2 \right) + \eta\phi_2 \right\} \alpha_t \left(1 + \frac{1}{\log m} \right)^t C_6 \frac{\sqrt{n \log^{15} m}}{m} \\
&\le \alpha_{t+1} \left(1 + \frac{1}{\log m} \right)^{t+1} C_6 \frac{\sqrt{n \log^{15} m}}{m},
\end{aligned}$$

for some $|\phi_2| \ll \frac{1}{\log m}$, as long as η is sufficiently small and

$$\frac{\sqrt{n \log^3 m}}{m} (\alpha_t + \beta_t) \ll \frac{1}{\log m} \alpha_t \left(1 + \frac{1}{\log m} \right)^t C_6 \frac{\sqrt{n \log^{15} m}}{m}.$$

This is satisfied since, according to Lemma 1,

$$\frac{\sqrt{n \log^3 m}}{m} (\alpha_t + \beta_t) \lesssim \frac{\sqrt{n \log^3 m}}{m} \lesssim \frac{\sqrt{n \log^{13} m}}{m} \alpha_t \ll \frac{1}{\log m} \alpha_t \left(1 + \frac{1}{\log m} \right)^t C_6 \frac{\sqrt{n \log^{15} m}}{m},$$

as long as $C_6 > 0$ is sufficiently large.

J Proof of Lemma 12

Without loss of generality, it suffices to consider all the unit vectors $\boldsymbol{z}$ obeying $\|\boldsymbol{z}\|_2 = 1$. To begin with, for any given $\boldsymbol{z}$, we can express the quantities of interest as $\frac{1}{m} \sum_{i=1}^{m} \left(g_i(\boldsymbol{z}) - G(\boldsymbol{z}) \right)$, where $g_i(\boldsymbol{z})$ depends only on $\boldsymbol{z}$ and $\boldsymbol{a}_i$. Note that

$$g_i(\boldsymbol{z}) = a_{i,1}^{\theta_1} \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{z} \right)^{\theta_2}$$

for different $\theta_1, \theta_2 \in \{1, 2, 3, 4, 6\}$ in each of the cases considered herein. It can be easily verified from Gaussianality that in all of these cases, for any fixed *unit* vector $\boldsymbol{z}$ one has

$$\mathbb{E}\left[g_i^2(\boldsymbol{z}) \right] \lesssim \left(\mathbb{E}\left[|g_i(\boldsymbol{z})| \right] \right)^2; \tag{149}$$

$$\mathbb{E}\left[|g_i(\boldsymbol{z})| \right] \asymp 1; \tag{150}$$

$$\left| \mathbb{E}\left[g_i(\boldsymbol{z}) \mathbb{1}_{\left\{ |\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}| \le \beta \|\boldsymbol{z}\|_2, |a_{i,1}| \le 5\sqrt{\log m} \right\}} \right] - \mathbb{E}\left[g_i(\boldsymbol{z}) \right] \right| \le \frac{1}{n} \mathbb{E}\left[|g_i(\boldsymbol{z})| \right]. \tag{151}$$

In addition, on the event $\left\{ \max_{1 \le i \le m} \|\boldsymbol{a}_i\|_2 \le \sqrt{6n} \right\}$ which has probability at least $1 - O\left(m e^{-1.5n} \right)$, one has, for any fixed unit vectors $\boldsymbol{z}, \boldsymbol{z}_0$, that

$$|g_i(\boldsymbol{z}) - g_i(\boldsymbol{z}_0)| \le n^{\alpha} \|\boldsymbol{z} - \boldsymbol{z}_0\|_2 \tag{152}$$

for some parameter $\alpha = O(1)$ in all cases. In light of these properties, we will proceed by controlling $\frac{1}{m} \sum_{i=1}^{m} g_i(\boldsymbol{z}) - \mathbb{E}\left[g_i(\boldsymbol{z}) \right]$ in a unified manner.

We start by looking at any fixed vector $\boldsymbol{z}$ independent of $\{\boldsymbol{a}_i\}$. Recognizing that

$$\frac{1}{m} \sum_{i=1}^{m} g_i(\boldsymbol{z}) \mathbb{1}_{\left\{ |\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}| \le \beta \|\boldsymbol{z}\|_2, |a_{i,1}| \le 5\sqrt{\log m} \right\}} - \mathbb{E}\left[g_i(\boldsymbol{z}) \mathbb{1}_{\left\{ |\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}| \le \beta \|\boldsymbol{z}\|_2, |a_{i,1}| \le 5\sqrt{\log m} \right\}} \right]$$

is a sum of m i.i.d. random variables, one can thus apply the Bernstein inequality to obtain

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}-\mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right]\right|\geq\tau\right\}\\
&\quad\leq 2\exp\left(-\frac{\tau^2/2}{V+\tau B/3}\right),
\end{aligned}$$

where the two quantities V and B obey

$$V:=\frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\left[g_i^2\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right]\leq\frac{1}{m}\mathbb{E}\left[g_i^2\left(\boldsymbol{z}\right)\right]\lesssim\frac{1}{m}\left(\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right]\right)^2; \tag{153}$$

$$B:=\frac{1}{m}\max_{1\leq i\leq m}\left\{\left|g_i\left(\boldsymbol{z}\right)\right|\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right\}. \tag{154}$$

Here the penultimate relation of (153) follows from (149). Taking $\tau=\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right]$, we can deduce that

$$\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}-\mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right]\right|\leq\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right] \tag{155}$$

with probability exceeding $1-2\min\left\{\exp\left(-c_1 m\epsilon^2\right),\exp\left(-\frac{c_2\epsilon\mathbb{E}\left[\left|g_i(\boldsymbol{z})\right|\right]}{B}\right)\right\}$ for some constants $c_1,c_2>0$. In particular, when $m\epsilon^2/(n\log n)$ and $\epsilon\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right]/(Bn\log n)$ are both sufficiently large, the inequality (155) holds with probability exceeding $1-2\exp\left(-c_3 n\log n\right)$ for some constant $c_3>0$ sufficiently large.

We then move on to extending this result to a uniform bound. Let $\mathcal{N}_\theta$ be a θ-net of the unit sphere with cardinality $|\mathcal{N}_\theta|\leq\left(1+\frac{2}{\theta}\right)^n$ such that for any $\boldsymbol{z}$ on the unit sphere, one can find a point $\boldsymbol{z}_0\in\mathcal{N}_\theta$ such that $\|\boldsymbol{z}-\boldsymbol{z}_0\|_2\leq\theta$. Apply the triangle inequality to obtain

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}-\mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right]\right|\\
&\quad\leq\underbrace{\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\leq\beta\|\boldsymbol{z}_0\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}-\mathbb{E}\left[g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\leq\beta\|\boldsymbol{z}_0\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right]\right|}_{:=I_1}\\
&\qquad+\underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}-g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\leq\beta\|\boldsymbol{z}_0\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right]\right|}_{:=I_2},
\end{aligned}$$

where the second line arises from the fact that

$$\mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right]=\mathbb{E}\left[g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\leq\beta\|\boldsymbol{z}_0\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right].$$

With regard to the first term I_1, by the union bound, with probability at least $1-2\left(1+\frac{2}{\theta}\right)^n\exp\left(-c_3 n\log n\right)$, one has

$$I_1\leq\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}_0\right)\right|\right].$$

It remains to bound I_2. Denoting $\mathcal{S}_i=\left\{\boldsymbol{z}\mid\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\left\|\boldsymbol{z}\right\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}$, we have

$$\begin{aligned}
I_2&=\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\boldsymbol{z}\in\mathcal{S}_i\right\}}-g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\boldsymbol{z}_0\in\mathcal{S}_i\right\}}\right|\\
&\leq\left|\frac{1}{m}\sum_{i=1}^{m}\left(g_i\left(\boldsymbol{z}\right)-g_i\left(\boldsymbol{z}_0\right)\right)\mathbb{1}_{\left\{\boldsymbol{z}\in\mathcal{S}_i,\boldsymbol{z}_0\in\mathcal{S}_i\right\}}\right|+\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\boldsymbol{z}\in\mathcal{S}_i,\boldsymbol{z}_0\notin\mathcal{S}_i\right\}}\right|+\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\boldsymbol{z}\notin\mathcal{S}_i,\boldsymbol{z}_0\in\mathcal{S}_i\right\}}\right|
\end{aligned}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\left|g_i\left(\boldsymbol{z}\right)-g_i\left(\boldsymbol{z}_0\right)\right|+\frac{1}{m}\max_{1\leq i\leq m}\left|g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i\}}\right|\cdot\sum_{i=1}^{m}\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i,\boldsymbol{z}_0\notin\mathcal{S}_i\}}$$
$$+\frac{1}{m}\max_{1\leq i\leq m}\left|g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\{\boldsymbol{z}_0\in\mathcal{S}_i\}}\right|\cdot\sum_{i=1}^{m}\mathbb{1}_{\{\boldsymbol{z}\notin\mathcal{S}_i,\boldsymbol{z}_0\in\mathcal{S}_i\}}. \tag{156}$$

For the first term in (156), it follows from (152) that

$$\frac{1}{m}\sum_{i=1}^{m}\left|g_i\left(\boldsymbol{z}\right)-g_i\left(\boldsymbol{z}_0\right)\right|\leq n^{\alpha}\left\|\boldsymbol{z}-\boldsymbol{z}_0\right\|_2\leq n^{\alpha}\theta.$$

For the second term of (156), we have

$$\begin{aligned}
\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i,\boldsymbol{z}_0\notin\mathcal{S}_i\}} &\leq \mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta,\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\geq\beta\right\}}\\
&= \mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\right\}}\left(\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\geq\beta+\sqrt{6n}\theta\right\}}+\mathbb{1}_{\left\{\beta\leq\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|<\beta+\sqrt{6n}\theta\right\}}\right)\\
&= \mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\right\}}\mathbb{1}_{\left\{\beta\leq\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\leq\beta+\sqrt{6n}\theta\right\}} \qquad (157)\\
&\leq \mathbb{1}_{\left\{\beta\leq\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\leq\beta+\sqrt{6n}\theta\right\}}.
\end{aligned}$$

Here, the identity (157) holds due to the fact that

$$\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\right\}}\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\geq\beta+\sqrt{6n}\theta\right\}}=0;$$

in fact, under the condition $\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\geq\beta+\sqrt{6n}\theta$ one has

$$\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\geq\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|-\left|\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{z}-\boldsymbol{z}_0\right)\right|\geq\beta+\sqrt{6n}\theta-\left\|\boldsymbol{a}_{i,\perp}\right\|_2\left\|\boldsymbol{z}-\boldsymbol{z}_0\right\|_2>\beta+\sqrt{6n}\theta-\sqrt{6n}\theta\geq\beta,$$

which is contradictory to $\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta$. As a result, one can obtain

$$\sum_{i=1}^{m}\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i,\boldsymbol{z}_0\notin\mathcal{S}_i\}}\leq\sum_{i=1}^{m}\mathbb{1}_{\left\{\beta\leq\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\leq\beta+\sqrt{6n}\theta\right\}}\leq 2Cn\log n,$$

with probability at least $1-e^{-\frac{2}{3}Cn\log n}$ for a sufficiently large constant $C>0$, where the last inequality follows from the Chernoff bound (see Lemma 10). This together with the union bound reveals that with probability exceeding $1-\left(1+\frac{2}{\theta}\right)^{n}e^{-\frac{2}{3}Cn\log n}$,

$$\frac{1}{m}\max_{1\leq i\leq m}\left|g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i\}}\right|\cdot\sum_{i=1}^{m}\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i,\boldsymbol{z}_0\notin\mathcal{S}_i\}}\leq B\cdot 2Cn\log n$$

with B defined in (154). Similarly, one can show that

$$\frac{1}{m}\max_{1\leq i\leq m}\left|g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\{\boldsymbol{z}_0\in\mathcal{S}_i\}}\right|\cdot\sum_{i=1}^{m}\mathbb{1}_{\{\boldsymbol{z}\notin\mathcal{S}_i,\boldsymbol{z}_0\in\mathcal{S}_i\}}\leq B\cdot 2Cn\log n.$$

Combine the above bounds to reach that

$$I_1+I_2\leq\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}_0\right)\right|\right]+n^{\alpha}\theta+4B\cdot Cn\log n\leq 2\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right],$$

as long as

$$n^{\alpha}\theta\leq\frac{\epsilon}{2}\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right]\qquad\text{and}\qquad 4B\cdot Cn\log n\leq\frac{\epsilon}{2}\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right].$$

In view of the fact (150), one can take $\theta\asymp\epsilon n^{-\alpha}$ to conclude that

$$\left|\frac{1}{m}\sum_{i=1}^{m}g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}-\mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\leq 5\sqrt{\log m}\right\}}\right]\right|\leq 2\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right] \tag{158}$$

holds for all $\boldsymbol{z} \in \mathbb{R}^n$ with probability at least $1 - 2\exp\left(-c_4 n \log n\right)$ for some constant $c_4 > 0$, with the proviso that $\epsilon \geq \frac{1}{n}$ and that $\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right] / \left(Bn \log n\right)$ sufficiently large.

Further, we note that $\{\max_i |a_{i,1}| \leq 5\sqrt{\log m}\}$ occurs with probability at least $1 - O(m^{-10})$. Therefore, on an event of probability at least $1 - O(m^{-10})$, one has

$$\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right) = \frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right) \mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right| \leq \beta\|\boldsymbol{z}\|_2, |a_{i,1}| \leq 5\sqrt{\log m}\right\}} \tag{159}$$

for all $\boldsymbol{z} \in \mathbb{R}^{n-1}$ obeying $\max_i \left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right| \leq \beta \left\|\boldsymbol{z}\right\|_2$. On this event, one can use the triangle inequality to obtain

$$\begin{aligned}
\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right) - \mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\right]\right| &= \left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right) \mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right| \leq \beta\|\boldsymbol{z}\|_2, |a_{i,1}| \leq 5\sqrt{\log m}\right\}} - \mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\right]\right| \\
&\leq \left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right) \mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right| \leq \beta\|\boldsymbol{z}\|_2, |a_{i,1}| \leq 5\sqrt{\log m}\right\}} - \mathbb{E}\left[g_i\left(\boldsymbol{z}\right) \mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right| \leq \beta\|\boldsymbol{z}\|_2, |a_{i,1}| \leq 5\sqrt{\log m}\right\}}\right]\right| \\
&\quad + \left|\mathbb{E}\left[g_i\left(\boldsymbol{z}\right) \mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right| \leq \beta\|\boldsymbol{z}\|_2, |a_{i,1}| \leq 5\sqrt{\log m}\right\}}\right] - \mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\right]\right| \\
&\leq 2\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right] + \frac{1}{n}\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right] \\
&\leq 3\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right],
\end{aligned}$$

as long as $\epsilon > 1/n$, where the penultimate line follows from (151). This leads to the desired uniform upper bound for $\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right) - \mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\right]$, namely, with probability at least $1 - O\left(m^{-10}\right)$,

$$\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right) - \mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\right]\right| \leq 3\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right]$$

holds uniformly for all $\boldsymbol{z} \in \mathbb{R}^{n-1}$ obeying $\max_i \left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right| \leq \beta \left\|\boldsymbol{z}\right\|_2$, provided that

$$m\epsilon^2/(n\log n) \quad \text{and} \quad \epsilon\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right] / \left(Bn\log n\right)$$

are both sufficiently large (with B defined in (154)).

To finish up, we provide the bounds on B and the resulting sample complexity conditions for each case as follows.

- For $g_i\left(\boldsymbol{z}\right) = a_{i,1}^3 \boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}$, one has $B \lesssim \frac{1}{m}\beta \log^{\frac{3}{2}} m$, and hence we need $m \gg \max\left\{\frac{1}{\epsilon^2} n\log n,\ \frac{1}{\epsilon}\beta n \log^{\frac{5}{2}} m\right\}$;
- For $g_i\left(\boldsymbol{z}\right) = a_{i,1}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right)^3$, one has $B \lesssim \frac{1}{m}\beta^3 \log^{\frac{1}{2}} m$, and hence we need $m \gg \max\left\{\frac{1}{\epsilon^2} n\log n,\ \frac{1}{\epsilon}\beta^3 n \log^{\frac{3}{2}} m\right\}$
- For $g_i\left(\boldsymbol{z}\right) = a_{i,1}^2\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right)^2$, we have $B \lesssim \frac{1}{m}\beta^2 \log m$, and hence $m \gg \max\left\{\frac{1}{\epsilon^2} n\log n,\ \frac{1}{\epsilon}\beta^2 n \log^2 m\right\}$;
- For $g_i\left(\boldsymbol{z}\right) = a_{i,1}^6\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right)^2$, we have $B \lesssim \frac{1}{m}\beta^2 \log^3 m$, and hence $m \gg \max\left\{\frac{1}{\epsilon^2} n\log n,\ \frac{1}{\epsilon}\beta^2 n \log^4 m\right\}$;
- For $g_i\left(\boldsymbol{z}\right) = a_{i,1}^2\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right)^6$, one has $B \lesssim \frac{1}{m}\beta^6 \log m$, and hence $m \gg \max\left\{\frac{1}{\epsilon^2} n\log n,\ \frac{1}{\epsilon}\beta^6 n \log^2 m\right\}$;
- For $g_i\left(\boldsymbol{z}\right) = a_{i,1}^2\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right)^4$, one has $B \lesssim \frac{1}{m}\beta^4 \log m$, and hence $m \gg \max\left\{\frac{1}{\epsilon^2} n\log n,\ \frac{1}{\epsilon}\beta^4 n \log^2 m\right\}$.

Given that ϵ can be arbitrary quantity above $1/n$, we establish the advertised results.

K Proof of Lemma 14

Note that if the second claim (59) holds, we can readily use it to justify the first one (58) by observing that

$$\max_{1\le i\le m}\left|\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right| \le 5\sqrt{\log m}\left\|\boldsymbol{x}^\natural\right\|_2$$

holds with probability at least $1-O\left(m^{-10}\right)$. As a consequence, the proof is devoted to justifying the second claim in the lemma.

First, notice that it suffices to consider all $\boldsymbol{z}$'s with unit norm, i.e. $\|\boldsymbol{z}\|_2 = 1$. We can then apply the triangle inequality to obtain

$$\begin{aligned}\left\|\frac{1}{m}\sum_{i=1}^m\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^2\boldsymbol{a}_i\boldsymbol{a}_i^\top - \boldsymbol{I}_n - 2\boldsymbol{z}\boldsymbol{z}^\top\right\| &\le \underbrace{\left\|\frac{1}{m}\sum_{i=1}^m\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^2\boldsymbol{a}_i\boldsymbol{a}_i^\top \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top\boldsymbol{z}\right|\le c_2\sqrt{\log m}\right\}} - \left(\beta_1\boldsymbol{I}_n + \beta_2\boldsymbol{z}\boldsymbol{z}^\top\right)\right\|}_{:=\theta_1} \\ &\quad + \underbrace{\left\|\beta_1\boldsymbol{I}_n + \beta_2\boldsymbol{z}\boldsymbol{z}^\top - \left(\boldsymbol{I}_n + 2\boldsymbol{z}\boldsymbol{z}^\top\right)\right\|}_{:=\theta_2},\end{aligned}$$

where

$$\beta_1 := \mathbb{E}\left[\xi^2\,\mathbb{1}_{\left\{|\xi|\le c_2\sqrt{\log m}\right\}}\right] \qquad \text{and} \qquad \beta_2 := \mathbb{E}\left[\xi^4\,\mathbb{1}_{\left\{|\xi|\le c_2\sqrt{\log m}\right\}}\right] - \beta_1$$

with $\xi \sim N(0,1)$.

- For the second term θ_2, we can further bound it as follows

$$\begin{aligned}\theta_2 &\le \left\|\beta_1\boldsymbol{I}_n - \boldsymbol{I}_n\right\| + \left\|\beta_2\boldsymbol{z}\boldsymbol{z}^\top - 2\boldsymbol{z}\boldsymbol{z}^\top\right\| \\ &\le \left|\beta_1 - 1\right| + \left|\beta_2 - 2\right|,\end{aligned}$$

  which motivates us to bound $|\beta_1 - 1|$ and $|\beta_2 - 2|$. Towards this end, simple calculation yields

$$\begin{aligned}1-\beta_1 &= \sqrt{\frac{2}{\pi}}\cdot c_2\sqrt{\log m}e^{-\frac{c_2^2\log m}{2}} + \operatorname{erfc}\left(\frac{c_2\sqrt{\log m}}{2}\right) \\ &\overset{\text{(i)}}{\le} \sqrt{\frac{2}{\pi}}\cdot c_2\sqrt{\log m}e^{-\frac{c_2^2\log m}{2}} + \frac{1}{\sqrt{\pi}}\frac{2}{c_2\sqrt{\log m}}e^{-\frac{c_2^2\log m}{4}} \\ &\overset{\text{(ii)}}{\le} \frac{1}{m},\end{aligned}$$

  where (i) arises from the fact that for all $x>0$, $\operatorname{erfc}(x) \le \frac{1}{\sqrt{\pi}}\frac{1}{x}e^{-x^2}$ and (ii) holds as long as $c_2>0$ is sufficiently large. Similarly, for the difference $|\beta_2 - 2|$, one can easily show that

$$\left|\beta_2 - 2\right| \le \left|\mathbb{E}\left[\xi^4\,\mathbb{1}_{\left\{|\xi|\le c_2\sqrt{\log m}\right\}}\right] - 3\right| + \left|\beta_1 - 1\right| \le \frac{2}{m}. \tag{160}$$

  Take the previous two bounds collectively to reach

$$\theta_2 \le \frac{3}{m}.$$

- With regards to θ_1, we resort to the standard covering argument. First, fix some $\boldsymbol{x},\boldsymbol{z}\in\mathbb{R}^n$ with $\|\boldsymbol{x}\|_2 = \|\boldsymbol{z}\|_2 = 1$ and notice that

$$\frac{1}{m}\sum_{i=1}^m\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^2\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2\mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top\boldsymbol{z}\right|\le c_2\sqrt{\log m}\right\}} - \beta_1 - \beta_2\left(\boldsymbol{z}^\top\boldsymbol{x}\right)^2$$

  is a sum of m i.i.d. random variables with bounded sub-exponential norms. To see this, one has

$$\left\|\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^2\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2\mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top\boldsymbol{z}\right|\le c_2\sqrt{\log m}\right\}}\right\|_{\psi_1} \le c_2^2\log m\left\|\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2\right\|_{\psi_1} \le c_2^2\log m,$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm [Ver12]. This further implies that

$$\left\| \left(\boldsymbol{a}_i^\top \boldsymbol{z}\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}\right| \le c_2\sqrt{\log m}\right\}} - \beta_1 - \beta_2 \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^2 \right\|_{\psi_1} \le 2c_2^2 \log m.$$

Apply the Bernstein's inequality to show that for any $0 \le \epsilon \le 1$,

$$\mathbb{P}\left(\left| \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{z}\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}\right| \le c_2\sqrt{\log m}\right\}} - \beta_1 - \beta_2 \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^2 \right| \ge 2\epsilon c_2^2 \log m \right) \le 2\exp\left(-c\epsilon^2 m\right),$$

where $c > 0$ is some absolute constant. Taking $\epsilon \asymp \sqrt{\frac{n \log m}{m}}$ reveals that with probability exceeding $1 - 2\exp\left(-c_{10} n \log m\right)$ for some $c_{10} > 0$, one has

$$\left| \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{z}\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}\right| \le c_2\sqrt{\log m}\right\}} - \beta_1 - \beta_2 \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^2 \right| \lesssim c_2^2 \sqrt{\frac{n \log^3 m}{m}}. \tag{161}$$

One can then apply the covering argument to extend the above result to all unit vectors $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$. Let $\mathcal{N}_\theta$ be a θ-net of the unit sphere, which has cardinality at most $\left(1 + \frac{2}{\theta}\right)^n$. Then for every $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}$ with unit norm, we can find $\boldsymbol{x}_0, \boldsymbol{z}_0 \in \mathcal{N}_\theta$ such that $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \le \theta$ and $\|\boldsymbol{z} - \boldsymbol{z}_0\|_2 \le \theta$. The triangle inequality reveals that

$$\begin{aligned}
&\left| \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{z}\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}\right| \le c_2\sqrt{\log m}\right\}} - \beta_1 - \beta_2 \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^2 \right| \\
&\quad \le \underbrace{\left| \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}_0\right| \le c_2\sqrt{\log m}\right\}} - \beta_1 - \beta_2 \left(\boldsymbol{z}_0^\top \boldsymbol{x}_0\right)^2 \right|}_{:=I_1} + \underbrace{\beta_2 \left| \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{z}_0^\top \boldsymbol{x}_0\right)^2 \right|}_{:=I_2} \\
&\qquad + \underbrace{\left| \frac{1}{m}\sum_{i=1}^m \left[\left(\boldsymbol{a}_i^\top \boldsymbol{z}\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}\right| \le c_2\sqrt{\log m}\right\}} - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}_0\right| \le c_2\sqrt{\log m}\right\}} \right] \right|}_{:=I_3}.
\end{aligned}$$

Regarding I_1, one sees from (161) and the union bound that with probability at least $1 - 2(1 + \frac{2}{\theta})^{2n} \exp\left(-c_{10} n \log m\right)$, one has

$$I_1 \lesssim c_2^2 \sqrt{\frac{n \log^3 m}{m}}.$$

For the second term I_2, we can deduce from (160) that $\beta_2 \le 3$ and

$$\begin{aligned}
\left| \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{z}_0^\top \boldsymbol{x}_0\right)^2 \right| &= \left| \boldsymbol{z}^\top \boldsymbol{x} - \boldsymbol{z}_0^\top \boldsymbol{x}_0 \right| \left| \boldsymbol{z}^\top \boldsymbol{x} + \boldsymbol{z}_0^\top \boldsymbol{x}_0 \right| \\
&= \left| \left(\boldsymbol{z} - \boldsymbol{z}_0\right)^\top \boldsymbol{x} + \boldsymbol{z}_0 \left(\boldsymbol{x} - \boldsymbol{x}_0\right) \right| \left| \boldsymbol{z}^\top \boldsymbol{x} + \boldsymbol{z}_0^\top \boldsymbol{x}_0 \right| \\
&\le 2\left(\left\|\boldsymbol{z} - \boldsymbol{z}_0\right\|_2 + \left\|\boldsymbol{x} - \boldsymbol{x}_0\right\|_2 \right) \le 2\theta,
\end{aligned}$$

where the last line arises from the Cauchy-Schwarz inequality and the fact that $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}_0, \boldsymbol{z}_0$ are all unit norm vectors. This further implies

$$I_2 \le 6\theta.$$

Now we move on to control the last term I_3. Denoting

$$\mathcal{S}_i := \left\{ \boldsymbol{u} \mid \left|\boldsymbol{a}_i^\top \boldsymbol{u}\right| \le c_2\sqrt{\log m} \right\}$$

allows us to rewrite I_3 as

$$I_3 = \left| \frac{1}{m}\sum_{i=1}^m \left[\left(\boldsymbol{a}_i^\top \boldsymbol{z}\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 \mathbb{1}_{\left\{\boldsymbol{z} \in \mathcal{S}_i\right\}} - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0\right)^2 \mathbb{1}_{\left\{\boldsymbol{z}_0 \in \mathcal{S}_i\right\}} \right] \right|$$

$$
\begin{aligned}
&\leq \left| \frac{1}{m} \sum_{i=1}^{m} \left[\left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0 \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0 \right)^2 \right] \mathbb{1}_{\{ \boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \in \mathcal{S}_i \}} \right| \\
&\quad + \left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\{ \boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \notin \mathcal{S}_i \}} \right| + \left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0 \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0 \right)^2 \mathbb{1}_{\{ \boldsymbol{z}_0 \in \mathcal{S}_i, \boldsymbol{z} \notin \mathcal{S}_i \}} \right|. \qquad (162)
\end{aligned}
$$

Here the decomposition is similar to what we have done in (156). For the first term in (162), one has

$$
\begin{aligned}
\left| \frac{1}{m} \sum_{i=1}^{m} \left[\left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0 \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0 \right)^2 \right] \mathbb{1}_{\{ \boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \in \mathcal{S}_i \}} \right| &\leq \frac{1}{m} \sum_{i=1}^{m} \left| \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0 \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0 \right)^2 \right| \\
&\leq n^\alpha \theta,
\end{aligned}
$$

for some $\alpha = O(1)$. Here the last line follows from the smoothness of the function $g(\boldsymbol{x}, \boldsymbol{z}) = \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2$. Proceeding to the second term in (162), we see from (157) that

$$
\mathbb{1}_{\{ \boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \notin \mathcal{S}_i \}} \leq \mathbb{1}_{\left\{ c_2 \sqrt{\log m} \leq \left| \boldsymbol{a}_i^\top \boldsymbol{z}_0 \right| \leq c_2 \sqrt{\log m} + \sqrt{6n}\theta \right\}},
$$

which implies that

$$
\begin{aligned}
\left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\{ \boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \notin \mathcal{S}_i \}} \right| &\leq \max_{1 \leq i \leq m} \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \mathbb{1}_{\{ \boldsymbol{z} \in \mathcal{S}_i \}} \left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\{ \boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \notin \mathcal{S}_i \}} \right| \\
&\leq c_2^2 \log m \left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\left\{ c_2 \sqrt{\log m} \leq \left| \boldsymbol{a}_i^\top \boldsymbol{z}_0 \right| \leq c_2 \sqrt{\log m} + \sqrt{6n}\theta \right\}} \right|.
\end{aligned}
$$

With regard to the above quantity, we have the following claim.

Claim 2. With probability at least $1 - c_2 e^{-c_3 n \log m}$ for some constants $c_2, c_3 > 0$, one has

$$
\left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\left\{ c_2 \sqrt{\log m} \leq \left| \boldsymbol{a}_i^\top \boldsymbol{z}_0 \right| \leq c_2 \sqrt{\log m} + \sqrt{6n}\theta \right\}} \right| \lesssim \sqrt{\frac{n \log m}{m}}
$$

for all $\boldsymbol{x} \in \mathbb{R}^n$ with unit norm and for all $\boldsymbol{z}_0 \in \mathcal{N}_\theta$.

With this claim in place, we arrive at

$$
\left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\{ \boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \notin \mathcal{S}_i \}} \right| \lesssim c_2^2 \sqrt{\frac{n \log^3 m}{m}}
$$

with high probability. Similar arguments lead us to conclude that with high probability

$$
\left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0 \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0 \right)^2 \mathbb{1}_{\{ \boldsymbol{z}_0 \in \mathcal{S}_i, \boldsymbol{z} \notin \mathcal{S}_i \}} \right| \lesssim c_2^2 \sqrt{\frac{n \log^3 m}{m}}.
$$

Taking the above bounds collectively and setting $\theta \asymp m^{-\alpha-1}$ yield with high probability for all unit vectors $\boldsymbol{z}$'s and $\boldsymbol{x}$'s

$$
\left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_i^\top \boldsymbol{z} \right| \leq c_2 \sqrt{\log m} \right\}} - \beta_1 - \beta_2 \left(\boldsymbol{z}^\top \boldsymbol{x} \right)^2 \right| \lesssim c_2^2 \sqrt{\frac{n \log^3 m}{m}},
$$

which is equivalent to saying that

$$
\theta_1 \lesssim c_2^2 \sqrt{\frac{n \log^3 m}{m}}.
$$

The proof is complete by combining the upper bounds on θ_1 and θ_2, and the fact $\frac{1}{m} = o\left(\sqrt{\frac{n \log^3 m}{m}} \right)$.

Proof of Claim 2. We first apply the triangle inequality to get

$$\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}\right| \le \underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^\top\boldsymbol{x}_0\right)^2\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}\right|}_{:=J_1}$$
$$+\underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2-\left(\boldsymbol{a}_i^\top\boldsymbol{x}_0\right)^2\right]\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}\right|}_{:=J_2},$$

where $\boldsymbol{x}_0\in\mathcal{N}_\theta$ and $\|\boldsymbol{x}-\boldsymbol{x}_0\|_2\le\theta$. The second term can be controlled as follows

$$J_2\le\frac{1}{m}\sum_{i=1}^{m}\left|\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2-\left(\boldsymbol{a}_i^\top\boldsymbol{x}_0\right)^2\right|\le n^{O(1)}\theta,$$

where we utilize the smoothness property of the function $h(\boldsymbol{x})=\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2$. It remains to bound J_1, for which we first fix $\boldsymbol{x}_0$ and $\boldsymbol{z}_0$. Take the Bernstein inequality to get

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^\top\boldsymbol{x}_0\right)^2\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}-\mathbb{E}\left[\left(\boldsymbol{a}_i^\top\boldsymbol{x}_0\right)^2\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}\right]\right|\ge\tau\right)$$
$$\le 2e^{-cm\tau^2}$$

for some constant $c>0$ and any sufficiently small $\tau>0$. Taking $\tau\asymp\sqrt{\frac{n\log m}{m}}$ reveals that with probability exceeding $1-2e^{-Cn\log m}$ for some large enough constant $C>0$,

$$J_1\lesssim\mathbb{E}\left[\left(\boldsymbol{a}_i^\top\boldsymbol{x}_0\right)^2\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}\right]+\sqrt{\frac{n\log m}{m}}.$$

Regarding the expectation term, it follows from Cauchy-Schwarz that

$$\begin{aligned}\mathbb{E}\left[\left(\boldsymbol{a}_i^\top\boldsymbol{x}_0\right)^2\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}\right] &\le\sqrt{\mathbb{E}\left[\left(\boldsymbol{a}_i^\top\boldsymbol{x}_0\right)^4\right]}\sqrt{\mathbb{E}\left[\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}\right]}\\ &\asymp\mathbb{E}\left[\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}\right]\\ &\le 1/m,\end{aligned}$$

as long as θ is sufficiently small. Combining the preceding bounds with the union bound, we can see that with probability at least $1-2\left(1+\frac{2}{\theta}\right)^{2n}e^{-Cn\log m}$

$$J_1\lesssim\sqrt{\frac{n\log m}{m}}+\frac{1}{m}.$$

Picking $\theta\asymp m^{-c_1}$ for some large enough constant $c_1>0$, we arrive at with probability at least $1-c_2e^{-c_3n\log m}$

$$\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2\mathbb{1}_{\left\{c_2\sqrt{\log m}\le\left|\boldsymbol{a}_i^\top\boldsymbol{z}_0\right|\le c_2\sqrt{\log m}+\sqrt{6n\theta}\right\}}\right|\lesssim\sqrt{\frac{n\log m}{m}}$$

for all unit vectors $\boldsymbol{x}$'s and for all $\boldsymbol{z}_0\in\mathcal{N}_\theta$, where $c_2,c_3>0$ are some absolute constants. □

L Proof of Lemma 15

Recall that the Hessian matrix is given by

$$\nabla^2 f(\boldsymbol{z})=\frac{1}{m}\sum_{i=1}^{m}\left[3\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^2-\left(\boldsymbol{a}_i^\top\boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_i\boldsymbol{a}_i^\top.$$

Lemma 14 implies that with probability at least $1-O\left(m^{-10}\right)$,

$$\left\|\nabla^{2} f(\boldsymbol{z})-6 \boldsymbol{z} \boldsymbol{z}^{\top}-3\|\boldsymbol{z}\|_{2}^{2} \boldsymbol{I}_{n}+2 \boldsymbol{x}^{\natural} \boldsymbol{x}^{\natural \top}+\left\|\boldsymbol{x}^{\natural}\right\|_{2}^{2} \boldsymbol{I}_{n}\right\| \lesssim \sqrt{\frac{n \log ^{3} m}{m}} \max \left\{\|\boldsymbol{z}\|_{2}^{2},\left\|\boldsymbol{x}^{\natural}\right\|_{2}^{2}\right\} \tag{163}$$

holds simultaneously for all $\boldsymbol{z}$ obeying $\max_{1 \leq i \leq m}\left|\boldsymbol{a}_{i}^{\top} \boldsymbol{z}\right| \leq c_{0} \sqrt{\log m}\|\boldsymbol{z}\|_{2}$, with the proviso that $m \gg n \log ^{3} m$. This together with the fact $\left\|\boldsymbol{x}^{\natural}\right\|_{2}=1$ leads to

$$\begin{aligned}
-\nabla^{2} f(\boldsymbol{z}) & \succeq-6 \boldsymbol{z} \boldsymbol{z}^{\top}-\left\{3\|\boldsymbol{z}\|_{2}^{2}-1+O\left(\sqrt{\frac{n \log ^{3} m}{m}} \max \left\{\|\boldsymbol{z}\|_{2}^{2}, 1\right\}\right)\right\} \boldsymbol{I}_{n} \\
& \succeq-\left\{9\|\boldsymbol{z}\|_{2}^{2}-1+O\left(\sqrt{\frac{n \log ^{3} m}{m}} \max \left\{\|\boldsymbol{z}\|_{2}^{2}, 1\right\}\right)\right\} \boldsymbol{I}_{n} .
\end{aligned}$$

As a consequence, if we pick $0<\eta<\frac{c_{2}}{\max \left\{\|\boldsymbol{z}\|_{2}^{2}, 1\right\}}$ for $c_{2}>0$ sufficiently small, then $\boldsymbol{I}_{n}-\eta \nabla^{2} f(\boldsymbol{z}) \succeq \mathbf{0}$. This combined with (163) gives

$$\left\|\left(\boldsymbol{I}_{n}-\eta \nabla^{2} f(\boldsymbol{z})\right)-\left\{\left(1-3 \eta\|\boldsymbol{z}\|_{2}^{2}+\eta\right) \boldsymbol{I}_{n}+2 \eta \boldsymbol{x}^{\natural} \boldsymbol{x}^{\natural \top}-6 \eta \boldsymbol{z} \boldsymbol{z}^{\top}\right\}\right\| \lesssim \sqrt{\frac{n \log ^{3} m}{m}} \max \left\{\|\boldsymbol{z}\|_{2}^{2}, 1\right\} .$$

Additionally, it follows from (163) that

$$\begin{aligned}
\left\|\nabla^{2} f(\boldsymbol{z})\right\| & \leq\left\|6 \boldsymbol{z} \boldsymbol{z}^{\top}+3\|\boldsymbol{z}\|_{2}^{2} \boldsymbol{I}_{n}+2 \boldsymbol{x}^{\natural} \boldsymbol{x}^{\natural \top}+\left\|\boldsymbol{x}^{\natural}\right\|_{2}^{2} \boldsymbol{I}_{n}\right\|+O\left(\sqrt{\frac{n \log ^{3} m}{m}}\right) \max \left\{\|\boldsymbol{z}\|_{2}^{2},\left\|\boldsymbol{x}^{\natural}\right\|_{2}^{2}\right\} \\
& \leq 9\|\boldsymbol{z}\|_{2}^{2}+3+O\left(\sqrt{\frac{n \log ^{3} m}{m}}\right) \max \left\{\|\boldsymbol{z}\|_{2}^{2}, 1\right\} \\
& \leq 10\|\boldsymbol{z}\|_{2}^{2}+4
\end{aligned}$$

as long as $m \gg n \log ^{3} m$.

M Proof of Lemma 16

Note that when $t \lesssim \log n$, one naturally has

$$\left(1+\frac{1}{\log m}\right)^{t} \lesssim 1. \tag{164}$$

Regarding the first set of consequences (61), one sees via the triangle inequality that

$$\begin{aligned}
\max_{1 \leq l \leq m}\left\|\boldsymbol{x}^{t,(l)}\right\|_{2} \leq\left\|\boldsymbol{x}^{t}\right\|_{2}+\max_{1 \leq l \leq m}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}\right\|_{2} & \stackrel{\text{(i)}}{\leq} C_{5}+\beta_{t}\left(1+\frac{1}{\log m}\right)^{t} C_{1} \eta \frac{\sqrt{n \log ^{5} m}}{m} \\
& \stackrel{\text{(ii)}}{\leq} C_{5}+O\left(\frac{\sqrt{n \log ^{5} m}}{m}\right) \\
& \stackrel{\text{(iii)}}{\leq} 2 C_{5},
\end{aligned}$$

where (i) follows from the induction hypotheses (40a) and (40e). The second inequality (ii) holds true since $\beta_t \lesssim 1$ and (164). The last one (iii) is valid as long as $m \gg \sqrt{n \log^5 m}$. Similarly, for the lower bound, one can show that for each $1 \le l \le m$,

$$\begin{aligned}\left\|\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2 &\ge \left\|\boldsymbol{x}_{\perp}^{t}\right\|_2 - \left\|\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,(l)}\right\|_2 \\ &\ge \left\|\boldsymbol{x}_{\perp}^{t}\right\|_2 - \max_{1\le l\le m}\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)}\right\|_2 \\ &\ge c_5 - \beta_t\left(1+\frac{1}{\log m}\right)^t C_1\eta\frac{\sqrt{n\log^3 m}}{m} \ge \frac{c_5}{2},\end{aligned}$$

as long as $m \gg \sqrt{n \log^5 m}$. Using similar arguments ($\alpha_t \lesssim 1$), we can prove the lower and upper bounds for $\boldsymbol{x}^{t,\mathrm{sgn}}$ and $\boldsymbol{x}^{t,\mathrm{sgn},(l)}$.

For the second set of consequences (62), namely the incoherence consequences, first notice that it is sufficient to show that the inner product (for instance $|\boldsymbol{a}_l^\top \boldsymbol{x}^t|$) is upper bounded by $C_7 \log m$ in magnitude for some absolute constants $C_7 > 0$. To see this, suppose for now

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\boldsymbol{x}^t\right| \le C_7\sqrt{\log m}. \tag{165}$$

One can further utilize the lower bound on $\|\boldsymbol{x}^t\|_2$ to deduce that

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\boldsymbol{x}^t\right| \le \frac{C_7}{c_5}\sqrt{\log m}\left\|\boldsymbol{x}^t\right\|_2.$$

This justifies the claim that we only need to obtain bounds as in (165). Once again we can invoke the triangle inequality to deduce that with probability at least $1 - O\left(m^{-10}\right)$,

$$\begin{aligned}\max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\boldsymbol{x}^t\right| &\le \max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right)\right| + \max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right| \\ &\overset{\text{(i)}}{\le} \max_{1\le l\le m}\left\|\boldsymbol{a}_l\right\|_2 \max_{1\le l\le m}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 + \max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right| \\ &\overset{\text{(ii)}}{\lesssim} \sqrt{n}\beta_t\left(1+\frac{1}{\log m}\right)^t C_1\eta\frac{\sqrt{n\log^5 m}}{m} + \sqrt{\log m}\max_{1\le l\le m}\left\|\boldsymbol{x}^{t,(l)}\right\|_2 \\ &\lesssim \frac{n\log^{5/2} m}{m} + C_5\sqrt{\log m} \lesssim C_5\sqrt{\log m}.\end{aligned}$$

Here, the first relation (i) results from the Cauchy-Schwarz inequality and (ii) utilizes the induction hypothesis (40a), the fact (57) and the standard Gaussian concentration, namely, $\max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right| \lesssim \sqrt{\log m}\max_{1\le l\le m}\left\|\boldsymbol{x}^{t,(l)}\right\|_2$ with probability at least $1 - O\left(m^{-10}\right)$. The last line is a direct consequence of the fact (61a) established above and (164). In regard to the incoherence w.r.t. $\boldsymbol{x}^{t,\mathrm{sgn}}$, we resort to the leave-one-out sequence $\boldsymbol{x}^{t,\mathrm{sgn},(l)}$. Specifically, we have

$$\begin{aligned}\left|\boldsymbol{a}_l^\top\boldsymbol{x}^{t,\mathrm{sgn}}\right| &\le \left|\boldsymbol{a}_l^\top\boldsymbol{x}^t\right| + \left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^t\right)\right| \\ &\le \left|\boldsymbol{a}_l^\top\boldsymbol{x}^t\right| + \left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn},(l)} + \boldsymbol{x}^{t,(l)}\right)\right| + \left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t,\mathrm{sgn},(l)} - \boldsymbol{x}^{t,(l)}\right)\right| \\ &\lesssim \sqrt{\log m} + \sqrt{n}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_4\frac{\sqrt{n\log^9 m}}{m} + \sqrt{\log m} \\ &\lesssim \sqrt{\log m}.\end{aligned}$$

The remaining incoherence conditions can be obtained through similar arguments. For the sake of conciseness, we omit the details here.

With regard to the third set of consequences (63), we can directly use the induction hypothesis and obtain

$$\max_{1\le l\le m}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \le \beta_t\left(1+\frac{1}{\log m}\right)^t C_1\frac{\sqrt{n\log^3 m}}{m}$$

$$\lesssim \frac{\sqrt{n\log^3 m}}{m} \lesssim \frac{1}{\log m},$$

as long as $m \gg \sqrt{n\log^5 m}$. Apply similar arguments to get the claimed bound on $\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\|_2$. For the remaining one, we have

$$\begin{aligned}
\max_{1\le l\le m} \left| x_{\|}^{t,(l)} \right| &\le \max_{1\le l\le m} \left| x_{\|}^{t} \right| + \max_{1\le l\le m} \left| x_{\|}^{t,(l)} - x_{\|}^{t} \right| \\
&\le \alpha_t + \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_2 \eta \frac{\sqrt{n\log^{12} m}}{m} \\
&\le 2\alpha_t,
\end{aligned}$$

with the proviso that $m \gg \sqrt{n\log^{12} m}$.

N Proof of Theorem 3

A key observation is that in the proof of Theorem 2, we do not require independence between $\boldsymbol{x}^0$ and the data $\{\boldsymbol{a}_i, y_i\}_{1\le i\le m}$. Instead, what we really need are:

1. $\boldsymbol{x}^{0,\mathrm{sgn}}$ is independent of $\{\xi_i = \mathrm{sgn}(a_{i,1})\}_{1\le i\le m}$;
2. $\boldsymbol{x}^{0,(l)}$ is independent of $(\boldsymbol{a}_l, y_l)$ for all $1 \le l \le m$ and
3. $\boldsymbol{x}^{0,\mathrm{sgn},(l)}$ is independent of both $\{\xi_i\}_{1\le i\le m}$ and $\{\boldsymbol{a}_l, y_l\}$ for all $1 \le l \le m$.

With this observation in mind, one can see that the claim on the convergence holds true as long as the initialization $\boldsymbol{x}^0$ satisfies (14) and we can construct $\boldsymbol{x}^{0,\mathrm{sgn}}$, $\boldsymbol{x}^{0,(l)}$ and $\boldsymbol{x}^{0,\mathrm{sgn},(l)}$, which obey the required independence mentioned above as well as the base case specified in (40). In the following, we show that for

$$\boldsymbol{x}^0 = \sqrt{\frac{1}{m}\sum_{i=1}^m y_i} \cdot \boldsymbol{u},$$

where $\boldsymbol{u}$ is uniformly distributed over the unit sphere, the requirements can all be satisfied.

1. The first restriction (14) can be easily verified by concentration inequalities for spherical distribution and the fact that $\frac{1}{m}\sum_{i=1}^m y_i$ sharply concentrates around $\|\boldsymbol{x}^\natural\|_2^2$.

2. Next, we move on to demonstrating how to construct $\boldsymbol{x}^{0,\mathrm{sgn}}$, $\boldsymbol{x}^{0,(l)}$ and $\boldsymbol{x}^{0,\mathrm{sgn},(l)}$ with prescribed independence. In view of the initialization, we have

$$\boldsymbol{x}^0 = \lambda \cdot \boldsymbol{u},$$

where $\boldsymbol{u}$ is a unit vector uniformly distributed over the unit sphere in $\mathbb{R}^n$ and $\lambda = \sqrt{\sum_{i=1}^m y_i/m}$. Moreover, one has λ is independent of $\boldsymbol{u}$. This together with the fact that

$$y_i = \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2 = |a_{i,1}|^2$$

reveals that λ depends on $\{|a_{i,1}|\}_{1\le i\le m}$ only and $\boldsymbol{u}$ is independent of the data $\{\boldsymbol{a}_i, y_i\}_{1\le i\le m}$. Therefore, one can set

$$\boldsymbol{x}^{0,(l)} = \lambda^{(l)} \cdot \boldsymbol{u},$$

where $\boldsymbol{u}$ is the same vector as in $\boldsymbol{x}^0$ and $\lambda^{(l)} = \sqrt{\sum_{i:i\ne l}^m y_i/m}$. One can see from this construction that $\boldsymbol{x}^{0,(l)}$is independent of $\{\boldsymbol{a}_l, y_l\}$. Regarding $\boldsymbol{x}^{0,\mathrm{sgn}}$ and $\boldsymbol{x}^{0,\mathrm{sgn},(l)}$, we set

$$\boldsymbol{x}^{0,\mathrm{sgn}} = \boldsymbol{x}^0, \qquad \text{and} \qquad \boldsymbol{x}^{0,\mathrm{sgn},(l)} = \boldsymbol{x}^{0,(l)}.$$

Since $\boldsymbol{x}^0$ is independent of $\{\xi_i = \mathrm{sgn}(a_{i,1})\}_{1\le i\le m}$, so is $\boldsymbol{x}^{0,\mathrm{sgn}}$. The same reasoning can be applied to show independence between $\boldsymbol{x}^{0,\mathrm{sgn},(l)}$ and $\{\xi_i\}_{1\le i\le m}$ and $\{\boldsymbol{a}_l, y_l\}$.

3. We are left with checking the base case, i.e. (40):

(a) For the difference between $\boldsymbol{x}^0$ and $\boldsymbol{x}^{0,(l)}$, we have

$$\begin{aligned}\left\|\boldsymbol{x}^0-\boldsymbol{x}^{0,(l)}\right\|_2 &= \left\|\lambda\boldsymbol{u}-\lambda^{(l)}\boldsymbol{u}\right\|_2=\left|\lambda-\lambda^{(l)}\right| \\ &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}y_i}-\sqrt{\frac{1}{m}\sum_{i:i\neq l}^{m}y_i} \\ &= \frac{\frac{1}{m}y_l}{\sqrt{\frac{1}{m}\sum_{i=1}^{m}y_i}+\sqrt{\frac{1}{m}\sum_{i:i\neq l}^{m}y_i}},\end{aligned}$$

where the last relation holds due to the basic identity $\sqrt{a}-\sqrt{b}=(a-b)/(\sqrt{a}+\sqrt{b})$ for $a,b>0$. Noting that $\frac{1}{m}\sum_{i=1}^{m}y_i$ sharply concentrates around 1 and $|y_l|\lesssim\log m$ with high probability, one arrives at

$$\left\|\boldsymbol{x}^0-\boldsymbol{x}^{0,(l)}\right\|_2=\left|\lambda-\lambda^{(l)}\right|\lesssim\frac{\log m}{m}\leq\beta_0C_1\frac{\sqrt{n\log^5 m}}{m}.$$

This finishes the proof of (40a).

(b) The base case for (40b) can be easily deduced due to

$$\left|x_{\|}^0-x_{\|}^{0,(l)}\right|\leq\left\|\boldsymbol{x}^0-\boldsymbol{x}^{0,(l)}\right\|_2\lesssim\frac{\log m}{m}\leq\alpha_0C_2\frac{\sqrt{n\log^{12} m}}{m}.$$

(c) By construction, we have $\boldsymbol{x}^{0,\mathrm{sgn}}=\boldsymbol{x}^0$ and $\boldsymbol{x}^{0,\mathrm{sgn},(l)}=\boldsymbol{x}^{0,(l)}$. Therefore (40c) and (40d) trivially hold.

(d) The last two relations (40e) and (40f) can be verified using (14).

Combining all and repeating the proof of Theorem 2, we finish the proof of Theorem 3.