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*

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

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.

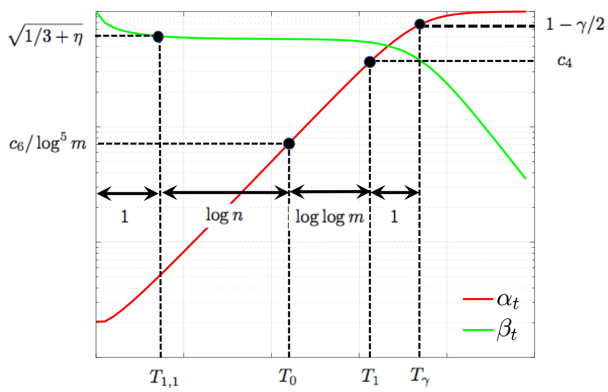

Figure 9: Illustration of the substages for the proof of Lemma 1.

– **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*

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

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

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

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,

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

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*

$$\alpha_t^2 + \beta_t^2 \le 2; \tag{73}$$

$$\frac{\alpha_{t+1}/\beta_{t+1}}{\alpha_t/\beta_t} \ge 1+\eta; \tag{74}$$

$$\alpha_{t+1} \ge \{1 - 3.1\eta\}\alpha_t; \tag{75}$$

$$\beta_{t+1} \ge \{1 - 5.1\eta\}\beta_t. \tag{76}$$

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 $T_{1,1} < t \le T_1$. □

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 $\beta_{t+1} \ge \{1 - 5.1\eta\}\,\beta_t$. □

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

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

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

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}a_{i,1}^4 - 3\right| \geq \tau\right) \leq e^2 e^{-c_1 m^{1/4}\tau^{1/2}}$$

  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

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

  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

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

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 $\|\boldsymbol{x}_\perp^t\|_2^2 \leq \|\boldsymbol{x}^t\|_2^2 \lesssim 1$.

- 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

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

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 $\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}\leq\left\|\boldsymbol{x}_{\perp}^{t}\right\|_{2}+\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_{2}$ and the fact that $\sqrt{\frac{\log m}{m}}\leq\sqrt{\frac{n\log^{3}m}{m}}$.

- 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

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

where R_1 is the residual term obeying

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

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

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

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

$$\sqrt{\frac{n\log^3 m}{m}} \ll \frac{1}{\log m} \tag{103a}$$

$$\sqrt{\frac{\log m}{m}}\beta_t \ll \frac{1}{\log m}\alpha_t \tag{103b}$$

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

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

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 $\beta_t \geq c_5$ (see (40e)). This finishes the proof.

D Proof of Lemma 4

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

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

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 $0 \leq \tau \leq 1$, 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

$$\|\boldsymbol{x}_{\perp}(\tau)\|_2 \geq \left\|\boldsymbol{x}_{\perp}^t\right\|_2 - \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \geq c_5 - \log^{-1} m \geq c_5/4. \tag{111}$$

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

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

This further allows one to derive

$$\begin{aligned}
&\left\|\left\{\boldsymbol{I}_n - \eta\nabla^2 f(\boldsymbol{x}(\tau))\right\}\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right)\right\|_2 \\
&\leq \left\|\left\{\left(1 - 3\eta\|\boldsymbol{x}(\tau)\|_2^2 + \eta\right)\boldsymbol{I}_n + 2\eta\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top} - 6\eta\boldsymbol{x}(\tau)\boldsymbol{x}(\tau)^{\top}\right\}\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right)\right\|_2 + O\left(\eta\sqrt{\frac{n\log^3 m}{m}}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2\right).
\end{aligned}$$

Moreover, we can apply the triangle inequality to get

$$\begin{aligned}
&\left\|\left\{\left(1 - 3\eta\|\boldsymbol{x}(\tau)\|_2^2 + \eta\right)\boldsymbol{I}_n + 2\eta\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top} - 6\eta\boldsymbol{x}(\tau)\boldsymbol{x}(\tau)^{\top}\right\}\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right)\right\|_2 \\
&\quad\leq \left\|\left\{\left(1 - 3\eta\|\boldsymbol{x}(\tau)\|_2^2 + \eta\right)\boldsymbol{I}_n - 6\eta\boldsymbol{x}(\tau)\boldsymbol{x}(\tau)^{\top}\right\}\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right)\right\|_2 + \left\|2\eta\boldsymbol{x}^{\natural}\boldsymbol{x}^{\natural\top}\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right)\right\|_2 \\
&\quad\overset{\text{(i)}}{=} \left\|\left\{\left(1 - 3\eta\|\boldsymbol{x}(\tau)\|_2^2 + \eta\right)\boldsymbol{I}_n - 6\eta\boldsymbol{x}(\tau)\boldsymbol{x}(\tau)^{\top}\right\}\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right)\right\|_2 + 2\eta\left|x_{\|}^t - x_{\|}^{t,(l)}\right| \\
&\quad\overset{\text{(ii)}}{\leq} \left(1 - 3\eta\|\boldsymbol{x}(\tau)\|_2^2 + \eta\right)\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 + 2\eta\left|x_{\|}^t - x_{\|}^{t,(l)}\right|,
\end{aligned}$$

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

$$\left(1 - 3\eta\|\boldsymbol{x}(\tau)\|_2^2 + \eta\right)\boldsymbol{I}_n - 6\eta\boldsymbol{x}(\tau)\boldsymbol{x}(\tau)^{\top} \succeq \boldsymbol{0},$$

as long as $\eta \leq 1/(18C_5)$. This further reveals

$$\begin{aligned}
&\left\|\left\{\boldsymbol{I}_n - \eta\nabla^2 f(\boldsymbol{x}(\tau))\right\}\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right)\right\|_2 \\
&\quad\leq \left\{1 + \eta\left(1 - 3\|\boldsymbol{x}(\tau)\|_2^2\right) + O\left(\eta\sqrt{\frac{n\log^3 m}{m}}\right)\right\}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 + 2\eta\left|x_{\|}^t - x_{\|}^{t,(l)}\right| \\
&\quad\overset{\text{(i)}}{\leq} \left\{1 + \eta\left(1 - 3\left\|\boldsymbol{x}^t\right\|_2^2\right) + O\left(\eta\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2\right) + O\left(\eta\sqrt{\frac{n\log^3 m}{m}}\right)\right\}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 + 2\eta\left|x_{\|}^t - x_{\|}^{t,(l)}\right| \\
&\quad\overset{\text{(ii)}}{\leq} \left\{1 + \eta\left(1 - 3\left\|\boldsymbol{x}^t\right\|_2^2\right) + \eta\phi_1\right\}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 + 2\eta\left|x_{\|}^t - x_{\|}^{t,(l)}\right|,
\end{aligned} \tag{112}$$

for some $|\phi_1| \ll \frac{1}{\log m}$, where (i) holds since for every $0 \leq \tau \leq 1$

$$\begin{aligned}
\|\boldsymbol{x}(\tau)\|_2^2 &\geq \left\|\boldsymbol{x}^t\right\|_2^2 - \left|\|\boldsymbol{x}(\tau)\|_2^2 - \left\|\boldsymbol{x}^t\right\|_2^2\right| \\
&\geq \left\|\boldsymbol{x}^t\right\|_2^2 - \left\|\boldsymbol{x}(\tau) - \boldsymbol{x}^t\right\|_2\left(\|\boldsymbol{x}(\tau)\|_2 + \left\|\boldsymbol{x}^t\right\|_2\right) \\
&\geq \left\|\boldsymbol{x}^t\right\|_2^2 - O\left(\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2\right),
\end{aligned} \tag{113}$$

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

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

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

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

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

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

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

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 $\omega_1(\tau)$ and $\omega_2(\tau)$ separately.

- For $\omega_1(\tau)$, 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

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

with the residual term r_1 obeying

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

- 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

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

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

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

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

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

which implies that

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

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

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

First note that

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

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

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

where we have used the fact that $\|\boldsymbol{x}_\perp(\tau)\|_2 \lesssim 1$ (see (110)). Regarding $\theta_2(\tau)$, one obtains

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

Apply the Cauchy-Schwarz inequality to reach

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

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

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

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

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

Similar to (118), one can obtain

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

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

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

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

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

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

which implies

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

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

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

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

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

where the last inequality utilizes the triangle inequality

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

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

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

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

- Combine the bounds to arrive at

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

$$+O\left(\eta\frac{\log^{2}m}{m}\left\Vert \boldsymbol{x}^{t,(l)}\right\Vert _{2}\right)+O\left(\eta\sqrt{\log m}\left\Vert \boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}+\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\Vert _{2}\right)$$

$$+O\left(\eta\sup_{0\leq\tau\leq1}\left\{ \left|x_{\|}\left(\tau\right)\right|+\sqrt{\log m}\left\Vert \boldsymbol{x}\left(\tau\right)-\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right\Vert _{2}+\sqrt{\frac{\log^{3}m}{m}}\right\} \left\Vert \boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\Vert _{2}\right).$$

To simplify the above bound, notice that for the last term, for any $t<T_0\lesssim\log n$ and $0\leq\tau\leq1$, one has

$$\left|x_{\|}\left(\tau\right)\right|\leq\left|x_{\|}^{t}\right|+\left|x_{\|}^{t,(l)}-x_{\|}^{t}\right|\leq\alpha_{t}+\alpha_{t}\left(1+\frac{1}{\log m}\right)^{t}C_{2}\frac{\sqrt{n\log^{12}m}}{m}\lesssim\alpha_{t},$$

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

$$\sqrt{\log m}\left\Vert \boldsymbol{x}\left(\tau\right)-\boldsymbol{x}^{\mathrm{sgn}}\left(\tau\right)\right\Vert _{2}\leq\sqrt{\log m}\left(\left\Vert \boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\Vert _{2}+\left\Vert \boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\Vert _{2}\right)$$
$$\lesssim\alpha_{t}\sqrt{\log m}\left(\sqrt{\frac{n\log^{5}m}{m}}+\frac{\sqrt{n\log^{9}m}}{m}\right)\lesssim\alpha_{t},$$

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

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

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

$$\left|x_{\|}^{t+1}-x_{\|}^{t+1,(l)}\right|\leq\left\{ 1+3\eta\left(1-\left\Vert \boldsymbol{x}^{t}\right\Vert _{2}^{2}\right)+\eta\phi_{1}\right\} \alpha_{t}\left(1+\frac{1}{\log m}\right)^{t}C_{2}\frac{\sqrt{n\log^{12}m}}{m}$$
$$+O\left(\frac{\eta\log^{2}m}{m}\left(\alpha_{t}+\beta_{t}\right)\right)+O\left(\eta\sqrt{\log m}\cdot\alpha_{t}\left(1+\frac{1}{\log m}\right)^{t}C_{4}\frac{\sqrt{n\log^{9}m}}{m}\right)$$
$$+O\left(\eta\alpha_{t}\beta_{t}\left(1+\frac{1}{\log m}\right)^{t}C_{1}\frac{\sqrt{n\log^{5}m}}{m}\right)$$
$$\overset{\text{(i)}}{\leq}\left\{ 1+3\eta\left(1-\left\Vert \boldsymbol{x}^{t}\right\Vert _{2}^{2}\right)+\eta\phi_{2}\right\} \alpha_{t}\left(1+\frac{1}{\log m}\right)^{t}C_{2}\frac{\sqrt{n\log^{12}m}}{m}$$
$$\overset{\text{(ii)}}{\leq}\alpha_{t+1}\left(1+\frac{1}{\log m}\right)^{t+1}C_{2}\frac{\sqrt{n\log^{12}m}}{m}$$

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

$$\frac{\log^{2}m}{m}\left(\alpha_{t}+\beta_{t}\right)\ll\frac{1}{\log m}\alpha_{t}C_{2}\frac{\sqrt{n\log^{12}m}}{m}\tag{124a}$$

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

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

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

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

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

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

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

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

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

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

and hence we can further upper bound (126) as

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

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

$$\le \left\{1 + 3\eta\left(1 - \left\|\boldsymbol{x}^t\right\|_2^2\right) + \eta\phi_1\right\}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2,$$

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 $f^{\mathrm{sgn}}(\cdot)$, 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 $\boldsymbol{a}_{i,\perp}^{\mathrm{sgn}} = \boldsymbol{a}_{i,\perp}$, $a_{i,1}^{\mathrm{sgn}} = \xi_i^{\mathrm{sgn}}|a_{i,1}|$ and $a_{i,1} = \xi_1|a_{i,1}|$. These taken together yield

$$\boldsymbol{a}_i^{\mathrm{sgn}}\boldsymbol{a}_i^{\mathrm{sgn}\top} - \boldsymbol{a}_i\boldsymbol{a}_i^{\top} = \left(\xi_i^{\mathrm{sgn}} - \xi_i\right)|a_{i,1}|\begin{bmatrix} 0 & \boldsymbol{a}_{i,\perp}^{\top} \\ \boldsymbol{a}_{i,\perp} & \boldsymbol{0}\end{bmatrix}, \tag{127}$$

and hence $\boldsymbol{r}_2$ can be rewritten as

$$\boldsymbol{r}_2 = \begin{bmatrix} \frac{1}{m}\sum_{i=1}^m\left(\xi_i^{\mathrm{sgn}} - \xi_i\right)|a_{i,1}|^3\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} \\ x_{\parallel}^{t,\mathrm{sgn}}\cdot\frac{1}{m}\sum_{i=1}^m\left(\xi_i^{\mathrm{sgn}} - \xi_i\right)|a_{i,1}|^3\boldsymbol{a}_{i,\perp}\end{bmatrix}. \tag{128}$$

For the first entry of $\boldsymbol{r}_2$, the triangle inequality gives

$$\begin{aligned}
\left|\frac{1}{m}\sum_{i=1}^m\left(\xi_i^{\mathrm{sgn}} - \xi_i\right)|a_{i,1}|^3\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right| &\le \underbrace{\left|\frac{1}{m}\sum_{i=1}^m|a_{i,1}|^3\xi_i\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right|}_{:=\phi_1} + \underbrace{\left|\frac{1}{m}\sum_{i=1}^m|a_{i,1}|^3\xi_i^{\mathrm{sgn}}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right|}_{:=\phi_2}\\
&\quad + \underbrace{\left|\frac{1}{m}\sum_{i=1}^m|a_{i,1}|^3\xi_i^{\mathrm{sgn}}\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t}\right)\right|}_{:=\phi_3}.
\end{aligned}$$

Regarding ϕ_1, we make use of the independence between ξ_i and $|a_{i,1}|^3\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}$ 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

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

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

$$\phi_3 \le \left\| \frac{1}{m} \sum_{i=1}^m \left| a_{i,1} \right|^3 \xi_i^{\text{sgn}} \boldsymbol{a}_{i,\perp} \right\|_2 \left\| \boldsymbol{x}_\perp^{t,\text{sgn}} - \boldsymbol{x}_\perp^t \right\|_2 \lesssim \sqrt{\frac{n \log^3 m}{m}} \left\| \boldsymbol{x}_\perp^{t,\text{sgn}} - \boldsymbol{x}_\perp^t \right\|_2 ,$$

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 $\left\| \boldsymbol{x}_\perp^{t,\text{sgn}} \right\|_2 \le \left\| \boldsymbol{x}_\perp^t \right\|_2 + \left\| \boldsymbol{x}_\perp^{t,\text{sgn}} - \boldsymbol{x}_\perp^t \right\|_2$ and the fact that $\sqrt{\frac{\log m}{m}} \le \sqrt{\frac{n \log^3 m}{m}}$. In addition, for the second through the nth entries of $\boldsymbol{r}_2$, one can again invoke Lemma 14 to obtain

$$\begin{aligned} \left\| \frac{1}{m} \sum_{i=1}^m \left| a_{i,1} \right|^3 \left(\xi_i^{\text{sgn}} - \xi_i \right) \boldsymbol{a}_{i,\perp} \right\|_2 &\le \left\| \frac{1}{m} \sum_{i=1}^m \left| a_{i,1} \right|^3 \xi_i^{\text{sgn}} \boldsymbol{a}_{i,\perp} \right\|_2 + \left\| \frac{1}{m} \sum_{i=1}^m \left| a_{i,1} \right|^3 \xi_i \boldsymbol{a}_{i,\perp} \right\|_2 \\ &\lesssim \sqrt{\frac{n \log^3 m}{m}}. \end{aligned} \tag{130}$$

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

$$\begin{aligned}
\boldsymbol{v}_3 &= \frac{1}{m}\sum_{i=1}^{m}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\left[\left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2 - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\boldsymbol{x}^{t,\mathrm{sgn}}\right)^2\right]\boldsymbol{a}_{i,\perp} \\
&= 2x_{\|}^{t,\mathrm{sgn}}\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i - \xi_i^{\mathrm{sgn}}\right)\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right)\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp} \\
&= 2x_{\|}^{t,\mathrm{sgn}}\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i\xi_i^{\mathrm{sgn}} - 1\right)\left|a_{i,1}\right|^2\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}. \qquad (131)
\end{aligned}$$

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

$$\boldsymbol{v}_6 = x_{\|}^{t,\text{sgn}} \frac{1}{m} \sum_{i=1}^{m} \left(\xi_i \xi_i^{\text{sgn}} - 1\right) |a_{i,1}|^2 \, \boldsymbol{a}_{i,\perp} \boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}}.$$

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

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \left\{ \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}}\right)^3 \boldsymbol{a}_{i,\perp}^{\text{sgn}} - \left(\boldsymbol{a}_i^{\top} \boldsymbol{x}^{t,\text{sgn}}\right)^3 \boldsymbol{a}_{i,\perp} \right\} \right\|_2 \lesssim \|\boldsymbol{v}_1\|_2 + \|\boldsymbol{v}_2\|_2 \lesssim \sqrt{\frac{n \log^3 m}{m}} \left| x_{\|}^{t,\text{sgn}} \right|.$$

Regarding the first entry of $\boldsymbol{r}_1$, one has

$$\begin{aligned}
&\left| \frac{1}{m} \sum_{i=1}^{m} \left\{ \left(\boldsymbol{a}_i^{\text{sgn}\top} \boldsymbol{x}^{t,\text{sgn}}\right)^3 a_{i,1}^{\text{sgn}} - \left(\boldsymbol{a}_i^{\top} \boldsymbol{x}^{t,\text{sgn}}\right)^3 a_{i,1} \right\} \right| \\
&\quad = \left| \frac{1}{m} \sum_{i=1}^{m} \left\{ \left(\xi_i^{\text{sgn}} |a_{i,1}| \, x_{\|}^{t,\text{sgn}} + \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^3 \xi_i^{\text{sgn}} |a_{i,1}| - \left(\xi_i |a_{i,1}| \, x_{\|}^{t,\text{sgn}} + \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^3 \xi_i |a_{i,1}| \right\} \right| \\
&\quad = \left| \frac{1}{m} \sum_{i=1}^{m} \left(\xi_i^{\text{sgn}} - \xi_i\right) |a_{i,1}| \left\{ 3 |a_{i,1}|^2 \left| x_{\|}^{t,\text{sgn}} \right|^2 \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}} + \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^3 \right\} \right|.
\end{aligned}$$

In view of the independence between ξ_i and $|a_{i,1}| \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^3$, from the Bernstein's inequality (see Lemma 11), we have that

$$\left| \frac{1}{m} \sum_{i=1}^{m} \xi_i |a_{i,1}| \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^3 \right| \lesssim \frac{1}{m} \left(\sqrt{V_2 \log m} + B_2 \log m \right)$$

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

$$V_2 := \sum_{i=1}^{m} |a_{i,1}|^2 \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^6 \qquad \text{and} \qquad B_2 := \max_{1 \le i \le m} |a_{i,1}| \left| \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}} \right|^3.$$

It is straightforward to check that $V_2 \lesssim m \left\| \boldsymbol{x}_{\perp}^{t,\text{sgn}} \right\|_2^6$ and $B_2 \lesssim \log^2 m \left\| \boldsymbol{x}_{\perp}^{t,\text{sgn}} \right\|_2^3$, which further implies

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

as long as $m \gg \log^5 m$. For the term involving ξ_i^{sgn}, we have

$$\frac{1}{m} \sum_{i=1}^{m} \xi_i^{\text{sgn}} |a_{i,1}| \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^3 = \underbrace{\frac{1}{m} \sum_{i=1}^{m} \xi_i^{\text{sgn}} |a_{i,1}| \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right)^3}_{:=\theta_1} + \underbrace{\frac{1}{m} \sum_{i=1}^{m} \xi_i^{\text{sgn}} |a_{i,1}| \left[\left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right)^3 - \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^3 \right]}_{:=\theta_2}.$$

Similarly one can obtain

$$|\theta_1| \lesssim \sqrt{\frac{\log m}{m}} \left\| \boldsymbol{x}_{\perp}^{t} \right\|_2^3.$$

Expand θ_2 using the elementary identity $a^3 - b^3 = (a - b)\left(a^2 + ab + b^2\right)$ to get

$$\begin{aligned}
\theta_2 &= \frac{1}{m} \sum_{i=1}^{m} \xi_i^{\text{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^{\top} \left(\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right) \left[\left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right)^2 + \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^2 + \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right) \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right) \right] \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t}\right)^2 \xi_i^{\text{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^{\top} \left(\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right) \\
&\quad + \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)^2 \xi_i^{\text{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^{\top} \left(\boldsymbol{x}_{\perp}^{t} - \boldsymbol{x}_{\perp}^{t,\text{sgn}}\right)
\end{aligned}$$

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

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

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

where the last inequality comes from Lemma 14. Combine the preceding bounds to reach

$$\left|\theta_{2}\right|\lesssim\sqrt{\frac{n\log^{3}m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{\mathrm{sgn}}\right\|_{2}+\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{\mathrm{sgn}}\right\|_{2}^{2}.$$

Applying the similar arguments as above we get

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

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

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

which can be further simplified to

$$\left\|\boldsymbol{r}_{1}\right\|_{2}\lesssim\sqrt{\frac{n\log^{3}m}{m}}\left|x_{\|}^{t}\right|+\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_{2}+\sqrt{\frac{n\log^{3}m}{m}}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{\mathrm{sgn}}\right\|_{2}+\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{\mathrm{sgn}}\right\|_{2}^{2}.$$

- Combine all of the above estimates to reach

$$\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,\mathrm{sgn}}\right\|_2 \leq \left\|\left\{\boldsymbol{I}-\eta\int_0^1 \nabla^2 f\left(\tilde{\boldsymbol{x}}(\tau)\right)\mathrm{d}\tau\right\}\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right)\right\|_2 + \eta\left\|\nabla f^{\mathrm{sgn}}\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)-\nabla f\left(\boldsymbol{x}^{t,\mathrm{sgn}}\right)\right\|_2$$

$$\leq \left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_2\right\}\left\|\boldsymbol{x}^t-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 + O\left(\eta\sqrt{\frac{\log m}{m}}\left\|\boldsymbol{x}_\perp^t\right\|_2\right)+\eta\sqrt{\frac{n\log^3 m}{m}}\left|x_\|^t\right|$$

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

$$\begin{aligned}
\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,\mathrm{sgn}}\right\|_2 &\leq \left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_2\right\}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}} \\
&\quad + \eta\sqrt{\frac{\log m}{m}}\beta_t + \eta\sqrt{\frac{n\log^3 m}{m}}\alpha_t \\
&\overset{\text{(i)}}{\leq} \left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_3\right\}\alpha_t\left(1+\frac{1}{\log m}\right)^t C_3\sqrt{\frac{n\log^5 m}{m}} \\
&\overset{\text{(ii)}}{\leq} \alpha_{t+1}\left(1+\frac{1}{\log m}\right)^{t+1} C_3\sqrt{\frac{n\log^5 m}{m}},
\end{aligned}$$

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 ($1 \leq l \leq m$). 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

$$\begin{aligned}
\nabla f\left(\boldsymbol{x}^t\right)-\nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right) &= \nabla f\left(\boldsymbol{x}^t\right)-\nabla f\left(\boldsymbol{x}^{t,(l)}\right)+\nabla f\left(\boldsymbol{x}^{t,(l)}\right)-\nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right) \\
&= \left(\int_0^1 \nabla^2 f\left(\boldsymbol{x}(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}^t-\boldsymbol{x}^{t,(l)}\right)+\frac{1}{m}\left[\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2-\left(\boldsymbol{a}_l^\top\boldsymbol{x}^\natural\right)^2\right]\boldsymbol{a}_l\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)},
\end{aligned} \tag{133}$$

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

$$\begin{aligned}
&\nabla f\left(\boldsymbol{x}^t\right) - \nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right) - \nabla f^{\text{sgn}}\left(\boldsymbol{x}^{t,\text{sgn}}\right) + \nabla f^{\text{sgn},(l)}\left(\boldsymbol{x}^{t,\text{sgn},(l)}\right) \\
&= \underbrace{\left(\int_0^1 \nabla^2 f\left(\boldsymbol{x}(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right) - \left(\int_0^1 \nabla^2 f^{\text{sgn}}\left(\tilde{\boldsymbol{x}}(\tau)\right)\mathrm{d}\tau\right)\left(\boldsymbol{x}^{t,\text{sgn}} - \boldsymbol{x}^{t,\text{sgn},(l)}\right)}_{:=\boldsymbol{v}_1} \\
&\quad + \underbrace{\frac{1}{m}\left[\left(\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t,(l)}\right)^2 - \left(\boldsymbol{a}_l^{\top}\boldsymbol{x}^{\natural}\right)^2\right]\boldsymbol{a}_l\boldsymbol{a}_l^{\top}\boldsymbol{x}^{t,(l)} - \frac{1}{m}\left[\left(\boldsymbol{a}_l^{\text{sgn}\top}\boldsymbol{x}^{t,\text{sgn},(l)}\right)^2 - \left(\boldsymbol{a}_l^{\text{sgn}\top}\boldsymbol{x}^{\natural}\right)^2\right]\boldsymbol{a}_l^{\text{sgn}}\boldsymbol{a}_l^{\text{sgn}\top}\boldsymbol{x}^{t,\text{sgn},(l)}}_{:=\boldsymbol{v}_2}.
\end{aligned} \tag{135}$$

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

$$\begin{aligned}
\left|\left(\boldsymbol{a}_l^{\text{sgn}\top}\boldsymbol{x}^{t,\text{sgn},(l)}\right)^2 - \left|a_{l,1}\right|^2\right| &\le \left(\boldsymbol{a}_l^{\text{sgn}\top}\boldsymbol{x}^{t,\text{sgn},(l)}\right)^2 + \left|a_{l,1}\right|^2 \\
&\lesssim \log m \cdot \left\|\boldsymbol{x}^{t,\text{sgn},(l)}\right\|_2^2 + \log m \lesssim \log m
\end{aligned} \tag{136}$$

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

$$\begin{aligned}
\left\|\boldsymbol{\theta}_2\right\|_2 &\le \left|\xi_l - \xi_l^{\text{sgn}}\right|\left|a_{l,1}\right|\left(\left|\boldsymbol{a}_{l,\perp}^{\top}\boldsymbol{x}_{\perp}^{t,(l)}\right| + \left|x_{\parallel}^{t,(l)}\right|\left\|\boldsymbol{a}_{l,\perp}\right\|_2\right) + \left\|\boldsymbol{a}_l^{\text{sgn}}\right\|_2\left|\boldsymbol{a}_l^{\text{sgn}\top}\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\text{sgn},(l)}\right)\right| \\
&\lesssim \sqrt{\log m}\left(\sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2 + \sqrt{n}\left|x_{\parallel}^{t,(l)}\right|\right) + \sqrt{n\log m}\left\|\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\text{sgn},(l)}\right\|_2
\end{aligned}$$

$$\lesssim \log m \left\|\boldsymbol{x}_{\perp}^{t,(l)}\right\|_2 + \sqrt{n\log m}\left(\left|x_{\|}^{t,(l)}\right| + \left\|\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\right). \tag{137}$$

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

$$
\begin{aligned}
\left\|\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right) - \nabla^2 f\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right)\right\| &= \left\| \frac{3}{m} \sum_{i=1}^m \left[\left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau)\right)^2 - \left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}(\tau)\right)^2 \right] \boldsymbol{a}_i \boldsymbol{a}_i^\top \right\| \\
&\leq \max_{1\leq i\leq m} \left| \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau)\right)^2 - \left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}(\tau)\right)^2 \right| \left\| \frac{3}{m} \sum_{i=1}^m \boldsymbol{a}_i \boldsymbol{a}_i^\top \right\| \\
&\leq \max_{1\leq i\leq m} \left|\boldsymbol{a}_i^\top \left(\boldsymbol{x}\left(\tau\right) - \tilde{\boldsymbol{x}}\left(\tau\right)\right)\right| \max_{1\leq i\leq m} \left|\boldsymbol{a}_i^\top \left(\boldsymbol{x}\left(\tau\right) + \tilde{\boldsymbol{x}}\left(\tau\right)\right)\right| \left\| \frac{3}{m} \sum_{i=1}^m \boldsymbol{a}_i \boldsymbol{a}_i^\top \right\| \\
&\lesssim \max_{1\leq i\leq m} \left|\boldsymbol{a}_i^\top \left(\boldsymbol{x}\left(\tau\right) - \tilde{\boldsymbol{x}}\left(\tau\right)\right)\right| \sqrt{\log m}, \qquad (141)
\end{aligned}
$$

where the last line makes use of Lemma 13 as well as the incoherence conditions

$$
\max_{1\leq i\leq m} \left|\boldsymbol{a}_i^\top \left(\boldsymbol{x}\left(\tau\right) + \tilde{\boldsymbol{x}}\left(\tau\right)\right)\right| \leq \max_{1\leq i\leq m} \left|\boldsymbol{a}_i^\top \boldsymbol{x}\left(\tau\right)\right| + \max_{1\leq i\leq m} \left|\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}\left(\tau\right)\right| \lesssim \sqrt{\log m}. \qquad (142)
$$

Note that

$$
\begin{aligned}
\boldsymbol{x}(\tau) - \tilde{\boldsymbol{x}}(\tau) &= \boldsymbol{x}^t + \tau\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^t\right) - \left[\boldsymbol{x}^{t,\mathrm{sgn}} + \tau\left(\boldsymbol{x}^{t,\mathrm{sgn},(l)} - \boldsymbol{x}^{t,\mathrm{sgn}}\right)\right] \\
&= (1-\tau)\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,\mathrm{sgn}}\right) + \tau\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right).
\end{aligned}
$$

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 $\boldsymbol{w}_2\left(\tau\right)$:

$$\begin{aligned}
&\|\boldsymbol{w}_2(\tau)\|_2 \le \left\|\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right) - \nabla^2 f\left(\tilde{\boldsymbol{x}}\left(\tau\right)\right)\right\| \cdot \left\|\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\lesssim \left\{\log m \left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 + \sqrt{n\log m}\max_{1\le l\le m}\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,(l)} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\right\}\left\|\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2.
\end{aligned}$$

– It remains to control $\boldsymbol{w}_3\left(\tau\right)$. To this end, one has

$$\begin{aligned}
\boldsymbol{w}_3(\tau) = &\frac{1}{m}\sum_{i=1}^m \underbrace{\left[3\left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}\left(\tau\right)\right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right)^2\right]}_{:=\rho_i} \boldsymbol{a}_i\boldsymbol{a}_i^\top\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right) \\
&- \frac{1}{m}\sum_{i=1}^m \underbrace{\left[3\left(\boldsymbol{a}_i^{\mathrm{sgn}\top} \tilde{\boldsymbol{x}}\left(\tau\right)\right)^2 - \left(\boldsymbol{a}_i^{\mathrm{sgn}\top} \boldsymbol{x}^\natural\right)^2\right]}_{:=\rho_i^{\mathrm{sgn}}} \boldsymbol{a}_i^{\mathrm{sgn}}\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right).
\end{aligned}$$

We consider the first entry of $\boldsymbol{w}_3\left(\tau\right)$, i.e. $w_{3,\|}\left(\tau\right)$, and the 2nd through the nth entries, $\boldsymbol{w}_{3,\perp}\left(\tau\right)$, separately. For the first entry $w_{3,\|}\left(\tau\right)$, we obtain

$$w_{3,\|}\left(\tau\right) = \frac{1}{m}\sum_{i=1}^m \rho_i\xi_i\left|a_{i,1}\right|\boldsymbol{a}_i^\top\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right) - \frac{1}{m}\sum_{i=1}^m \rho_i^{\mathrm{sgn}}\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_i^{\mathrm{sgn}\top}\left(\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right). \quad (143)$$

Use the expansions

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

to further obtain

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

The identity (122) reveals that

$$\rho_i - \rho_i^{\mathrm{sgn}} = 6\left(\xi_i - \xi_i^{\mathrm{sgn}}\right)|a_{i,1}|\,\tilde{x}_{\parallel}(\tau)\,\boldsymbol{a}_{i,\perp}^{\top}\tilde{\boldsymbol{x}}_{\perp}(\tau), \tag{144}$$

and hence

$$\theta_1(\tau) = \tilde{x}_{\parallel}(\tau)\cdot\frac{6}{m}\sum_{i=1}^{m}\left(\xi_i - \xi_i^{\mathrm{sgn}}\right)|a_{i,1}|^3\,\boldsymbol{a}_{i,\perp}^{\top}\tilde{\boldsymbol{x}}_{\perp}(\tau)\left(x_{\parallel}^{t,\mathrm{sgn}} - x_{\parallel}^{t,\mathrm{sgn},(l)}\right),$$

which together with (130) implies

$$\begin{aligned}
|\theta_1(\tau)| &\le 6\left|\tilde{x}_{\parallel}(\tau)\right|\left|x_{\parallel}^{t,\mathrm{sgn}} - x_{\parallel}^{t,\mathrm{sgn},(l)}\right|\left\|\tilde{\boldsymbol{x}}_{\perp}(\tau)\right\|_2\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i - \xi_i^{\mathrm{sgn}}\right)|a_{i,1}|^3\,\boldsymbol{a}_{i,\perp}^{\top}\right\| \\
&\lesssim \sqrt{\frac{n\log^3 m}{m}}\left|\tilde{x}_{\parallel}(\tau)\right|\left|x_{\parallel}^{t,\mathrm{sgn}} - x_{\parallel}^{t,\mathrm{sgn},(l)}\right|\left\|\tilde{\boldsymbol{x}}_{\perp}(\tau)\right\|_2 \\
&\lesssim \sqrt{\frac{n\log^3 m}{m}}\left|\tilde{x}_{\parallel}(\tau)\right|\left|x_{\parallel}^{t,\mathrm{sgn}} - x_{\parallel}^{t,\mathrm{sgn},(l)}\right|,
\end{aligned}$$

where the penultimate inequality arises from (130) and the last inequality utilizes the fact that

$$\left\|\tilde{\boldsymbol{x}}_{\perp}(\tau)\right\|_2 \le \left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}\right\|_2 + \left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2 \lesssim 1.$$

Again, we can use (144) and the identity $\left(\xi_i - \xi_i^{\mathrm{sgn}}\right)\left(\xi_i + \xi_i^{\mathrm{sgn}}\right) = 0$ to deduce that

$$\theta_2(\tau) = 0.$$

When it comes to $\theta_3(\tau)$, we exploit the independence between ξ_i and $\rho_i^{\mathrm{sgn}}|a_{i,1}|\,\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)$ and apply the Bernstein inequality (see Lemma 11) to obtain that with probability exceeding $1 - O\left(m^{-10}\right)$

$$|\theta_3(\tau)| \lesssim \frac{1}{m}\left(\sqrt{V_1\log m} + B_1\log m\right),$$

where

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

Combine the fact $|\rho_i^{\mathrm{sgn}}| \lesssim \log m$ and Lemma 14 to see that

$$V_1 \lesssim \left(m\log^2 m\right)\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2^2.$$

In addition, the facts $|\rho_i^{\mathrm{sgn}}| \lesssim \log m$, (56) and (57) tell us that

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

Continue the derivation to reach

$$|\theta_3(\tau)| \lesssim \left(\sqrt{\frac{\log^3 m}{m}} + \frac{\sqrt{n\log^5 m}}{m}\right)\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2 \lesssim \sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_2, \tag{145}$$

provided that $m \gtrsim n\log^2 m$. This further allows us to obtain

$$|\theta_4(\tau)| = \left|\frac{1}{m}\sum_{i=1}^{m}\left[3\left(\boldsymbol{a}_i^{\top}\tilde{\boldsymbol{x}}(\tau)\right)^2 - \left(\boldsymbol{a}_i^{\top}\boldsymbol{x}^{\natural}\right)^2\right]\xi_i^{\mathrm{sgn}}|a_{i,1}|\,\boldsymbol{a}_{i,\perp}^{\top}\left(\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}} - \boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right)\right|$$

$$\leq \left| \frac{1}{m} \sum_{i=1}^{m} \left\{ 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 - |a_{i,1}|^2 \right\} \xi_i^{\mathrm{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} \right) \right|$$

$$+ \left| \frac{1}{m} \sum_{i=1}^{m} \left\{ 3 \left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}(\tau) \right)^2 - 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 \right\} \xi_i^{\mathrm{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)} \right) \right|$$

$$+ \left| \frac{1}{m} \sum_{i=1}^{m} \left\{ 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 - |a_{i,1}|^2 \right\} \xi_i^{\mathrm{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)} \right) \right|$$

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

$$+ \frac{1}{\log^{3/2} m} \left\| \boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn}(l)} \right\|_2 . \tag{146}$$

To justify the last inequality, we first use similar bounds as in (145) to show that with probability exceeding $1 - O\left(m^{-10}\right)$,

$$\left| \frac{1}{m} \sum_{i=1}^{m} \left\{ 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 - |a_{i,1}|^2 \right\} \xi_i^{\mathrm{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} \right) \right| \lesssim \sqrt{\frac{\log^3 m}{m}} \left\| \boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} \right\|_2 .$$

In addition, we can invoke the Cauchy-Schwarz inequality to get

$$\left| \frac{1}{m} \sum_{i=1}^{m} \left\{ 3 \left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}(\tau) \right)^2 - 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 \right\} \xi_i^{\mathrm{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)} \right) \right|$$

$$\leq \sqrt{ \left(\frac{1}{m} \sum_{i=1}^{m} \left\{ 3 \left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}(\tau) \right)^2 - 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 \right\}^2 |a_{i,1}|^2 \right) \left(\frac{1}{m} \sum_{i=1}^{m} \left| \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)} \right) \right|^2 \right) }$$

$$\lesssim \sqrt{ \frac{1}{m} \sum_{i=1}^{m} \left\{ \left(\boldsymbol{a}_i^\top \tilde{\boldsymbol{x}}(\tau) \right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 \right\}^2 |a_{i,1}|^2 } \left\| \boldsymbol{x}_\perp^{t,\mathrm{sgn}} - \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)} \right\|_2 ,$$

where the last line arises from Lemma 13. For the remaining term in the expression above, we have

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

$$\overset{\text{(i)}}{\lesssim} \sqrt{ \frac{\log m}{m} \sum_{i=1}^{m} |a_{i,1}|^2 \left[\boldsymbol{a}_i^\top \left(\boldsymbol{x}(\tau) - \tilde{\boldsymbol{x}}(\tau) \right) \right]^2 }$$

$$\overset{\text{(ii)}}{\lesssim} \sqrt{\log m} \left\| \boldsymbol{x}(\tau) - \tilde{\boldsymbol{x}}(\tau) \right\|_2 .$$

Here, (i) makes use of the incoherence condition (142), whereas (ii) comes from Lemma 14. Regarding the last line in (146), we have

$$\left| \frac{1}{m} \sum_{i=1}^{m} \left\{ 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 - |a_{i,1}|^2 \right\} \xi_i^{\mathrm{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^\top \left(\boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)} \right) \right|$$

$$\leq \left\| \frac{1}{m} \sum_{i=1}^{m} \left\{ 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 - |a_{i,1}|^2 \right\} \xi_i^{\mathrm{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^\top \right\|_2 \left\| \boldsymbol{x}_\perp^t - \boldsymbol{x}_\perp^{t,(l)} - \boldsymbol{x}_\perp^{t,\mathrm{sgn}} + \boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)} \right\|_2 .$$

Since ξ_i^{sgn} is independent of $\left\{ 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 - |a_{i,1}|^2 \right\} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^\top$, one can apply the Bernstein inequality (see Lemma 11) to deduce that

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \left\{ 3 \left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau) \right)^2 - |a_{i,1}|^2 \right\} \xi_i^{\mathrm{sgn}} |a_{i,1}| \, \boldsymbol{a}_{i,\perp}^\top \right\|_2 \lesssim \frac{1}{m} \left(\sqrt{V_2 \log m} + B_2 \log m \right),$$

where

$$V_2 := \sum_{i=1}^{m} \left\{3\left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau)\right)^2 - |a_{i,1}|^2\right\}^2 |a_{i,1}|^2 \boldsymbol{a}_{i,\perp}^\top \boldsymbol{a}_{i,\perp} \lesssim mn\log^3 m;$$

$$B_2 := \max_{1\leq i\leq m} \left|3\left(\boldsymbol{a}_i^\top \boldsymbol{x}(\tau)\right)^2 - |a_{i,1}|^2\right| |a_{i,1}| \left\|\boldsymbol{a}_{i,\perp}\right\|_2 \lesssim \sqrt{n}\log^{3/2} m.$$

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

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

and the fact that $\sqrt{\frac{\log^3 m}{m}} \leq \frac{1}{\log^{3/2} m}$ 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

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2\xi_i^{\mathrm{sgn}}\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}=2\tilde{x}_1\left(\tau\right)\tilde{\boldsymbol{x}}_\perp\left(\tau\right)+O\left(\sqrt{\frac{n\log^3 m}{m}}\right).$$

Besides, in view of (130), we have

$$\left\|\frac{3}{m}\sum_{i=1}^{m}\left|a_{i,1}\right|^2\left(\xi_i-\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}\right\|_2\lesssim\sqrt{\frac{n\log^3 m}{m}}.$$

We are left with controlling $\left\|\frac{3}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^\top\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2-\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2\right]\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right\|_2$. To this end, one can see from (144) that

$$\begin{aligned}
&\left\|\frac{3}{m}\sum_{i=1}^{m}\left[\left(\boldsymbol{a}_i^\top\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2-\left(\boldsymbol{a}_i^{\mathrm{sgn}\top}\tilde{\boldsymbol{x}}\left(\tau\right)\right)^2\right]\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right\|_2\\
&\quad=\left\|\tilde{x}_\parallel\left(\tau\right)\cdot\frac{6}{m}\sum_{i=1}^{m}\left(\xi_i-\xi_i^{\mathrm{sgn}}\right)\left|a_{i,1}\right|\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\tilde{\boldsymbol{x}}_\perp\left(\tau\right)\boldsymbol{a}_{i,\perp}^\top\left(\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right)\right\|\\
&\quad\le 12\max_{1\le i\le m}\left|a_{i,1}\right|\left|\tilde{x}_\parallel\left(\tau\right)\right|\max_{1\le i\le m}\left|\boldsymbol{a}_{i,\perp}^\top\tilde{\boldsymbol{x}}_\perp\left(\tau\right)\right|\left\|\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{a}_{i,\perp}\boldsymbol{a}_{i,\perp}^\top\right\|\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2\\
&\quad\lesssim\log m\left|\tilde{x}_\parallel\left(\tau\right)\right|\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2,
\end{aligned}$$

where the last relation arises from (56), the incoherence condition $\max_{1\le i\le m}\left|\boldsymbol{a}_{i,\perp}^\top\tilde{\boldsymbol{x}}_\perp\left(\tau\right)\right|\lesssim\sqrt{\log m}$ and Lemma 13. Hence the 2nd through the nth entries of $\boldsymbol{w}_3\left(\tau\right)$ obey

$$\left\|\boldsymbol{w}_{3,\perp}\left(\tau\right)\right\|_2\lesssim\sqrt{\frac{n\log^3 m}{m}}\left|x_\parallel^{t,\mathrm{sgn}}-x_\parallel^{t,\mathrm{sgn},(l)}\right|+\log m\left|\tilde{x}_\parallel\left(\tau\right)\right|\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2.$$

Combine the above estimates to arrive at

$$\begin{aligned}
\left\|\boldsymbol{w}_3\left(\tau\right)\right\|_2&\le\left|w_{3,\parallel}\left(\tau\right)\right|+\left\|\boldsymbol{w}_{3,\perp}\left(\tau\right)\right\|_2\\
&\le\log m\left|\tilde{x}_\parallel\left(\tau\right)\right|\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2+\sqrt{\frac{n\log^3 m}{m}}\left|x_\parallel^{t,\mathrm{sgn}}-x_\parallel^{t,\mathrm{sgn},(l)}\right|\\
&\quad+\sqrt{\frac{\log^3 m}{m}}\left\|\boldsymbol{x}_\perp^{t}-\boldsymbol{x}_\perp^{t,(l)}\right\|_2+\sqrt{\log m}\left\|\boldsymbol{x}_\perp^{t,\mathrm{sgn}}-\boldsymbol{x}_\perp^{t,\mathrm{sgn},(l)}\right\|_2\left\|\boldsymbol{x}\left(\tau\right)-\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_2\\
&\quad+\frac{1}{\log^{3/2} m}\left\|\boldsymbol{x}_\perp^{t}-\boldsymbol{x}_\perp^{t,(l)}-\boldsymbol{x}_\perp^{t,\mathrm{sgn}}+\boldsymbol{x}_\perp^{t,\mathrm{sgn}(l)}\right\|_2.
\end{aligned}$$

- Putting together the preceding bounds on $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ ($\boldsymbol{w}_1\left(\tau\right)$, $\boldsymbol{w}_2\left(\tau\right)$ and $\boldsymbol{w}_3\left(\tau\right)$), we can deduce that

$$\begin{aligned}
&\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^{t+1,\mathrm{sgn}}+\boldsymbol{x}^{t+1,\mathrm{sgn},(l)}\right\|_2\\
&=\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}-\eta\left(\int_0^1\boldsymbol{w}_1\left(\tau\right)\mathrm{d}\tau+\int_0^1\boldsymbol{w}_2\left(\tau\right)\mathrm{d}\tau+\int_0^1\boldsymbol{w}_3\left(\tau\right)\mathrm{d}\tau\right)-\eta\boldsymbol{v}_2\right\|_2\\
&\le\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}-\eta\int_0^1\boldsymbol{w}_1\left(\tau\right)\mathrm{d}\tau\right\|_2+\eta\sup_{0\le\tau\le1}\left\|\boldsymbol{w}\left(\tau\right)\right\|_2+\eta\sup_{0\le\tau\le1}\left\|\boldsymbol{w}_3\left(\tau\right)\right\|_2+\eta\left\|\boldsymbol{v}_2\right\|_2\\
&\le\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^t\right\|_2^2\right)+\eta\phi_1\right\}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2
\end{aligned}$$

$$+O\left(\eta\left\{\sqrt{n\log m}\max_{1\leq l\leq m}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}-\boldsymbol{x}^{t,(l)}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_{2}+\log m\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_{2}\right\}\left\|\boldsymbol{x}^{t,\mathrm{sgn}}-\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_{2}\right)$$
$$+O\left(\eta\log m\sup_{0\leq\tau\leq1}\left|\tilde{x}_{\|}\left(\tau\right)\right|\left\|\boldsymbol{x}^{t,\mathrm{sgn}}-\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_{2}\right)+O\left(\eta\sqrt{\frac{n\log^{3}m}{m}}\left|x_{\|}^{t,\mathrm{sgn}}-x_{\|}^{t,\mathrm{sgn},(l)}\right|\right)$$
$$+O\left(\eta\sqrt{\frac{\log^{3}m}{m}}\left\|\boldsymbol{x}_{\perp}^{t}-\boldsymbol{x}_{\perp}^{t,(l)}\right\|_{2}\right)+O\left(\eta\sqrt{\log m}\left\|\boldsymbol{x}_{\perp}^{t,\mathrm{sgn}}-\boldsymbol{x}_{\perp}^{t,\mathrm{sgn},(l)}\right\|_{2}\sup_{0\leq\tau\leq1}\left\|\boldsymbol{x}\left(\tau\right)-\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_{2}\right).$$
$$+O\left(\eta\frac{\log^{2}m}{m}\left\|\boldsymbol{x}_{\perp}^{t,(l)}\right\|_{2}\right)+O\left(\eta\frac{\sqrt{n\log^{3}m}}{m}\left(\left|x_{\|}^{t,(l)}\right|+\left\|\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_{2}\right)\right). \tag{147}$$

To simplify the preceding bound, we first make the following claim, whose proof is deferred to the end of this subsection.

Claim 1. For $t\leq T_0$, the following inequalities hold:

$$\sqrt{n\log m}\left\|\boldsymbol{x}^{t,\mathrm{sgn}}-\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_{2}\ll\frac{1}{\log m};$$
$$\log m\sup_{0\leq\tau\leq1}\left|\tilde{x}_{\|}\left(\tau\right)\right|+\log m\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_{2}+\sqrt{\log m}\sup_{0\leq\tau\leq1}\left\|\boldsymbol{x}\left(\tau\right)-\tilde{\boldsymbol{x}}\left(\tau\right)\right\|_{2}+\frac{\sqrt{n\log^{3}m}}{m}\lesssim\alpha_{t}\log m;$$
$$\alpha_{t}\log m\ll\frac{1}{\log m}.$$

Armed with Claim 1, one can rearrange terms in (147) to obtain for some $|\phi_2|,|\phi_3|\ll\frac{1}{\log m}$

$$\begin{aligned}
&\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^{t+1,\mathrm{sgn}}+\boldsymbol{x}^{t+1,\mathrm{sgn},(l)}\right\|_{2}\\
&\quad\leq\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^{t}\right\|_{2}^{2}\right)+\eta\phi_{2}\right\}\max_{1\leq l\leq m}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_{2}\\
&\qquad+\eta O\left(\log m\cdot\alpha_{t}+\sqrt{\frac{\log^{3}m}{m}}+\frac{\log^{2}m}{m}\right)\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}\right\|_{2}\\
&\qquad+\eta O\left(\sqrt{\frac{n\log^{3}m}{m}}+\frac{\sqrt{n\log^{3}m}}{m}\right)\left|x_{\|}^{t}-x_{\|}^{t,(l)}\right|+\eta\frac{\log^{2}m}{m}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_{2}\\
&\qquad+\eta O\left(\frac{\sqrt{n\log^{3}m}}{m}\right)\left(\left|x_{\|}^{t}\right|+\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_{2}\right)\\
&\quad\leq\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^{t}\right\|_{2}^{2}\right)+\eta\phi_{3}\right\}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^{t,\mathrm{sgn}}+\boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_{2}\\
&\qquad+O\left(\eta\log m\right)\cdot\alpha_{t}\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,(l)}\right\|_{2}\\
&\qquad+O\left(\eta\sqrt{\frac{n\log^{3}m}{m}}\right)\left|x_{\|}^{t}-x_{\|}^{t,(l)}\right|+O\left(\eta\frac{\log^{2}m}{m}\right)\left\|\boldsymbol{x}_{\perp}^{t}\right\|_{2}\\
&\qquad+O\left(\eta\frac{\sqrt{n\log^{3}m}}{m}\right)\left(\left|x_{\|}^{t}\right|+\left\|\boldsymbol{x}^{t}-\boldsymbol{x}^{t,\mathrm{sgn}}\right\|_{2}\right).
\end{aligned}$$

Substituting in the hypotheses (40), we can arrive at

$$\begin{aligned}
&\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^{t+1,\mathrm{sgn}}+\boldsymbol{x}^{t+1,\mathrm{sgn},(l)}\right\|_{2}\\
&\quad\leq\left\{1+3\eta\left(1-\left\|\boldsymbol{x}^{t}\right\|_{2}^{2}\right)+\eta\phi_{3}\right\}\alpha_{t}\left(1+\frac{1}{\log m}\right)^{t}C_{4}\frac{\sqrt{n\log^{9}m}}{m}
\end{aligned}$$

$$+ O\left(\eta \log m\right) \alpha_t \beta_t \left(1 + \frac{1}{\log m}\right)^t C_1 \frac{\sqrt{n \log^5 m}}{m}$$
$$+ O\left(\eta \sqrt{\frac{\log^3 m}{m}}\right) \beta_t \left(1 + \frac{1}{\log m}\right)^t C_1 \frac{\sqrt{n \log^5 m}}{m}$$
$$+ O\left(\eta \sqrt{\frac{n \log^3 m}{m}}\right) \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_2 \frac{\sqrt{n \log^{12} m}}{m}$$
$$+ O\left(\eta \frac{\log^2 m}{m}\right) \beta_t + O\left(\eta \frac{\sqrt{n \log^3 m}}{m}\right) \alpha_t$$
$$+ O\left(\eta \frac{\sqrt{n \log^3 m}}{m}\right) \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_3 \sqrt{\frac{n \log^5 m}{m}}$$
$$\overset{\text{(i)}}{\leq} \left\{1 + 3\eta\left(1 - \left\|\boldsymbol{x}^t\right\|_2^2\right) + \eta\phi_4\right\} \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n \log^9 m}}{m}$$
$$\overset{\text{(ii)}}{\leq} \alpha_{t+1} \left(1 + \frac{1}{\log m}\right)^{t+1} C_4 \frac{\sqrt{n \log^9 m}}{m}$$

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

$$(\log m) \alpha_t \beta_t \left(1 + \frac{1}{\log m}\right)^t C_1 \frac{\sqrt{n \log^5 m}}{m} \ll \frac{1}{\log m} \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n \log^9 m}}{m}; \tag{148a}$$
$$\sqrt{\frac{n \log^3 m}{m}} \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_2 \frac{\sqrt{n \log^{12} m}}{m} \ll \frac{1}{\log m} \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n \log^9 m}}{m}; \tag{148b}$$
$$\frac{\log^2 m}{m} \beta_t \ll \frac{1}{\log m} \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n \log^9 m}}{m}; \tag{148c}$$
$$\frac{\sqrt{n \log^3 m}}{m} \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_3 \sqrt{\frac{n \log^5 m}{m}} \ll \frac{1}{\log m} \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n \log^9 m}}{m}; \tag{148d}$$
$$\frac{\sqrt{n \log^3 m}}{m} \alpha_t \ll \frac{1}{\log m} \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n \log^9 m}}{m}, \tag{148e}$$

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

$$\sqrt{n \log m} \left\|\boldsymbol{x}^{t,\mathrm{sgn}} - \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2$$
$$\leq \sqrt{n \log m} \left(\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 + \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2\right)$$
$$\leq \sqrt{n \log m} \beta_t \left(1 + \frac{1}{\log m}\right)^t C_1 \frac{\sqrt{n \log^5 m}}{m} + \sqrt{n \log m} \alpha_t \left(1 + \frac{1}{\log m}\right)^t C_4 \frac{\sqrt{n \log^9 m}}{m}$$
$$\lesssim \frac{n \log^3 m}{m} + \frac{n \log^5 m}{m} \ll \frac{1}{\log m},$$

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

$$\begin{aligned}
\left|\tilde{x}_{\|}(\tau)\right| &\leq \left|x_{\|}^{t,\mathrm{sgn}}\right| + \left|x_{\|}^{t,\mathrm{sgn},(l)}\right| \leq 2\left|x_{\|}^{t,\mathrm{sgn}}\right| + \left|x_{\|}^{t,\mathrm{sgn}} - x_{\|}^{t,\mathrm{sgn},(l)}\right| \\
&\leq 2\left|x_{\|}^{t}\right| + 2\left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,\mathrm{sgn}}\right\|_2 + \left|x_{\|}^{t} - x_{\|}^{t,(l)}\right| + \left\|\boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^{t,\mathrm{sgn}} + \boldsymbol{x}^{t,\mathrm{sgn},(l)}\right\|_2 \\
&\lesssim \alpha_t \left(1 + \sqrt{\frac{n \log^5 m}{m}} + \frac{\sqrt{n \log^{12} m}}{m} + \frac{\sqrt{n \log^9 m}}{m}\right) \lesssim \alpha_t,
\end{aligned}$$

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). □

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

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

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

$$\begin{aligned}
|J_2| &\leq \eta\left|1 - 3\big(x_{\|}^{t}\big)^2\right|\left|\frac{1}{m}\sum_{i=1}^{m} a_{i,1}^3 \boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right| \ll \eta\frac{1}{\log^6 m}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2 \ll \eta\frac{1}{\log m}\alpha_t; \\
|J_4| &\leq \eta\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{x}_{\perp}^{t}\right) a_{i,1}\right| \ll \eta\frac{1}{\log^6 m}\left\|\boldsymbol{x}_{\perp}^{t}\right\|_2^3 \ll \eta\frac{1}{\log m}\alpha_t
\end{aligned}$$

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 $\alpha_t \geq \frac{c}{\log^5 m}$ (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

$$\left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)}\right\|_2 \leq \left\{1 + 3\eta\left(1 - \left\|\boldsymbol{x}^t\right\|_2^2\right) + \eta\phi_1\right\}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 + O\left(\eta\frac{\sqrt{n \log^3 m}}{m}\left\|\boldsymbol{x}^t\right\|_2\right),$$

for some $|\phi_1| \ll \frac{1}{\log m}$, where we use the trivial upper bound

$$2\eta \left| x_{\|}^{t} - x_{\|}^{t,(l)} \right| \leq 2\eta \left\| \boldsymbol{x}^{t} - \boldsymbol{x}^{t,(l)} \right\|_2 .$$

Under the hypotheses (48a), we can obtain

$$\begin{aligned}
\left\| \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)} \right\|_2 &\leq \left\{ 1 + 3\eta \left(1 - \left\| \boldsymbol{x}^{t} \right\|_2^2 \right) + \eta\phi_1 \right\} \alpha_t \left(1 + \frac{1}{\log m} \right)^t C_6 \frac{\sqrt{n \log^{15} m}}{m} + O\left(\eta \frac{\sqrt{n \log^3 m}}{m} (\alpha_t + \beta_t) \right) \\
&\leq \left\{ 1 + 3\eta \left(1 - \left\| \boldsymbol{x}^{t} \right\|_2^2 \right) + \eta\phi_2 \right\} \alpha_t \left(1 + \frac{1}{\log m} \right)^t C_6 \frac{\sqrt{n \log^{15} m}}{m} \\
&\leq \alpha_{t+1} \left(1 + \frac{1}{\log m} \right)^{t+1} C_6 \frac{\sqrt{n \log^{15} m}}{m},
\end{aligned}$$

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

$$\mathbb{E}\left[g_i^2(\boldsymbol{z}) \right] \lesssim \left(\mathbb{E}\left[|g_i(\boldsymbol{z})| \right] \right)^2; \tag{149}$$

$$\mathbb{E}\left[|g_i(\boldsymbol{z})| \right] \asymp 1; \tag{150}$$

$$\left| \mathbb{E}\left[g_i(\boldsymbol{z}) \mathbb{1}_{\left\{ \left| \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{z} \right| \leq \beta \|\boldsymbol{z}\|_2, |a_{i,1}| \leq 5\sqrt{\log m} \right\}} \right] - \mathbb{E}\left[g_i(\boldsymbol{z}) \right] \right| \leq \frac{1}{n} \mathbb{E}\left[|g_i(\boldsymbol{z})| \right]. \tag{151}$$

In addition, on the event $\left\{ \max_{1 \leq i \leq m} \|\boldsymbol{a}_i\|_2 \leq \sqrt{6n} \right\}$ 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

$$\left| g_i(\boldsymbol{z}) - g_i(\boldsymbol{z}_0) \right| \leq n^{\alpha} \left\| \boldsymbol{z} - \boldsymbol{z}_0 \right\|_2 \tag{152}$$

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

$$\frac{1}{m} \sum_{i=1}^{m} g_i(\boldsymbol{z}) \mathbb{1}_{\left\{ \left| \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{z} \right| \leq \beta \|\boldsymbol{z}\|_2, |a_{i,1}| \leq 5\sqrt{\log m} \right\}} - \mathbb{E}\left[g_i(\boldsymbol{z}) \mathbb{1}_{\left\{ \left| \boldsymbol{a}_{i,\perp}^{\top} \boldsymbol{z} \right| \leq \beta \|\boldsymbol{z}\|_2, |a_{i,1}| \leq 5\sqrt{\log m} \right\}} \right]$$

is a sum of m i.i.d. random variables, one can thus apply the Bernstein inequality to obtain

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}} - \mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}}\right]\right| \ge \tau\right\} \\
&\quad \le 2\exp\left(-\frac{\tau^2/2}{V+\tau B/3}\right),
\end{aligned}$$

where the two quantities V and B obey

$$V := \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\left[g_i^2\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}}\right] \le \frac{1}{m}\mathbb{E}\left[g_i^2\left(\boldsymbol{z}\right)\right] \lesssim \frac{1}{m}\left(\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right]\right)^2; \tag{153}$$

$$B := \frac{1}{m}\max_{1\le i\le m}\left\{\left|g_i\left(\boldsymbol{z}\right)\right|\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}}\right\}. \tag{154}$$

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

$$\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}} - \mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}}\right]\right| \le \epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right] \tag{155}$$

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 $|\mathcal{N}_\theta| \le \left(1+\frac{2}{\theta}\right)^n$ such that for any $\boldsymbol{z}$ on the unit sphere, one can find a point $\boldsymbol{z}_0 \in \mathcal{N}_\theta$ such that $\|\boldsymbol{z}-\boldsymbol{z}_0\|_2 \le \theta$. Apply the triangle inequality to obtain

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}} - \mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}}\right]\right| \\
&\quad \le \underbrace{\left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\le\beta\|\boldsymbol{z}_0\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}} - \mathbb{E}\left[g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\le\beta\|\boldsymbol{z}_0\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}}\right]\right|}_{:=I_1} \\
&\qquad + \underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}} - g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\le\beta\|\boldsymbol{z}_0\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}}\right]\right|}_{:=I_2},
\end{aligned}$$

where the second line arises from the fact that

$$\mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\le\beta\|\boldsymbol{z}\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}}\right] = \mathbb{E}\left[g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}_0\right|\le\beta\|\boldsymbol{z}_0\|_2,|a_{i,1}|\le 5\sqrt{\log m}\right\}}\right].$$

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

$$I_1 \le \epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}_0\right)\right|\right].$$

It remains to bound I_2. Denoting $\mathcal{S}_i = \left\{\boldsymbol{z} \mid \left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right| \le \beta\|\boldsymbol{z}\|_2, |a_{i,1}| \le 5\sqrt{\log m}\right\}$, we have

$$\begin{aligned}
I_2 &= \left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i\}} - g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\{\boldsymbol{z}_0\in\mathcal{S}_i\}}\right| \\
&\le \left|\frac{1}{m}\sum_{i=1}^{m}\left(g_i\left(\boldsymbol{z}\right)-g_i\left(\boldsymbol{z}_0\right)\right)\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i,\boldsymbol{z}_0\in\mathcal{S}_i\}}\right| + \left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i,\boldsymbol{z}_0\notin\mathcal{S}_i\}}\right| + \left|\frac{1}{m}\sum_{i=1}^{m} g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\{\boldsymbol{z}\notin\mathcal{S}_i,\boldsymbol{z}_0\in\mathcal{S}_i\}}\right|
\end{aligned}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\left|g_i\left(\boldsymbol{z}\right)-g_i\left(\boldsymbol{z}_0\right)\right|+\frac{1}{m}\max_{1\leq i\leq m}\left|g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i\}}\right|\cdot\sum_{i=1}^{m}\mathbb{1}_{\{\boldsymbol{z}\in\mathcal{S}_i,\boldsymbol{z}_0\notin\mathcal{S}_i\}}$$
$$+\frac{1}{m}\max_{1\leq i\leq m}\left|g_i\left(\boldsymbol{z}_0\right)\mathbb{1}_{\{\boldsymbol{z}_0\in\mathcal{S}_i\}}\right|\cdot\sum_{i=1}^{m}\mathbb{1}_{\{\boldsymbol{z}\notin\mathcal{S}_i,\boldsymbol{z}_0\in\mathcal{S}_i\}}. \tag{156}$$

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

$$\left|\frac{1}{m}\sum_{i=1}^{m}g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,\left|a_{i,1}\right|\leq 5\sqrt{\log m}\right\}}-\mathbb{E}\left[g_i\left(\boldsymbol{z}\right)\mathbb{1}_{\left\{\left|\boldsymbol{a}_{i,\perp}^{\top}\boldsymbol{z}\right|\leq\beta\|\boldsymbol{z}\|_2,\left|a_{i,1}\right|\leq 5\sqrt{\log m}\right\}}\right]\right|\leq 2\epsilon\,\mathbb{E}\left[\left|g_i\left(\boldsymbol{z}\right)\right|\right] \tag{158}$$

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

$$\max_{1\leq i\leq m}\left|\boldsymbol{a}_i^\top \boldsymbol{x}^\natural\right| \leq 5\sqrt{\log m}\left\|\boldsymbol{x}^\natural\right\|_2$$

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

$$\begin{aligned}\left\|\frac{1}{m}\sum_{i=1}^m\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^2\boldsymbol{a}_i\boldsymbol{a}_i^\top - \boldsymbol{I}_n - 2\boldsymbol{z}\boldsymbol{z}^\top\right\| &\leq \underbrace{\left\|\frac{1}{m}\sum_{i=1}^m\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^2\boldsymbol{a}_i\boldsymbol{a}_i^\top\mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top\boldsymbol{z}\right|\leq c_2\sqrt{\log m}\right\}} - \left(\beta_1\boldsymbol{I}_n+\beta_2\boldsymbol{z}\boldsymbol{z}^\top\right)\right\|}_{:=\theta_1} \\ &\quad + \underbrace{\left\|\beta_1\boldsymbol{I}_n+\beta_2\boldsymbol{z}\boldsymbol{z}^\top - \left(\boldsymbol{I}_n+2\boldsymbol{z}\boldsymbol{z}^\top\right)\right\|}_{:=\theta_2},\end{aligned}$$

where

$$\beta_1 := \mathbb{E}\left[\xi^2\,\mathbb{1}_{\left\{|\xi|\leq c_2\sqrt{\log m}\right\}}\right] \qquad \text{and} \qquad \beta_2 := \mathbb{E}\left[\xi^4\,\mathbb{1}_{\left\{|\xi|\leq c_2\sqrt{\log m}\right\}}\right] - \beta_1$$

with $\xi \sim N(0,1)$.

- For the second term θ_2, we can further bound it as follows

$$\begin{aligned}\theta_2 &\leq \left\|\beta_1\boldsymbol{I}_n - \boldsymbol{I}_n\right\| + \left\|\beta_2\boldsymbol{z}\boldsymbol{z}^\top - 2\boldsymbol{z}\boldsymbol{z}^\top\right\| \\ &\leq |\beta_1 - 1| + |\beta_2 - 2|,\end{aligned}$$

  which motivates us to bound $|\beta_1-1|$ and $|\beta_2-2|$. Towards this end, simple calculation yields

$$\begin{aligned}1-\beta_1 &= \sqrt{\frac{2}{\pi}}\cdot c_2\sqrt{\log m}e^{-\frac{c_2^2\log m}{2}} + \operatorname{erfc}\left(\frac{c_2\sqrt{\log m}}{2}\right) \\ &\overset{\text{(i)}}{\leq} \sqrt{\frac{2}{\pi}}\cdot c_2\sqrt{\log m}e^{-\frac{c_2^2\log m}{2}} + \frac{1}{\sqrt{\pi}}\frac{2}{c_2\sqrt{\log m}}e^{-\frac{c_2^2\log m}{4}} \\ &\overset{\text{(ii)}}{\leq} \frac{1}{m},\end{aligned}$$

  where (i) arises from the fact that for all $x>0$, $\operatorname{erfc}(x) \leq \frac{1}{\sqrt{\pi}}\frac{1}{x}e^{-x^2}$ and (ii) holds as long as $c_2>0$ is sufficiently large. Similarly, for the difference $|\beta_2-2|$, one can easily show that

$$|\beta_2 - 2| \leq \left|\mathbb{E}\left[\xi^4\,\mathbb{1}_{\left\{|\xi|\leq c_2\sqrt{\log m}\right\}}\right] - 3\right| + |\beta_1 - 1| \leq \frac{2}{m}. \tag{160}$$

  Take the previous two bounds collectively to reach

$$\theta_2 \leq \frac{3}{m}.$$

- 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

$$\frac{1}{m}\sum_{i=1}^m\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^2\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2\mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top\boldsymbol{z}\right|\leq c_2\sqrt{\log m}\right\}} - \beta_1 - \beta_2\left(\boldsymbol{z}^\top\boldsymbol{x}\right)^2$$

  is a sum of m i.i.d. random variables with bounded sub-exponential norms. To see this, one has

$$\left\|\left(\boldsymbol{a}_i^\top\boldsymbol{z}\right)^2\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2\mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top\boldsymbol{z}\right|\leq c_2\sqrt{\log m}\right\}}\right\|_{\psi_1} \leq c_2^2\log m\left\|\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)^2\right\|_{\psi_1} \leq c_2^2\log m,$$

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

$$\begin{aligned}
&\left| \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{z}\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}\right| \le c_2\sqrt{\log m}\right\}} - \beta_1 - \beta_2 \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^2 \right| \\
&\quad \le \underbrace{\left| \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}_0\right| \le c_2\sqrt{\log m}\right\}} - \beta_1 - \beta_2 \left(\boldsymbol{z}_0^\top \boldsymbol{x}_0\right)^2 \right|}_{:=I_1} + \underbrace{\beta_2 \left| \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{z}_0^\top \boldsymbol{x}_0\right)^2 \right|}_{:=I_2} \\
&\quad\quad + \underbrace{\left| \frac{1}{m}\sum_{i=1}^m \left[\left(\boldsymbol{a}_i^\top \boldsymbol{z}\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}\right| \le c_2\sqrt{\log m}\right\}} - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0\right)^2 \mathbb{1}_{\left\{\left|\boldsymbol{a}_i^\top \boldsymbol{z}_0\right| \le c_2\sqrt{\log m}\right\}} \right] \right|}_{:=I_3}.
\end{aligned}$$

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

$$\begin{aligned}
\left| \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{z}_0^\top \boldsymbol{x}_0\right)^2 \right| &= \left| \boldsymbol{z}^\top \boldsymbol{x} - \boldsymbol{z}_0^\top \boldsymbol{x}_0 \right| \left| \boldsymbol{z}^\top \boldsymbol{x} + \boldsymbol{z}_0^\top \boldsymbol{x}_0 \right| \\
&= \left| \left(\boldsymbol{z} - \boldsymbol{z}_0\right)^\top \boldsymbol{x} + \boldsymbol{z}_0 \left(\boldsymbol{x} - \boldsymbol{x}_0\right) \right| \left| \boldsymbol{z}^\top \boldsymbol{x} + \boldsymbol{z}_0^\top \boldsymbol{x}_0 \right| \\
&\le 2\left(\|\boldsymbol{z} - \boldsymbol{z}_0\|_2 + \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \right) \le 2\theta,
\end{aligned}$$

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

$$I_3 = \left| \frac{1}{m}\sum_{i=1}^m \left[\left(\boldsymbol{a}_i^\top \boldsymbol{z}\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}\right)^2 \mathbb{1}_{\{\boldsymbol{z} \in \mathcal{S}_i\}} - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0\right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0\right)^2 \mathbb{1}_{\{\boldsymbol{z}_0 \in \mathcal{S}_i\}} \right] \right|$$

$$\leq \left| \frac{1}{m} \sum_{i=1}^{m} \left[\left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0 \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0 \right)^2 \right] \mathbb{1}_{\{\boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \in \mathcal{S}_i\}} \right|$$

$$+ \left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\{\boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \notin \mathcal{S}_i\}} \right| + \left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0 \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0 \right)^2 \mathbb{1}_{\{\boldsymbol{z}_0 \in \mathcal{S}_i, \boldsymbol{z} \notin \mathcal{S}_i\}} \right|. \qquad (162)$$

Here the decomposition is similar to what we have done in (156). For the first term in (162), one has

$$\left| \frac{1}{m} \sum_{i=1}^{m} \left[\left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0 \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0 \right)^2 \right] \mathbb{1}_{\{\boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \in \mathcal{S}_i\}} \right| \leq \frac{1}{m} \sum_{i=1}^{m} \left| \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 - \left(\boldsymbol{a}_i^\top \boldsymbol{z}_0 \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x}_0 \right)^2 \right|$$

$$\leq n^{\alpha} \theta,$$

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

$$\left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\{\boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \notin \mathcal{S}_i\}} \right| \leq \max_{1 \leq i \leq m} \left(\boldsymbol{a}_i^\top \boldsymbol{z} \right)^2 \mathbb{1}_{\{\boldsymbol{z} \in \mathcal{S}_i\}} \left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\{\boldsymbol{z} \in \mathcal{S}_i, \boldsymbol{z}_0 \notin \mathcal{S}_i\}} \right|$$

$$\leq c_2^2 \log m \left| \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\left\{ c_2 \sqrt{\log m} \leq \left| \boldsymbol{a}_i^\top \boldsymbol{z}_0 \right| \leq c_2 \sqrt{\log m} + \sqrt{6n}\theta \right\}} \right|.$$

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 $1 - O\left(m^{-10}\right)$,

$$\left\|\nabla^2 f\left(\boldsymbol{z}\right) - 6\boldsymbol{z}\boldsymbol{z}^\top - 3\left\|\boldsymbol{z}\right\|_2^2 \boldsymbol{I}_n + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top} + \left\|\boldsymbol{x}^\natural\right\|_2^2 \boldsymbol{I}_n\right\| \lesssim \sqrt{\frac{n\log^3 m}{m}} \max\left\{\left\|\boldsymbol{z}\right\|_2^2, \left\|\boldsymbol{x}^\natural\right\|_2^2\right\} \tag{163}$$

holds simultaneously for all $\boldsymbol{z}$ obeying $\max_{1\le i\le m}\left|\boldsymbol{a}_i^\top\boldsymbol{z}\right| \le c_0\sqrt{\log m}\left\|\boldsymbol{z}\right\|_2$, with the proviso that $m \gg n\log^3 m$. This together with the fact $\left\|\boldsymbol{x}^\natural\right\|_2 = 1$ leads to

$$\begin{aligned}
-\nabla^2 f\left(\boldsymbol{z}\right) &\succeq -6\boldsymbol{z}\boldsymbol{z}^\top - \left\{3\left\|\boldsymbol{z}\right\|_2^2 - 1 + O\left(\sqrt{\frac{n\log^3 m}{m}} \max\left\{\left\|\boldsymbol{z}\right\|_2^2, 1\right\}\right)\right\} \boldsymbol{I}_n \\
&\succeq -\left\{9\left\|\boldsymbol{z}\right\|_2^2 - 1 + O\left(\sqrt{\frac{n\log^3 m}{m}} \max\left\{\left\|\boldsymbol{z}\right\|_2^2, 1\right\}\right)\right\} \boldsymbol{I}_n.
\end{aligned}$$

As a consequence, if we pick $0 < \eta < \frac{c_2}{\max\left\{\left\|\boldsymbol{z}\right\|_2^2, 1\right\}}$ for $c_2 > 0$ sufficiently small, then $\boldsymbol{I}_n - \eta\nabla^2 f\left(\boldsymbol{z}\right) \succeq \boldsymbol{0}$. This combined with (163) gives

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

Additionally, it follows from (163) that

$$\begin{aligned}
\left\|\nabla^2 f\left(\boldsymbol{z}\right)\right\| &\le \left\|6\boldsymbol{z}\boldsymbol{z}^\top + 3\left\|\boldsymbol{z}\right\|_2^2 \boldsymbol{I}_n + 2\boldsymbol{x}^\natural\boldsymbol{x}^{\natural\top} + \left\|\boldsymbol{x}^\natural\right\|_2^2 \boldsymbol{I}_n\right\| + O\left(\sqrt{\frac{n\log^3 m}{m}}\right) \max\left\{\left\|\boldsymbol{z}\right\|_2^2, \left\|\boldsymbol{x}^\natural\right\|_2^2\right\} \\
&\le 9\|\boldsymbol{z}\|_2^2 + 3 + O\left(\sqrt{\frac{n\log^3 m}{m}}\right) \max\left\{\left\|\boldsymbol{z}\right\|_2^2, 1\right\} \\
&\le 10\|\boldsymbol{z}\|_2^2 + 4
\end{aligned}$$

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

$$\begin{aligned}
\max_{1\le l\le m}\left\|\boldsymbol{x}^{t,(l)}\right\|_2 &\le \left\|\boldsymbol{x}^t\right\|_2 + \max_{1\le l\le m}\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \\
&\overset{\text{(i)}}{\le} C_5 + \beta_t\left(1 + \frac{1}{\log m}\right)^t C_1\eta\frac{\sqrt{n\log^5 m}}{m} \\
&\overset{\text{(ii)}}{\le} C_5 + O\left(\frac{\sqrt{n\log^5 m}}{m}\right) \\
&\overset{\text{(iii)}}{\le} 2C_5,
\end{aligned}$$

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

$$\left\|\boldsymbol{x}^0 - \boldsymbol{x}^{0,(l)}\right\|_2 = \left|\lambda - \lambda^{(l)}\right| \lesssim \frac{\log m}{m} \leq \beta_0 C_1 \frac{\sqrt{n \log^5 m}}{m}.$$

This finishes the proof of (40a).

(b) The base case for (40b) can be easily deduced due to

$$\left|x_{\|}^0 - x_{\|}^{0,(l)}\right| \leq \left\|\boldsymbol{x}^0 - \boldsymbol{x}^{0,(l)}\right\|_2 \lesssim \frac{\log m}{m} \leq \alpha_0 C_2 \frac{\sqrt{n \log^{12} m}}{m}.$$

(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.