## A. Proof of Theorem 1

In this section, before presenting the proof of Theorem 1, we start with defining some useful notations. Recall that in (3), the empirical risk function for linear regression problem is defined as

$$\min_{\boldsymbol{W}}: \quad \hat{f}_{\Omega}(\boldsymbol{W}) = \frac{1}{2|\Omega|}\sum_{n\in\Omega}\Big|y_n - g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\Big|^2. \tag{22}$$

Population risk function, which is the expectation of the empirical risk function, is defined as

$$\min_{\boldsymbol{W}}: \quad f_{\Omega}(\boldsymbol{W}) = \mathbb{E}_{\boldsymbol{X}}\frac{1}{2|\Omega|}\sum_{n\in\Omega}\Big|y_n - g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\Big|^2. \tag{23}$$

Then, the road-map of the proof can be summarized in the following three steps.

First, we show the Hessian matrix of the population risk function $f_{\Omega_t}$ is positive-definite at ground-truth parameters $\boldsymbol{W}^*$ and then characterize the local convexity region of $f_{\Omega_t}$ near $\boldsymbol{W}^*$, which is summarized in Lemma 2.

Second, $\hat{f}_{\Omega_t}$ is non-smooth because of ReLU activation, but $f_{\Omega_t}$ is smooth. Hence, we characterize the gradient descent term as $\nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) = \langle\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}), \boldsymbol{W}^{(t)} - \boldsymbol{W}^*\rangle + \big(\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) - f_{\Omega_t}(\boldsymbol{W}^{(t)})\big)$. During this step, we need to apply concentration theorem to bound $\nabla\hat{f}_{\Omega_t}$ to its expectation $\nabla f_{\Omega_t}$ , which is summarized in Lemma 3.

Third, we take the momentum term of $\beta(\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)})$ into consideration and obtain the following recursive rule:

$$\begin{bmatrix}\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^*\end{bmatrix} = \boldsymbol{L}(\beta)\begin{bmatrix}\boldsymbol{W}^{(t)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^*\end{bmatrix}. \tag{24}$$

Then, we know iterates $\boldsymbol{W}^{(t)}$ converge to the ground-truth with a linear rate which is the largest singlar value of matrix $\boldsymbol{L}(\beta)$. Recall that AGD reduces to GD with $\beta = 0$, so our analysis applies to GD method as well. We are able to show the convergence rate of AGD is faster than GD by proving the largest singluar value of $\boldsymbol{L}(\beta)$ is smaller than $\boldsymbol{L}(0)$ for some $\beta > 0$. Lemma 4 provides the estimation error of $\boldsymbol{W}^{(0)}$ and sample complexity to guarantee $\|\boldsymbol{L}(\beta)\|_2$ is less than 1 for $t = 0$.

**Lemma 2.** *Let $f_{\Omega_t}$ be the population risk function in (23) for regression problems, then for any $\boldsymbol{W}$ that satisfies*

$$\|\boldsymbol{W}^* - \boldsymbol{W}\|_2 \le \frac{\varepsilon_0\sigma_K}{44\kappa^2\gamma K^2}, \tag{25}$$

*the second-order derivative of $f_{\Omega_t}$ is bounded as*

$$\frac{(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}\boldsymbol{I} \preceq \nabla^2 f_{\Omega_t}(\boldsymbol{W}) \preceq \frac{4\sigma_1^2(\boldsymbol{A})}{K}\boldsymbol{I}. \tag{26}$$

**Lemma 3.** *Let $\hat{f}_{\Omega_t}$ and $f_{\Omega_t}$ be the empirical and population risk functions in (22) and (23) for regression problems, respectively. Then, for any fixed point $\boldsymbol{W}$ satisfies (25), we have* [6]

$$\Big\|\nabla f_{\Omega_t}(\boldsymbol{W}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W})\Big\|_2 \lesssim \sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\|\boldsymbol{W} - \boldsymbol{W}^*\|_2, \tag{27}$$

*with probability at least $1 - K^2\cdot N^{-10}$.*

**Lemma 4.** *Assume the number of samples $|\Omega_t| \gtrsim \kappa^3(1+\delta^2)\sigma_1^4(\boldsymbol{A})Kd\log^4 N$, the tensor initialization method via Subroutine 1 outputs $\boldsymbol{W}^{(0)}$ such that*

$$\|\boldsymbol{W}^{(0)} - \boldsymbol{W}^*\|_2 \lesssim \kappa^6\sigma_1^2(\boldsymbol{A})\sqrt{\frac{K^4(1+\delta^2)d\log N}{|\Omega_t|}}\|\boldsymbol{W}^*\|_2 \tag{28}$$

*with probability at least $1 - N^{-10}$.*

[6] We use $f(d) \gtrsim (\text{ or } \lesssim, \eqsim) g(d)$ to denote there exists some positive constant $C$ such that $f(d) \ge (\text{ or } \le, =) C\cdot g(d)$ when $d$ is sufficiently large.

The proofs of Lemmas 2 and 3 are included in Appendix A.1 and A.2, respectively, while the proof of Lemma 4 can be found in Appendix D. With these three preliminary lemmas on hand, the proof of Theorem 1 is formally summarized in the following contents.

*Proof of Theorem 1.* The update rule of $\boldsymbol{W}^{(t)}$ is

$$\begin{aligned}\boldsymbol{W}^{(t+1)} =&\boldsymbol{W}^{(t)} - \eta\nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) + \beta(\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)})\\ =&\boldsymbol{W}^{(t)} - \eta\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) + \beta(\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)}) + \eta(\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})).\end{aligned} \tag{29}$$

Since $\nabla^2_{\Omega_t}$ is a smooth function, by the intermediate value theorem, we have

$$\begin{aligned}\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} &- \eta\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)})(\boldsymbol{W}^{(t)} - \boldsymbol{W}^*)\\ &+ \beta(\boldsymbol{W}^{(t)} - \boldsymbol{W}^{(t-1)})\\ &+ \eta\big(\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})\big),\end{aligned} \tag{30}$$

where $\widehat{\boldsymbol{W}}^{(t)}$ lies in the convex hull of $\boldsymbol{W}^{(t)}$ and $\boldsymbol{W}^*$.

Next, we have

$$\begin{bmatrix}\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^*\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} - \eta\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}) + \beta\boldsymbol{I} & \beta\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^*\end{bmatrix} + \eta\begin{bmatrix}\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})\\ 0\end{bmatrix}. \tag{31}$$

Let $\boldsymbol{L}(\beta) = \begin{bmatrix}\boldsymbol{I} - \eta\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}) + \beta\boldsymbol{I} & \beta\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}$, so we have

$$\left\|\begin{bmatrix}\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^*\end{bmatrix}\right\|_2 = \|\boldsymbol{L}(\beta)\|_2\left\|\begin{bmatrix}\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^*\end{bmatrix}\right\|_2 + \eta\left\|\begin{bmatrix}\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})\\ 0\end{bmatrix}\right\|_2.$$

From Lemma 3, we know that

$$\eta\left\|\nabla f_{\Omega_t}(\boldsymbol{W}^{(t)}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)})\right\|_2 \lesssim \eta\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\|\boldsymbol{W} - \boldsymbol{W}^*\|_2. \tag{32}$$

Then, we have

$$\begin{aligned}\|\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\|_2 \lesssim&\left(\|\boldsymbol{L}(\beta)\|_2 + \eta\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\right)\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2\\ :\eqsim&\nu(\beta)\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2.\end{aligned} \tag{33}$$

Let $\nabla^2 f(\widehat{\boldsymbol{W}}^{(t)}) = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^T$ be the eigen-decomposition of $\nabla^2 f(\widehat{\boldsymbol{W}}^{(t)})$. Then, we define

$$\widetilde{\boldsymbol{L}}(\beta) := \begin{bmatrix}\boldsymbol{S}^T & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}^T\end{bmatrix}\boldsymbol{L}(\beta)\begin{bmatrix}\boldsymbol{S} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} - \eta\boldsymbol{\Lambda} + \beta\boldsymbol{I} & \beta\boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}. \tag{34}$$

Since $\begin{bmatrix}\boldsymbol{S} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}\end{bmatrix}\begin{bmatrix}\boldsymbol{S}^T & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}^T\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}$, we know $\boldsymbol{L}(\beta)$ and $\widetilde{\boldsymbol{L}}(\beta)$ share the same eigenvalues. Let $\lambda_i$ be the $i$-th eigenvalue of $\nabla^2 f_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)})$, then the corresponding $i$-th eigenvalue of $\boldsymbol{L}(\beta)$, denoted by $\delta_i(\beta)$, satisfies

$$\delta_i^2 - (1 - \eta\lambda_i + \beta)\delta_i + \beta = 0. \tag{35}$$

Then, we have

$$\delta_i(\beta) = \frac{(1 - \eta\lambda_i + \beta) + \sqrt{(1 - \eta\lambda_i + \beta)^2 - 4\beta}}{2}, \tag{36}$$

and

$$|\delta_i(\beta)| = \begin{cases} \sqrt{\beta}, & \text{if} \quad \beta \geq \left(1-\sqrt{\eta\lambda_i}\right)^2, \\ \frac{1}{2}\left|(1-\eta\lambda_i+\beta)+\sqrt{(1-\eta\lambda_i+\beta)^2-4\beta}\right|, & \text{otherwise.} \end{cases} \tag{37}$$

Note that the other root of (35) is abandoned because the root in (36) is always no less than the other root with $|1-\eta\lambda_i|<1$. By simple calculations, we have

$$\delta_i(0) > \delta_i(\beta), \quad \text{for} \quad \forall \beta \in \left(0,(1-\eta\lambda_i)^2\right). \tag{38}$$

Moreover, $\delta_i$ achieves the minimum $\delta_i^* = |1-\sqrt{\eta\lambda_i}|$ when $\beta = \left(1-\sqrt{\eta\lambda_i}\right)^2$.

Let us first assume $\boldsymbol{W}^{(t)}$ satisfies (25), then from Lemma 2, we know that

$$0 < \frac{(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \leq \lambda_i \leq \frac{4\sigma_1^2(\boldsymbol{A})}{K}.$$

Let $\gamma_1 = \frac{(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}$ and $\gamma_2 = \frac{4\sigma_1^2(\boldsymbol{A})}{K}$. If we choose $\beta$ such that

$$\beta^* = \max\left\{(1-\sqrt{\eta\gamma_1})^2, (1-\sqrt{\eta\gamma_2})^2\right\}, \tag{39}$$

then we have $\beta \geq (1-\sqrt{\eta\lambda_i})^2$ and $\delta_i = \max\left\{|1-\sqrt{\eta\gamma_1}|, |1-\sqrt{\eta\gamma_2}|\right\}$ for any $i$.

Let $\eta = \frac{1}{2\gamma_2}$, then $\beta^*$ equals to $\left(1-\sqrt{\frac{\gamma_1}{2\gamma_2}}\right)^2$. Then, for any $\varepsilon_0 \in (0,1/2)$, we have

$$\|\boldsymbol{L}(\beta^*)\|_2 = \max_i \delta_i(\beta^*) = 1-\sqrt{\frac{\gamma_1}{2\gamma_2}} = 1-\sqrt{\frac{1-\varepsilon_0}{88\kappa^2\gamma K}} \leq 1-\frac{1-(3/4)\cdot\varepsilon_0}{\sqrt{88\kappa^2\gamma K}}. \tag{40}$$

Then, let

$$\eta\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}} \lesssim \frac{\varepsilon_0}{4\sqrt{88\kappa^2\gamma K}}, \tag{41}$$

we need $|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^2\gamma M(1+\delta^2)\sigma_1^2(\boldsymbol{A})K^3 d\log N$. Combining (40) and (41), we have

$$\nu(\beta^*) \leq 1-\frac{1-\varepsilon_0}{\sqrt{88\kappa^2\gamma K}}. \tag{42}$$

Let $\beta = 0$, we have

$$\nu(0) \geq \|\boldsymbol{A}(0)\|_2 = 1-\frac{1-\varepsilon_0}{88\kappa^2\gamma K},$$

$$\nu(0) \lesssim \|\boldsymbol{A}(0)\|_2 + \eta\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}} \leq 1-\frac{1-2\varepsilon_0}{88\kappa^2\gamma K}$$

if $|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^2\gamma M(1+\delta^2)\sigma_1^2(\boldsymbol{A})K^3 d\log N$.

Hence, with $\eta = \frac{1}{2\gamma_2}$ and $\beta = \left(1-\frac{\gamma_1}{2\gamma_2}\right)^2$, we have

$$\|\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\|_2 \leq \left(1-\frac{1-\varepsilon_0}{\sqrt{88\kappa^2\gamma K}}\right)\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2, \tag{43}$$

provided $\boldsymbol{W}^{(t)}$ satisfies (25), and

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^2\gamma(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^3 d\log N. \tag{44}$$

Then, we can start mathematical induction of (43) over $t$.

**Base case**: According to Lemma 4, we know that (25) holds for $\boldsymbol{W}^{(0)}$ if

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^9\gamma^2(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^8 d\log N. \tag{45}$$

According to Theorem 1, it is clear that the number of samples $|\Omega_t|$ satisfies (45), then (25) indeed holds for $t = 0$. Since (25) holds for $t = 0$ and $|\Omega_t|$ in Theorem 1 satisfies (44) as well, we have (43) holds for $t = 0$.

**Induction step**: Assuming (43) holds for $\boldsymbol{W}^{(t)}$, we need to show that (43) holds for $\boldsymbol{W}^{(t+1)}$. That is to say, we need $|\Omega_t|$ satisfies (44), which holds naturally from Theorem 1.

Therefore, when $|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^9\gamma^2(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^8 d\log N$, we know that (43) holds for all $0 \le t \le T-1$ with probability at least $1 - K^2T \cdot N^{-10}$. By simple calculations, we can obtain

$$\|\boldsymbol{W}^{(T)} - \boldsymbol{W}^*\|_2 \le \Big(1 - \frac{1-\varepsilon_0}{\sqrt{88\kappa^2\gamma K}}\Big)^T \|\boldsymbol{W}^{(0)} - \boldsymbol{W}^*\|_2 \tag{46}$$

□

### A.1. Proof of Lemma 2

In this section, we provide the proof of Lemma 2 which shows the local convexity of $f_{\Omega_t}$ in a small neighborhood of $\boldsymbol{W}^*$. The roadmap is to first bound the smallest eigenvalue of $\nabla^2 f_{\Omega_t}$ in the ground truth as shown in Lemma 5, then show that the difference of $\nabla^2 f_{\Omega_t}$ between any fixed point $\boldsymbol{W}$ in this region and the ground truth $\boldsymbol{W}^*$ is bounded in terms of $\|\boldsymbol{W} - \boldsymbol{W}^*\|_2$ by Lemma 6.

**Lemma 5.** *The second-order derivative of $f_{\Omega_t}$ at the ground truth $\boldsymbol{W}^*$ satisfies*

$$\frac{\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \preceq \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) \preceq \frac{3\sigma_1^2(\boldsymbol{A})}{K}. \tag{47}$$

**Lemma 6.** *Suppose $\boldsymbol{W}$ satisfies (25), we have*

$$\big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\big\|_2 \le 4\sigma_1^2(\boldsymbol{A})\frac{\|\boldsymbol{W}^* - \boldsymbol{W}\|_2}{\sigma_K}. \tag{48}$$

The proofs of Lemmas 5 and 6 can be found in Sec. A.3. With these two preliminary lemmas on hand, the proof of Lemma 2 is formally summarized in the following contents.

*Proof of Lemma 2.* By the triangle inequality, we have

$$\Big|\big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\big\|_2 - \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\big\|_2\Big| \le \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2,$$

and

$$\begin{aligned}\big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\big\|_2 &\le \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\big\|_2 + \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2,\\ \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\big\|_2 &\ge \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\big\|_2 - \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2.\end{aligned}$$

The error bound of $\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2$ can be derived from Lemma 6, and the error bound of $\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)$ is provided in Lemma 5.

Therefore, for any $\boldsymbol{W}$ satisfies (25), we have

$$\frac{(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \le \big\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\big\|_2 \le \frac{4\sigma_1^2(\boldsymbol{A})}{K}. \tag{49}$$

□

### A.2. Proof of Lemma 3

The proof of Lemma 3 is mainly to bound the concentration error of random variables $\boldsymbol{z}_n(j,k)$ as shown in (60). We first show that $\boldsymbol{z}_n(j,k)$ is a sub-exponential random variable, and the definitions of sub-Gaussian and sub-exponential random variables are provided in Definitions 1 and 2. Though Hoeffding's inequality provides the concentration error for sum of independent random variables, random variables $\boldsymbol{z}_n(j,k)$ with different $j,k$ are not independent. Hence, we introduce Lemma 7 to provide the upper bound for the moment generation function of the sum of partly dependent random variables and then apply standard Chernoff inequality. Lemmas 8 and 9 are standard tools in analyzing spectral norms of high-demensional random matrices.

**Definition 1** (Definition 5.7, (Vershynin, 2010)). *A random variable $X$ is called a sub-Gaussian random variable if it satisfies*

$$(\mathbb{E}|X|^p)^{1/p} \leq c_1\sqrt{p} \tag{50}$$

*for all $p \geq 1$ and some constant $c_1 > 0$. In addition, we have*

$$\mathbb{E}e^{s(X-\mathbb{E}X)} \leq e^{c_2\|X\|_{\psi_2}^2 s^2} \tag{51}$$

*for all $s \in \mathbb{R}$ and some constant $c_2 > 0$, where $\|X\|_{\psi_2}$ is the sub-Gaussian norm of $X$ defined as $\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$.*

*Moreover, a random vector $\boldsymbol{X} \in \mathbb{R}^d$ belongs to the sub-Gaussian distribution if one-dimensional marginal $\boldsymbol{\alpha}^T\boldsymbol{X}$ is sub-Gaussian for any $\boldsymbol{\alpha} \in \mathbb{R}^d$, and the sub-Gaussian norm of $\boldsymbol{X}$ is defined as $\|\boldsymbol{X}\|_{\psi_2} = \sup_{\|\boldsymbol{\alpha}\|_2=1} \|\boldsymbol{\alpha}^T\boldsymbol{X}\|_{\psi_2}$.*

**Definition 2** (Definition 5.13, (Vershynin, 2010)). *A random variable $X$ is called a sub-exponential random variable if it satisfies*

$$(\mathbb{E}|X|^p)^{1/p} \leq c_3 p \tag{52}$$

*for all $p \geq 1$ and some constant $c_3 > 0$. In addition, we have*

$$\mathbb{E}e^{s(X-\mathbb{E}X)} \leq e^{c_4\|X\|_{\psi_1}^2 s^2} \tag{53}$$

*for $s \leq 1/\|X\|_{\psi_1}$ and some constant $c_4 > 0$, where $\|X\|_{\psi_1}$ is the sub-exponential norm of $X$ defined as $\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$.*

**Lemma 7.** *Given a sampling set $\mathcal{X} = \{x_n\}_{n=1}^N$ that contains $N$ partly dependent random variables, for each $n \in [N]$, suppose $x_n$ is dependent with at most $d_{\mathcal{X}}$ random variables in $\mathcal{X}$ (including $x_n$ itself), and the moment generate function of $x_n$ satisfies $\mathbb{E}_{x_n} e^{sx_n} \leq e^{Cs^2}$ for some constant $C$ that may depend on $x_n$. Then, the moment generation function of $\sum_{n=1}^N x_n$ is bounded as*

$$\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^N x_n} \leq e^{Cd_{\mathcal{X}}Ns^2}. \tag{54}$$

**Lemma 8** (Lemma 5.2, (Vershynin, 2010)). *Let $\mathcal{B}(0,1) \in \{\boldsymbol{\alpha}\big|\|\boldsymbol{\alpha}\|_2 = 1, \boldsymbol{\alpha} \in \mathbb{R}^d\}$ denote a unit ball in $\mathbb{R}^d$. Then, a subset $\mathcal{S}_\xi$ is called a $\xi$-net of $\mathcal{B}(0,1)$ if every point $\boldsymbol{z} \in \mathcal{B}(0,1)$ can be approximated to within $\xi$ by some point $\boldsymbol{\alpha} \in \mathcal{B}(0,1)$, i.e. $\|\boldsymbol{z} - \boldsymbol{\alpha}\|_2 \leq \xi$. Then the minimal cardinality of a $\xi$-net $\mathcal{S}_\xi$ satisfies*

$$|\mathcal{S}_\xi| \leq (1 + 2/\xi)^d. \tag{55}$$

**Lemma 9** (Lemma 5.3, (Vershynin, 2010)). *Let $\boldsymbol{A}$ be an $N \times d$ matrix, and let $\mathcal{S}_\xi$ be a $\xi$-net of $\mathcal{B}(0,1)$ in $\mathbb{R}^d$ for some $\xi \in (0,1)$. Then*

$$\|\boldsymbol{A}\|_2 \leq (1-\xi)^{-1} \max_{\boldsymbol{\alpha}\in\mathcal{S}_\xi} |\boldsymbol{\alpha}^T\boldsymbol{A}\boldsymbol{\alpha}|. \tag{56}$$

The proof of Lemma 7 can be found in Appendex A.3. With these preliminary Lemmas and definition on hand, the proof of Lemma 3 is formally summarized in the following contents.

*Proof of Lemma 3* . We have

$$\hat{f}_{\Omega_t}(\boldsymbol{W}) = \frac{1}{2|\Omega_t|}\sum_{n\in\Omega_t}\Big|y_n - g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\Big|^2 = \frac{1}{2|\Omega_t|}\sum_{n\in\Omega_t}\Big|y_n - \sum_{j=1}^K \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\Big|^2, \tag{57}$$

and

$$f_{\Omega_t}(\boldsymbol{W}) = \mathbb{E}_{\boldsymbol{X}}\hat{f}_{\Omega_t}(\boldsymbol{W}) = \frac{1}{2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{x}}\Big|y_n - \sum_{j=1}^K \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\Big|^2. \tag{58}$$

The gradients of $\hat{f}_{\Omega_t}$ are

$$\begin{aligned}\frac{\partial \hat{f}_{\Omega_t}}{\partial \boldsymbol{w}_k}(\boldsymbol{W}) =& \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(y_n - \sum_{j=1}^{K}\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\Big)\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\\ =& \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(\sum_{j=1}^{K}\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \sum_{j=1}^{K}\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\Big)\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\\ =& \sum_{j=1}^{K}\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big)\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k).\end{aligned} \tag{59}$$

Let us define

$$\boldsymbol{z}_n(k,j) = \boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big), \tag{60}$$

then for any normalized $\boldsymbol{\alpha}\in\mathbb{R}^d$, we have

$$\begin{aligned}&p^{-1}\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big)\big|^p\Big)^{1/p}\\ \le& p^{-1}\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^{2p}\cdot\mathbb{E}_{\boldsymbol{X}}\big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big)\big|^{2p}\Big)^{1/2p}\\ \le& p^{-1}\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^{2p}\Big)^{1/2p}\cdot\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{a}_n^T\boldsymbol{X}(\boldsymbol{w}_j^* - \boldsymbol{w}_j)\big|^{2p}\Big)^{1/2p}\end{aligned} \tag{61}$$

where the first inequality comes from the Cauchy-Schwarz inequality. Furthermore, $\boldsymbol{a}_n^T\boldsymbol{X}$ belongs to the Gaussian distribution and thus is a sub-Gaussian random vector as well. Then, from Definition 1, we have

$$\begin{aligned}&\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^{2p}\Big)^{1/2p} \le (2p)^{1/2}\|\boldsymbol{X}^T\boldsymbol{a}_n\|_{\psi_2} \le (2p)^{1/2}\|\boldsymbol{a}_n\|_2,\\ \text{and}\quad&\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{a}_n^T\boldsymbol{X}(\boldsymbol{w}_j^* - \boldsymbol{w}_j)\big|^{2p}\Big)^{1/2p} \le (2p)^{1/2}\|\boldsymbol{a}_n\|_2\cdot\|\boldsymbol{w}_j^* - \boldsymbol{w}_j\|_2.\end{aligned} \tag{62}$$

Then, we have

$$\begin{aligned}&p^{-1}\Big(\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_k)\big(\phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*) - \phi(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j)\big)\big|^p\Big)^{1/p}\\ \le& p^{-1}\cdot 2p\|\boldsymbol{a}_n\|_2^2\cdot\|\boldsymbol{w}_j^* - \boldsymbol{w}_j\|_2\\ \le& 2\sigma_1^2(\boldsymbol{A})\cdot\|\boldsymbol{w}_j^* - \boldsymbol{w}_j\|_2.\end{aligned} \tag{63}$$

Therefore, from Definition 2, $\boldsymbol{z}_n(k,j)$ belongs to the sub-exponential distribution with

$$\|\boldsymbol{z}_n\|_{\phi_1} \le 2\sigma_1^2(\boldsymbol{A})\cdot\|\boldsymbol{w}_j^* - \boldsymbol{w}_j\|_2. \tag{64}$$

Recall that each node is connected with at most $\delta$ other nodes. Hence, for any fixed $\boldsymbol{z}_n$, there are at most $(1+\delta^2)$ (including $\boldsymbol{z}_n$ itself) elements in $\{\boldsymbol{z}_l | l\in\Omega_t\}$ are dependant with $\boldsymbol{z}_n$. From Lemma 7, the moment generation function of $\sum_{n\in\Omega_t}(\boldsymbol{z}_n - \mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n)$ satisfies

$$\mathbb{E}_{\boldsymbol{X}}e^{s\sum_{n\in\Omega_t}(\boldsymbol{z}_n - \mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n)} \le e^{C(1+\delta^2)|\Omega_t|s^2}. \tag{65}$$

By Chernoff inequality, we have

$$\text{Prob}\Big\{\Big\|\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\big(\boldsymbol{z}_n(k,j) - \mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n(k,j)\big)\Big\|_2 > t\Big\} \le \frac{e^{C(1+\delta^2)|\Omega_t|s^2}}{e^{|\Omega_t|ts}} \tag{66}$$

for any $s>0$.

Let $s = t/\big(C(1+\delta^2)\|\boldsymbol{z}_n\|_{\phi_1}^2\big)$ and $t = \sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\|\boldsymbol{z}_n\|_{\phi_1}$, we have

$$\begin{aligned}\Big\|\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\big(\boldsymbol{z}_n(k,j)-\mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n(k,j)\big)\Big\|_2 \leq & C\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\sigma_1^2(\boldsymbol{A})\cdot\|\boldsymbol{w}_j^*-\boldsymbol{w}_j\|_2\\ \leq & C\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\cdot\|\boldsymbol{W}^*-\boldsymbol{W}\|_2\end{aligned} \tag{67}$$

with probability at least $1-N^{-d}$.

In conclusion, by selecting $\xi=\frac{1}{2}$ in Lemmas 8 and 9, we have

$$\begin{aligned}\Big\|\frac{\partial\hat{f}_{\Omega_t}}{\partial\boldsymbol{w}_k}(\boldsymbol{W})-\frac{\partial f_{\Omega_t}}{\partial\boldsymbol{w}_k}(\boldsymbol{W})\Big\|_2 \leq & \sum_{k=1}^K\sum_{j=1}^K\frac{1}{K^2}\Big\|\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\boldsymbol{z}_n(k,j)-\mathbb{E}_{\boldsymbol{X}}\boldsymbol{z}_n(k,j)\Big\|_2\\ \leq & C\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\cdot\|\boldsymbol{W}^*-\boldsymbol{W}\|_2\end{aligned} \tag{68}$$

with probability at least $1-\big(\frac{5}{N}\big)^d$. □

### A.3. Proof of auxiliary lemmas for regression problems

#### A.3.1. Proof of Lemma 5

*Proof of Lemma 5* . For any normalized $\boldsymbol{\alpha}\in\mathbb{R}^{Kd}$, the lower bound of $\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)$ is derived from

$$\begin{aligned}\boldsymbol{\alpha}^T\nabla^2 f(\boldsymbol{W}^*)\boldsymbol{\alpha} = & \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}\Big[\Big(\sum_{j=1}^K\boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*)\Big)^2\Big]\\ \geq & \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\frac{\|\boldsymbol{a}_n\|_2^2}{11\kappa^2\gamma}\|\boldsymbol{\alpha}\|_2^2=\frac{\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2},\end{aligned} \tag{69}$$

where the last inequality can be derived from Lemma D.6 in (Zhong et al., 2017c). In spite that the error bound in (Zhong et al., 2017c) is given in terms of $\boldsymbol{x}_n$ instead of $\boldsymbol{X}^T\boldsymbol{a}_n$, both $\boldsymbol{x}_n$ and $\boldsymbol{X}^T\boldsymbol{a}_n$ belong to Gaussian distribution. Hence, we can follow the similar steps in (Zhong et al., 2017c) to derive the results for Gaussian random variable $\boldsymbol{X}^T\boldsymbol{a}_n$ with 0 mean and $\|\boldsymbol{a}_n\|_2^2$ variance.

Next, the upper bound of $\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)$ is derived from

$$\begin{aligned}&\boldsymbol{\alpha}^T\nabla^2 f(\boldsymbol{W}^*)\boldsymbol{\alpha}\\ =&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}\Big[\Big(\sum_{j=1}^K\boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_j^*)\Big)^2\Big]\\ =&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\sum_{j_1=1}^K\sum_{j_2=1}^K\mathbb{E}_{\boldsymbol{X}}\Big[\boldsymbol{\alpha}_{j_1}^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\boldsymbol{\alpha}_{j_2}^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*)\Big]\\ \leq&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\sum_{j_1=1}^K\sum_{j_2=1}^K\Big[\mathbb{E}_{\boldsymbol{X}}|\boldsymbol{\alpha}_{j_1}^T\boldsymbol{X}^T\boldsymbol{a}_n|^4\cdot\mathbb{E}_{\boldsymbol{X}}|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)|^4\cdot\mathbb{E}_{\boldsymbol{X}}|\boldsymbol{\alpha}_{j_2}^T\boldsymbol{X}^T\boldsymbol{a}_n|^4\cdot\mathbb{E}_{\boldsymbol{X}}|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*)|^4\Big]^{\frac{1}{4}}\\ \leq&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\sum_{j_1=1}^K\sum_{j_2=1}^K 3\sigma_1^2(\boldsymbol{A})\|\boldsymbol{\alpha}_{j_1}\|_2\|\boldsymbol{\alpha}_{j_2}\|_2\\ \leq&3\sigma_1^2(\boldsymbol{A})\frac{\|\boldsymbol{\alpha}\|_2^2}{K},\end{aligned} \tag{70}$$

which completes the proof. □

### A.3.2. PROOF OF LEMMA 6

*Proof of Lemma 6 .* The second-order derivative of $f_{\Omega_t}$ is written as

$$\begin{aligned}
&\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}) - \frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}^*)\\
=&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\Big[\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1})\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*)\Big]\\
=&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\big(\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\big)\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\\
&-\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\big(\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\big).
\end{aligned} \tag{71}$$

For any normalized $\boldsymbol{\alpha}\in\mathbb{R}^d$, we have

$$\begin{aligned}
&\Big|\boldsymbol{\alpha}^T\Big[\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}) - \frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}^*)\Big]\boldsymbol{\alpha}\Big|\\
\le&\Big|\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\big(\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\big)\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\Big|\\
&+\Big|\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\big(\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\big)\Big|\\
\le&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^2\cdot\Big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\Big|\\
&+\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^2\cdot\Big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2}^*) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_2})\Big|.
\end{aligned} \tag{72}$$

It is easy to verify there exists a basis such that $\mathcal{B}=\{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha}_4^\perp,\cdots,\boldsymbol{\alpha}_d^\perp\}$ with $\{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}\}$ spanning a subspace that contains $\boldsymbol{\alpha}$, $\boldsymbol{w}_{j_1}$ and $\boldsymbol{w}_{j_1}^*$. Then, for any $\boldsymbol{X}^T\boldsymbol{a}_n\in\mathbb{R}^d$, we have a unique $\boldsymbol{z}=\begin{bmatrix}z_1 & z_2 & \cdots & z_d\end{bmatrix}^T$ such that

$$\boldsymbol{X}^T\boldsymbol{a}_n = z_1\boldsymbol{\alpha}+z_2\boldsymbol{\beta}+z_3\boldsymbol{\gamma}+\cdots+z_d\boldsymbol{\alpha}_d^\perp.$$

Also, since $\boldsymbol{X}^T\boldsymbol{a}_n\sim\mathcal{N}(\boldsymbol{0},\|\boldsymbol{a}_n\|_2^2\boldsymbol{I}_d)$, we have $\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{0},\|\boldsymbol{a}_n\|_2^2\boldsymbol{I}_d)$. Then, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^2\cdot\Big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\Big|\\
=&\mathbb{E}_{z_1,z_2,z_3}|\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big) - \phi'\big({\boldsymbol{w}_{j_1}^*}^T\widetilde{\boldsymbol{x}}\big)|\cdot|\boldsymbol{a}^T\widetilde{\boldsymbol{x}}|^2\\
=&\int|\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big) - \phi'\big({\boldsymbol{w}_{j_1}^*}^T\widetilde{\boldsymbol{x}}\big)|\cdot|\boldsymbol{a}^T\widetilde{\boldsymbol{x}}|^2\cdot f_Z(z_1,z_2,z_3)dz_1dz_2dz_3,
\end{aligned}$$

where $\widetilde{\boldsymbol{x}}=z_1\boldsymbol{\alpha}+z_2\boldsymbol{\beta}+z_3\boldsymbol{\gamma}$ and $f_Z(z_1,z_2,z_3)$ is the probability density function of $(z_1,z_2,z_3)$. Next, we consider spherical coordinates with $z_1=r\cos\phi_1, z_2=r\sin\phi_1\sin\phi_2, z_3=z_2=r\sin\phi_1\cos\phi_2$. Hence,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\big|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\big|^2\cdot\Big|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}) - \phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\Big|\\
=&\int\int\int|\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big) - \phi'\big({\boldsymbol{w}_{j_1}^*}^T\widetilde{\boldsymbol{x}}\big)|\cdot|r\cos\phi_1|^2\cdot f_Z(r,\phi_1,\phi_2)r^2\sin\phi_1 dr d\phi_1 d\phi_2.
\end{aligned} \tag{73}$$

It is easy to verify that $\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big)$ only depends on the direction of $\widetilde{\boldsymbol{x}}$ and

$$f_Z(r,\phi_1,\phi_2)=\frac{1}{(2\pi\|\boldsymbol{a}_n\|_2^2)^{\frac{3}{2}}}e^{-\frac{x_1^2+x_2^2+x_3^2}{2\|\boldsymbol{a}_n\|_2^2}}=\frac{1}{(2\pi\|\boldsymbol{a}_n\|_2^2)^{\frac{3}{2}}}e^{-\frac{r^2}{2\|\boldsymbol{a}_n\|_2^2}}$$

only depends on $r$. Then, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\right|^2\cdot\left|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1})-\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\right|\\
=&\int\int\int|\phi'\big(\boldsymbol{w}_{j_1}^T(\widetilde{\boldsymbol{x}}/r)\big)-\phi'\big(\boldsymbol{w}_{j_1}^{*\,T}(\widetilde{\boldsymbol{x}}/r)\big)|\cdot|r\cos\phi_1|^2\cdot f_Z(r)r^2\sin\phi_1 dr d\phi_1 d\phi_2\\
=&\int_0^\infty r^4 f_z(r)dr\int_0^\pi\int_0^{2\pi}|\cos\phi_1|^2\cdot\sin\phi_1\cdot|\phi'\big(\boldsymbol{w}_{j_2}^T(\widetilde{\boldsymbol{x}}/r)\big)-\phi'\big(\boldsymbol{w}_{j_2}^{*\,T}(\widetilde{\boldsymbol{x}}/r)\big)|d\phi_1 d\phi_2\\
\leq&3\|\boldsymbol{a}_n\|_2^2\cdot\int_0^\infty r^2 f_z(r)dr\int_0^\pi\int_0^{2\pi}\sin\phi_1\cdot|\phi'\big(\boldsymbol{w}_{j_2}^T(\widetilde{\boldsymbol{x}}/r)\big)-\phi'\big(\boldsymbol{w}_{j_2}^{*\,T}(\widetilde{\boldsymbol{x}}/r)\big)|d\phi_1 d\phi_2\\
=&3\|\boldsymbol{a}_n\|_2^2\cdot\mathbb{E}_{z_1,z_2,z_3}\left|\phi'\big(\boldsymbol{w}_{j_1}^T\widetilde{\boldsymbol{x}}\big)-\phi'\big(\boldsymbol{w}_{j_1}^{*\,T}\widetilde{\boldsymbol{x}}\big)\right|\\
=&3\|\boldsymbol{a}_n\|_2^2\cdot\mathbb{E}_{\boldsymbol{X}}\left|\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}\big)-\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*\big)\right|
\end{aligned}\tag{74}$$

Define a set $\mathcal{A}_1=\{\boldsymbol{x}|(\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x})(\boldsymbol{w}_{j_1}{}^T\boldsymbol{x})<0\}$. If $\boldsymbol{x}\in\mathcal{A}_1$, then $\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}$ and $\boldsymbol{w}_{j_1}{}^T\boldsymbol{x}$ have different signs, which means the value of $\phi'(\boldsymbol{w}_{j_1}^T\boldsymbol{x})$ and $\phi'(\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x})$ are different. This is equivalent to say that

$$|\phi'(\boldsymbol{w}_{j_1}^T\boldsymbol{x})-\phi'(\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x})|=\begin{cases}1,\ \text{if}\ \boldsymbol{x}\in\mathcal{A}_1\\0,\ \text{if}\ \boldsymbol{x}\in\mathcal{A}_1^c\end{cases}.\tag{75}$$

Moreover, if $\boldsymbol{x}\in\mathcal{A}_1$, then we have

$$|\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}|\leq|\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}-\boldsymbol{w}_{j_1}{}^T\boldsymbol{x}|\leq\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|\cdot\|\boldsymbol{x}\|.\tag{76}$$

Define a set $\mathcal{A}_2$ such that

$$\mathcal{A}_2=\Big\{\boldsymbol{x}\Big|\frac{|\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}|}{\|\boldsymbol{w}_{j_1}^*\|\|\boldsymbol{x}\|}\leq\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}\Big\}=\Big\{\theta_{\boldsymbol{x},\boldsymbol{w}_{j_1}^*}\Big||\cos\theta_{\boldsymbol{x},\boldsymbol{w}_{j_1}^*}|\leq\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}\Big\}.\tag{77}$$

Hence, we have that

$$\mathbb{E}_{\boldsymbol{x}}|\phi'(\boldsymbol{w}_{j_1}^T\boldsymbol{x})-\phi'(\boldsymbol{w}_{j_1}^{*\,T}\boldsymbol{x}_{i_2})|=\text{Prob}(\boldsymbol{x}\in\mathcal{A}_1)\leq\text{Prob}(\boldsymbol{x}\in\mathcal{A}_2).\tag{78}$$

Since $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$, $\theta_{\boldsymbol{x},\boldsymbol{w}_{j_1}^*}$ belongs to the uniform distribution on $[-\pi,\pi]$, we have

$$\begin{aligned}
\text{Prob}(\boldsymbol{x}\in\mathcal{A}_2)=&\frac{\pi-\arccos\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}}{\pi}\\
\leq&\frac{1}{\pi}\tan(\pi-\arccos\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|})\\
=&\frac{1}{\pi}\cot(\arccos\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|})\\
\leq&\frac{2}{\pi}\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}.
\end{aligned}\tag{79}$$

Hence, (81) and (79) suggest that

$$\mathbb{E}_{\boldsymbol{X}}\left|\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}\big)-\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*\big)\right|\leq\frac{6}{\pi}\frac{\|\boldsymbol{w}_{j_1}^*-\boldsymbol{w}_{j_1}\|}{\|\boldsymbol{w}_{j_1}^*\|}.\tag{80}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left|\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n\right|^2\cdot\left|\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1})-\phi'(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*)\right|\\
=&3\|\boldsymbol{a}_n\|_2^2\cdot\mathbb{E}_{\boldsymbol{X}}\left|\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}\big)-\phi'\big(\boldsymbol{a}_n^T\boldsymbol{X}\boldsymbol{w}_{j_1}^*\big)\right|\\
\leq&\frac{6\|\boldsymbol{a}_n\|_2^2}{\pi}\cdot\frac{\|\boldsymbol{w}_{j_1}-\boldsymbol{w}_{j_1}^*\|_2}{\|\boldsymbol{w}_{j_1}^*\|_2},
\end{aligned}\tag{81}$$

All in all, we have

$$\begin{aligned}\left\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})-\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\right\|_2 \leq & \sum_{j_1}^{K}\sum_{j_2}^{K}\left\|\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W})-\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}^*)\right\|_2 \\ \leq & K^2\max_{j_1,j_2}\left\|\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W})-\frac{\partial^2 f_{\Omega_t}}{\partial \boldsymbol{w}_{j_1}\partial \boldsymbol{w}_{j_2}}(\boldsymbol{W}^*)\right\|_2 \\ \leq & K^2\cdot\frac{12\|\boldsymbol{a}_n\|_2^2}{\pi}\max_j\frac{\|\boldsymbol{w}_j-\boldsymbol{w}_j^*\|_2}{\|\boldsymbol{w}_j^*\|_2} \\ \leq & 4\sigma_1^2(\boldsymbol{A})\frac{\|\boldsymbol{W}^*-\boldsymbol{W}\|_2}{\sigma_K}.\end{aligned} \tag{82}$$

□

### A.3.3. Proof of Lemma 7

*Proof of Lemma 7*. According to the Definitions in (Janson, 2004), there exists a family of $\{(\mathcal{X}_j, w_j)\}_j$, where $\mathcal{X}_j \subseteq \mathcal{X}$ and $w_j \in [0,1]$, such that $\sum_j w_j \sum_{x_{n_j}\in\mathcal{X}_j} x_{n_j} = \sum_{n=1}^N x_n$, and $\sum_j w_j \leq d_{\mathcal{X}}$ by equations (2.1) and (2.2) in (Janson, 2004). Then, let $p_j$ be any positive numbers with $\sum_j p_j = 1$. By Jensen's inequality, for any $s \in \mathbb{R}$, we have

$$e^{s\sum_{n=1}^N x_n} = e^{\sum_j p_j \frac{sw_j}{p_j}X_j} \leq \sum_j p_j e^{\frac{sw_j}{p_j}X_j}, \tag{83}$$

where $X_j = \sum_{x_{n_j}\in\mathcal{X}_j} x_{n_j}$.

Then, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^N x_n} \leq & \mathbb{E}_{\mathcal{X}}\sum_j p_j e^{\frac{sw_j}{p_j}X_j} = \sum_j p_j \prod_{\mathcal{X}_j}\mathbb{E}_{\mathcal{X}} e^{\frac{sw_j}{p_j}x_{n_j}} \\ \leq & \sum_j p_j \prod_{\mathcal{X}_j} e^{\frac{Cw_j^2}{p_j^2}s^2} \\ \leq & \sum_j p_j e^{\frac{C|\mathcal{X}_j|w_j^2}{p_j^2}s^2}.\end{aligned} \tag{84}$$

Let $p_j = \frac{w_j|\mathcal{X}_j|^{1/2}}{\sum_j w_j|\mathcal{X}_j|^{1/2}}$, then we have

$$\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^N x_n} \leq \sum_j p_j e^{C\left(\sum_j w_j|\mathcal{X}_j|^{1/2}\right)^2 s^2} = e^{C\left(\sum_j w_j|\mathcal{X}_j|^{1/2}\right)^2 s^2}. \tag{85}$$

By Cauchy-Schwarz inequality, we have

$$\left(\sum_j w_j|\mathcal{X}_j|^{1/2}\right)^2 \leq \sum_j w_j \sum_j w_j|\mathcal{X}_j| \leq d_{\mathcal{X}}N. \tag{86}$$

Hence, we have

$$\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^N x_n} \leq e^{Cd_{\mathcal{X}}Ns^2}. \tag{87}$$

□

## B. Proof of Theorem 2

Recall that the empirical risk function in (4) is defined as

$$\min_{\boldsymbol{W}}: \quad \hat{f}_{\Omega}(\boldsymbol{W}) = \frac{1}{|\Omega|}\sum_{n\in\Omega} -y_n \log\big(g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big) - (1-y_n)\log\big(1-g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big). \tag{88}$$

The population risk function is defined as

$$\begin{aligned} f_{\Omega}(\boldsymbol{W}) :=& \mathbb{E}_{\boldsymbol{X},y_n}\hat{f}_{\Omega}(\boldsymbol{W}) \\ =& \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{y_n|\boldsymbol{X}}\Big[\frac{1}{|\Omega|}\sum_{n\in\Omega} -y_n \log\big(g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big) - (1-y_n)\log\big(1-g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big)\Big] \\ =& \mathbb{E}_{\boldsymbol{X}}\frac{1}{|\Omega|}\sum_{n\in\Omega} -g(\boldsymbol{W}^*;\boldsymbol{a}_n^T\boldsymbol{X})\log\big(g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big) - (1-g(\boldsymbol{W}^*;\boldsymbol{a}_n^T\boldsymbol{X}))\log\big(1-g(\boldsymbol{W};\boldsymbol{a}_n^T\boldsymbol{X})\big). \end{aligned} \tag{89}$$

The road-map of proof for Theorem 2 follows the similar three steps as those for Theorem 1. The major differences lie in three aspects: (i) in the second step, the objective function $\hat{f}_{\Omega_t}$ is smooth since the activation function $\phi(\cdot)$ is sigmoid. Hence, we can directly apply the mean value theorem as $\nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) = \langle\nabla^2\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}), \boldsymbol{W}^{(t)}-\boldsymbol{W}^*\rangle$ to characterize the effects of the gradient descent term in each iteration, and the error bound of $\nabla^2\hat{f}_{\Omega_t}$ is provided in Lemma 10; (ii) the objective function is the sum of cross-entry loss functions, which have more complex structure of derivatives than those of square loss functions; (iii) as the convergent point may not be the critical point of empirical loss function, we need to provide the distance from the convergent point to the ground-truth parameters additionally, where Lemma 11 is used.

Lemmas 10 and 11 are summarized in the following contents. Also, the notations $\lesssim$ and $\gtrsim$ follow the same definitions as in (27). The proofs of Lemmas 10 and 11 can be found in Appendix B.1 and B.2, respectively.

**Lemma 10.** *For any $\boldsymbol{W}$ that satisfies*

$$\|\boldsymbol{W}-\boldsymbol{W}^*\| \le \frac{2\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \tag{90}$$

*then the second-order derivative of the empirical risk function in (88) for binary classification problems is bounded as*

$$\frac{2(1-\varepsilon_0)}{11\kappa^2\gamma K^2}\sigma_1^2(\boldsymbol{A}) \preceq \nabla^2\hat{f}_{\Omega_t}(\boldsymbol{W}) \preceq \sigma_1^2(\boldsymbol{A}). \tag{91}$$

*provided the number of samples satisfies*

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}(1+\delta^2)\kappa^2\gamma\sigma_1^4(\boldsymbol{A})K^6 d\log N. \tag{92}$$

**Lemma 11.** *Let $\hat{f}_{\Omega_t}$ and $f_{\Omega_t}$ be the empirical and population risk function in (88) and (89) for binary classification problems, respectively, then the first-order derivative of $\hat{f}_{\Omega_t}$ is close to its expectation $f_{\Omega_t}$ with an upper bound as*

$$\|\nabla f_{\Omega_t}(\boldsymbol{W}) - \nabla\hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \lesssim K^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log d}{|\Omega_t|}} \tag{93}$$

*with probability at least $1-K^2N^{-10}$.*

With these preliminary lemmas, the proof of Theorem 2 is formally summarized in the following contents.

*Proof of Theorem 2.* The update rule of $\boldsymbol{W}^{(t)}$ is

$$\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta\nabla\hat{f}_{\Omega_t}(\boldsymbol{W}^{(t)}) + \beta(\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(t-1)}) \tag{94}$$

Since $\widehat{\boldsymbol{W}}$ is a critical point, then we have $\nabla\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}) = 0$. By the intermediate value theorem, we have

$$\begin{aligned} \boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta\nabla^2\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)})(\boldsymbol{W}^{(t)}-\widehat{\boldsymbol{W}}) \\ + \beta(\boldsymbol{W}^{(t)}-\boldsymbol{W}^{(t-1)}) \end{aligned} \tag{95}$$

where $\widehat{\boldsymbol{W}}^{(t)}$ lies in the convex hull of $\boldsymbol{W}^{(t)}$ and $\widehat{\boldsymbol{W}}$.

Next, we have

$$\begin{bmatrix} \boldsymbol{W}^{(t+1)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} - \eta \nabla^2 \hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}) + \beta \boldsymbol{I} & \beta \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{W}^{(t)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^* \end{bmatrix}. \tag{96}$$

Let $\boldsymbol{P}(\beta) = \begin{bmatrix} \boldsymbol{I} - \eta \nabla^2 \hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)}) + \beta \boldsymbol{I} & \beta \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix}$, so we have

$$\left\| \begin{bmatrix} \boldsymbol{W}^{(t+1)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t)} - \boldsymbol{W}^* \end{bmatrix} \right\|_2 = \|\boldsymbol{P}(\beta)\|_2 \left\| \begin{bmatrix} \boldsymbol{W}^{(t)} - \boldsymbol{W}^* \\ \boldsymbol{W}^{(t-1)} - \boldsymbol{W}^* \end{bmatrix} \right\|_2.$$

Then, we have

$$\|\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\|_2 \lesssim \|\boldsymbol{P}(\beta)\|_2 \|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2 \tag{97}$$

Let $\lambda_i$ be the $i$-th eigenvalue of $\nabla^2 \hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}^{(t)})$, and $\delta_i$ be the $i$-th eigenvalue of matrix $\boldsymbol{P}(\beta)$. Following the similar analysis in proof of Theorem 1, we have

$$\delta_i(0) > \delta_i(\beta), \quad \text{for} \quad \forall \beta \in \big(0, (1 - \eta\lambda_i)^2\big). \tag{98}$$

Moreover, $\delta_i$ achieves the minimum $\delta_i^* = |1 - \sqrt{\eta\lambda_i}|$ when $\beta = \big(1 - \sqrt{\eta\lambda_i}\big)^2$.

Let us first assume $\boldsymbol{W}^{(t)}$ satisfies (90) and the number of samples satisfies (92), then from Lemma 10, we know that

$$0 < \frac{2(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2} \le \lambda_i \le \sigma_1^2(\boldsymbol{A}).$$

We define $\gamma_1 = \frac{2(1-\varepsilon_0)\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}$ and $\gamma_2 = \sigma_1^2(\boldsymbol{A})$. Also, for any $\varepsilon_0 \in (0,1)$, we have

$$\nu(\beta^*) = \|\boldsymbol{P}(\beta^*)\|_2 = 1 - \sqrt{\frac{\gamma_1}{2\gamma_2}} = 1 - \sqrt{\frac{1-\varepsilon_0}{11\kappa^2\gamma K}} \tag{99}$$

Let $\beta = 0$, we have

$$\nu(0) = \|\boldsymbol{A}(0)\|_2 = 1 - \frac{1-\varepsilon_0}{11\kappa^2\gamma K}.$$

Hence, with probability at least $1 - K^2 \cdot N^{-10}$, we have

$$\|\boldsymbol{W}^{(t+1)} - \boldsymbol{W}^*\|_2 \le \Big(1 - \sqrt{\frac{1-\varepsilon_0}{11\kappa^2\gamma K}}\Big)\|\boldsymbol{W}^{(t)} - \boldsymbol{W}^*\|_2, \tag{100}$$

provided that $\boldsymbol{W}^{(t)}$ satisfies (25), and

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^2\gamma(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^6 d \log N. \tag{101}$$

According to Lemma 4, we know that (90) holds for $\boldsymbol{W}^{(0)}$ if

$$|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^8\gamma^2(1+\delta^2)K^8 d \log N. \tag{102}$$

Combining (101) and (102), we need $|\Omega_t| \gtrsim \varepsilon_0^{-2}\kappa^8\gamma^2(1+\delta^2)\sigma_1^4(\boldsymbol{A})K^8 d \log N$.

Finally, by the mean value theorem, we have

$$\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}) \le \hat{f}_{\Omega_t}(\boldsymbol{W}^*) + \nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*)^T(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*) + \frac{1}{2}(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*)^T \nabla^2 \hat{f}_{\Omega_t}(\widetilde{\boldsymbol{W}})(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*) \tag{103}$$

for some $\widetilde{\boldsymbol{W}}$ between $\widehat{\boldsymbol{W}}$ and $\boldsymbol{W}^*$. Since $\widehat{\boldsymbol{W}}$ is the local minima, we have $\hat{f}_{\Omega_t}(\widehat{\boldsymbol{W}}) \leq \hat{f}_{\Omega_t}(\boldsymbol{W}^*)$. That is to say

$$\nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*)^T(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*) + \frac{1}{2}(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*)^T \nabla^2 \hat{f}_{\Omega_t}(\widetilde{\boldsymbol{W}})(\widehat{\boldsymbol{W}} - \boldsymbol{W}^*) \leq 0 \tag{104}$$

which implies

$$\frac{1}{2}\|\nabla^2 \hat{f}_{\Omega_t}(\widetilde{\boldsymbol{W}})\|_2 \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_2^2 \leq \|\nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*)\|_2 \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_2. \tag{105}$$

From Lemma 10, we know that

$$\|\nabla^2 \hat{f}_{\Omega_t}(\widetilde{\boldsymbol{W}})\|_2 \geq \frac{2(1-\varepsilon_0)}{11\kappa^2\gamma K^2}\sigma^2(\boldsymbol{A}). \tag{106}$$

From Lemma 11, we know that

$$\|\nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*)\|_2 = \|\nabla \hat{f}_{\Omega_t}(\boldsymbol{W}^*) - \nabla f_{\Omega_t}(\boldsymbol{W}^*)\|_2 \lesssim K^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}. \tag{107}$$

Plugging inequalities (106) and (107) back into (105), we have

$$\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_2 \lesssim (1-\varepsilon_0)^{-1}\kappa^2\gamma K^4\sqrt{\frac{(1+\delta^2)d\log d}{|\Omega_t|}}. \tag{108}$$

□

### B.1. Proof of Lemma 10

The roadmap of proof for Lemma 10 follows the similar steps as those of Lemma 2 for regression problems. Lemmas 12, 13 and 14 are the preliminary lemmas, and their proofs can be found in Appendix B.2. The proof of Lemma 10 is summarized after these preliminary lemmas.

**Lemma 12.** *The second-order derivative of $f_{\Omega_t}$ at the ground truth $\boldsymbol{W}^*$ satisfies*

$$\frac{4\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}\boldsymbol{I} \preceq \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*) \preceq \frac{\sigma_1^2(\boldsymbol{A})}{4}\boldsymbol{I}. \tag{109}$$

**Lemma 13.** *Suppose $f_{\Omega_t}$ is the population loss function with respect to binary classification problems, then we have*

$$\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 \lesssim \|\boldsymbol{W} - \boldsymbol{W}^*\|_2. \tag{110}$$

**Lemma 14.** *Suppose $\hat{f}_{\Omega_t}$ is the empirical loss function with respect to binary classification problems, then the second-order derivative of $\hat{f}_{\Omega_t}$ is close to its expectation with an upper bound as*

$$\|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \lesssim K^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log d}{|\Omega_t|}} \tag{111}$$

*with probability at least $1 - K^2N^{-10}$.*

*Proof of Lemma 10 .* For any $\boldsymbol{W}$, we have

$$\Big|\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 - \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2\Big| \leq \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2. \tag{112}$$

That is

$$\begin{aligned} &\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 \leq \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 + \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 \\ \text{and}\quad &\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 \geq \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 - \|\nabla^2 f_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\|_2 \end{aligned} \tag{113}$$

Then, for any $\boldsymbol{W}$ that satisfies $\|\boldsymbol{W}-\boldsymbol{W}^*\| \le \frac{2\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}$, from Lemmas 12 and 13, we have

$$\frac{2}{11\kappa^2\gamma K^2}\sigma_1^2(\boldsymbol{A}) \preceq \nabla^2 f_{\Omega_t}(\boldsymbol{W}) \preceq \frac{1}{2}\sigma_1^2(\boldsymbol{A}). \tag{114}$$

Next, we have

$$\begin{aligned} &\|\nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \le \|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 + \|\nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 \\ \text{and}\quad &\|\nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \ge \|\nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 - \|\nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W}) - \nabla^2 f_{\Omega_t}(\boldsymbol{W})\|_2 \end{aligned} \tag{115}$$

Then, from (114) and Lemma 14, we have

$$\frac{2(1-\varepsilon_0)}{11\kappa^2\gamma K^2}\sigma_1^2(\boldsymbol{A}) \preceq \nabla^2 \hat{f}_{\Omega_t}(\boldsymbol{W}) \preceq \sigma_1^2(\boldsymbol{A}) \tag{116}$$

provided that the sample size $|\Omega_t| \gtrsim \varepsilon_0^{-2}(1+\delta^2)\kappa^2\gamma\sigma_1^4(\boldsymbol{A})K^6 d\log N$. □

### B.2. Proof of auxiliary lemmas for binary classification problems

#### B.2.1. Proof of Lemma 12

*Proof of Lemma 12* . Since $\mathbb{E}_{\boldsymbol{X}} y_n = g_n(\boldsymbol{W}^*;\boldsymbol{a}_n)$, then we have

$$\begin{aligned} \frac{\partial^2 f_{\Omega_t}(\boldsymbol{W}^*)}{\partial \boldsymbol{w}_j^* \partial \boldsymbol{w}_k^*} =& \mathbb{E}_{\boldsymbol{X}} \frac{\partial^2 \hat{f}_{\Omega_t}(\boldsymbol{W}^*)}{\partial \boldsymbol{w}_j^* \partial \boldsymbol{w}_k^*} \\ =& \mathbb{E}_{\boldsymbol{X}} \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t} \frac{1}{g(\boldsymbol{W};\boldsymbol{a}_n)\big(1-g(\boldsymbol{W};\boldsymbol{a}_n)\big)} \phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T, \end{aligned} \tag{117}$$

for any $j,k\in[K]$.

Then, for any $\boldsymbol{\alpha} = \begin{bmatrix}\boldsymbol{\alpha}_1^T, & \boldsymbol{\alpha}_2^T, & \cdots, & \boldsymbol{\alpha}_K^T\end{bmatrix}^T \in \mathbb{R}^{dk}$ with $\boldsymbol{\alpha}_j \in \mathbb{R}^d$, the lower bound can be obtained from

$$\begin{aligned} \boldsymbol{\alpha}^T\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\boldsymbol{\alpha} =& \mathbb{E}_{\boldsymbol{X}} \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t} \frac{\Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\Big)^2}{g(\boldsymbol{W}^*;\boldsymbol{a}_n)\big(1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)\big)} \\ \ge& \mathbb{E}_{\boldsymbol{X}} \frac{4}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\Big)^2 \\ \ge& \frac{4\sigma_1^2(\boldsymbol{A})}{11\kappa^2\gamma K^2}. \end{aligned} \tag{118}$$

Also, for the upper bound, we have

$$\begin{aligned} \boldsymbol{\alpha}^T\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\boldsymbol{\alpha} =& \mathbb{E}_{\boldsymbol{X}} \frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t} \frac{\Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\Big)^2}{g(\boldsymbol{W}^*;\boldsymbol{a}_n)\big(1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)\big)} \\ =& \mathbb{E}_{\boldsymbol{X}} \frac{1}{|\Omega_t|}\sum_{n\in\Omega_t} \frac{\Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\Big)^2}{\sum_{j_1=1}^K \phi(\boldsymbol{w}_{j_1}^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\sum_{j_2=1}^K\big(1-\phi(\boldsymbol{w}_{j_2}^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\big)} \\ \le& \mathbb{E}_{\boldsymbol{X}} \frac{1}{|\Omega_t|}\sum_{n\in\Omega_t} \frac{\sum_{j=1}^K\big(\boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n\big)^2\sum_{j=1}^K\big(\phi'(\boldsymbol{w}_j^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\big)^2}{\sum_{j_1=1}^K \phi(\boldsymbol{w}_{j_1}^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\sum_{j_2=1}^K\big(1-\phi(\boldsymbol{w}_{j_2}^{*T}\boldsymbol{X}^T\boldsymbol{a}_n)\big)}. \end{aligned} \tag{119}$$

For the denominator item, we have

$$\begin{aligned}\sum_{j_1=1}^{K}\phi({\boldsymbol{w}_{j_1}^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\sum_{j_2=1}^{K}\big(1-\phi({\boldsymbol{w}_{j_2}^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\big)\geq&\sum_{j=1}^{K}\phi({\boldsymbol{w}_j^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\big(1-\phi({\boldsymbol{w}_j^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\big)\\=&\sum_{j=1}^{K}\phi'({\boldsymbol{w}_j^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)\\\geq&4\sum_{j=1}^{K}\phi'({\boldsymbol{w}_j^*}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2.\end{aligned}\tag{120}$$

Hence, we have

$$\boldsymbol{\alpha}^T\nabla^2 f_{\Omega_t}(\boldsymbol{W}^*)\boldsymbol{\alpha}\leq\mathbb{E}_{\boldsymbol{X}}\frac{1}{4|\Omega_t|}\sum_{n\in\Omega_t}\sum_{j=1}^{K}(\boldsymbol{\alpha}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\leq\frac{1}{4}\sigma_1^2(\boldsymbol{A}).\tag{121}$$

□

### B.2.2. PROOF OF LEMMA 13

*Proof of Lemma 13* . Recall that

$$\begin{aligned}&\frac{\partial^2 f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\\=&\mathbb{E}_{\boldsymbol{X}}\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\left(\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\right)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T,\end{aligned}\tag{122}$$

and

$$\begin{aligned}\frac{\partial^2 f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j^2}=&\mathbb{E}_{\boldsymbol{X}}\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\left(\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\right)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)^2(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\\&-\mathbb{E}_{\boldsymbol{X}}\frac{1}{K|\Omega_t|}\sum_{n\in\Omega_t}\left(-\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\right)\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T.\end{aligned}\tag{123}$$

Let us denote $A_{j,k}(\boldsymbol{W};\boldsymbol{a}_n)$ as

$$A_{j,k}(\boldsymbol{W};\boldsymbol{a}_n)=\begin{cases}\frac{1}{K^2}\big(\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\big)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)\\\qquad-\frac{1}{K}\big(-\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\big)\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n),\quad\text{when}\quad j=k;\\\frac{1}{K^2}\big(\frac{g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-g(\boldsymbol{W}^*;\boldsymbol{a}_n)}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\big)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n),\quad\text{when}\quad j\neq k.\end{cases}\tag{124}$$

Further, let us define $M(\boldsymbol{W};\boldsymbol{a}_n)=\max\left\{\frac{2}{K^3}\frac{1}{g^3(\boldsymbol{W};\boldsymbol{a}_n)},\frac{2}{K^3}\frac{1}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^3},\frac{1}{K^2}\frac{1}{g^2(\boldsymbol{W};\boldsymbol{a}_n)},\frac{1}{K^2}\frac{1}{((1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\right\}$.

Then, by the mean value theorem, we have

$$A_{j,k}(\boldsymbol{W};\boldsymbol{a}_n)-A_{j,k}(\boldsymbol{W};\boldsymbol{a}_n)=\sum_{l=1}^{K}\langle\frac{\partial A_{j,k}}{\partial\boldsymbol{w}_l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n),\boldsymbol{w}_l-\boldsymbol{w}_l^*\rangle.\tag{125}$$

For $\frac{\partial A_{j,k}}{\partial\boldsymbol{w}_l}$, we have

$$\frac{\partial A_{j,k}}{\partial\boldsymbol{w}_l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)=B_{j,k,l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)\boldsymbol{X}^T\boldsymbol{a}_n\tag{126}$$

with

$$\begin{aligned}|B_{j,k,l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)|\leq&\frac{2}{K^3}\frac{1}{g^3(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)}+\frac{2}{K^3}\frac{1}{(1-g(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n))^3}+\frac{1}{K^2}\frac{1}{g(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)}+\frac{1}{K^2}\frac{1}{(1-g(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n))^2}\\ \leq&4M(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n).\end{aligned}\tag{127}$$

for all $j\in[K],k\in[K],l\in[K]$.

Therefore, for any $\boldsymbol{\alpha}\in\mathbb{R}^{Kd}$, we have

$$\begin{aligned}&\boldsymbol{\alpha}^T\nabla^2 f_{\Omega_t}(\boldsymbol{W})\boldsymbol{\alpha}\\ \leq&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}\sum_{j=1}^K\sum_{k=1}^K\mathbb{E}_{\boldsymbol{X}}\Big|\boldsymbol{\alpha}_j^T\frac{\partial f_{\Omega_t}}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}(\boldsymbol{W})\boldsymbol{\alpha}_k\Big|\\ =&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}\sum_{j=1}^K\sum_{k=1}^K\mathbb{E}_{\boldsymbol{X}}\Big|\sum_{l=1}^K|B_{j,k,l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)|\langle\boldsymbol{w}_l-\boldsymbol{w}_l^*,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\langle\boldsymbol{\alpha}_j,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\langle\boldsymbol{\alpha}_k,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\Big|\\ =&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}\sum_{j=1}^K\sum_{k=1}^K\Big(\sum_{l=1}^K\mathbb{E}_{\boldsymbol{X}}|B_{j,k,l}(\widetilde{\boldsymbol{W}};\boldsymbol{a}_n)|^2\Big)^{\frac{1}{2}}\Big(\sum_{l=1}^K\mathbb{E}_{\boldsymbol{X}}\big|\langle\boldsymbol{w}_l-\boldsymbol{w}_l^*,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\langle\boldsymbol{\alpha}_j,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\langle\boldsymbol{\alpha}_k,\boldsymbol{X}^T\boldsymbol{a}_n\rangle\big|^2\Big)^{\frac{1}{2}}\\ \leq&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}\sum_{j=1}^K\sum_{k=1}^K36K^{\frac{1}{2}}\Big(\mathbb{E}_{\boldsymbol{X}}M^2(\widehat{\boldsymbol{W}};\boldsymbol{a}_n)\Big)^{\frac{1}{2}}\cdot\Big(\sum_{l=1}^K\|\boldsymbol{w}_l-\boldsymbol{w}_l^*\|_2^2\Big)^{\frac{1}{2}}\|\boldsymbol{a}_j\|_2\|\boldsymbol{a}_k\|_2\\ \leq&\frac{1}{|\Omega_t|}\sum_{n\in|\Omega_t|}36K^3\Big(\mathbb{E}_{\boldsymbol{X}}M^2(\widehat{\boldsymbol{W}};\boldsymbol{a}_n)\Big)^{\frac{1}{2}}\|\boldsymbol{W}-\boldsymbol{W}^*\|_2\\ \overset{(a)}{\lesssim}&e^{\sigma_1^2(\boldsymbol{A})}\|\boldsymbol{W}-\boldsymbol{W}^*\|_2\\ \lesssim&\|\boldsymbol{W}-\boldsymbol{W}^*\|_2,\end{aligned}\tag{128}$$

where $(a)$ comes from Lemma 5 in (Fu et al., 2018). □

### B.2.3. Proof of Lemma 14

*Proof of Lemma 14* . Recall that

$$\begin{aligned}&\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\\ =&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(\frac{y_n}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\Big)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T,\end{aligned}\tag{129}$$

and

$$\begin{aligned}\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j^2}=&\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\Big(\frac{y_n}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\Big)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)^2(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T\\ &-\frac{1}{K|\Omega_t|}\sum_{n\in\Omega_t}\Big(-\frac{y_n}{g(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\Big)\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T.\end{aligned}\tag{130}$$

When $y_n=1$ and $j\neq k$, we have

$$\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}=\frac{1}{K^2|\Omega_t|}\sum_{n\in\Omega_t}\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}(\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T,\tag{131}$$

and

$$\begin{aligned}\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}=&\frac{\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n))\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))}{\big(\frac{1}{K}\sum_{l=1}^K\phi(\boldsymbol{w}_l^T\boldsymbol{X}^T\boldsymbol{a}_n)\big)^2}\\ \le&K^2\frac{\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n))\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))}{\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}\\ =&K^2(1-\phi(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n))(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))\le K^2.\end{aligned}\tag{132}$$

When $y_n=1$ and $j=k$, we have

$$\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}=\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\Big[\frac{1}{K^2}\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1}{K}\frac{\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}\Big](\boldsymbol{X}^T\boldsymbol{a}_n)(\boldsymbol{X}^T\boldsymbol{a}_n)^T,\tag{133}$$

and

$$\Big|\frac{\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}\Big|=\frac{\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))\cdot\big|1-2\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)\big|}{\frac{1}{K}\sum_{l=1}^K\phi(\boldsymbol{w}_l^T\boldsymbol{X}^T\boldsymbol{a}_n)}\le K.\tag{134}$$

Similar to (132) and (134), we can obtain the following inequality for $y_n=0$.

$$\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)}{\big(1-g(\boldsymbol{W};\boldsymbol{a}_n)\big)^2}\le K^2,\quad\text{and}\quad\Big|\frac{\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\Big|\le K.\tag{135}$$

Then, for any $\boldsymbol{\alpha}\in\mathbb{R}^d$, we have

$$\begin{aligned}\boldsymbol{\alpha}^T\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\boldsymbol{\alpha}=&\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}\bigg[\frac{1}{K^2}\bigg(\frac{y_n}{g^2(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{(1-g(\boldsymbol{W};\boldsymbol{a}_n))^2}\bigg)\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\phi'(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)\\ &-\frac{\mathbb{1}_{\{j=k\}}}{K}\bigg(-\frac{y_n}{g(\boldsymbol{W};\boldsymbol{a}_n)}+\frac{1-y_n}{1-g(\boldsymbol{W};\boldsymbol{a}_n)}\bigg)\phi''(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\bigg](\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\\ :=&\frac{1}{|\Omega_t|}\sum_{n\in\Omega_t}H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2.\end{aligned}\tag{136}$$

Next, we show that $H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2$ belongs to the sub-exponential distribution. For any $p\in\mathbb{N}^+$, we have

$$\begin{aligned}\Big(\mathbb{E}_{\boldsymbol{X},\boldsymbol{y}_n}\Big[\big|H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\big|^p\Big]\Big)^{1/p}\le&\Big(\mathbb{E}_{\boldsymbol{X}}\Big[\big|4(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\big|^p\Big]\Big)^{1/p}\\ \le&8\|\boldsymbol{a}_n\|_2^2p\le 8\sigma_1^2(\boldsymbol{A})p\end{aligned}\tag{137}$$

Hence, $H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2$ belongs to the sub-exponential distribution with $\|H_{j,k}(\boldsymbol{a}_n)(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2\|_{\psi_1}=8\sigma_1^2(\boldsymbol{A})$. Then, the moment generation function of $H_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2$ can be bounded as

$$\mathbb{E}e^{sH_{j,k}(\boldsymbol{a}_n)\cdot(\boldsymbol{\alpha}^T\boldsymbol{X}^T\boldsymbol{a}_n)^2}\le e^{C\sigma_1^2(\boldsymbol{A})s^2}\tag{138}$$

for some positive constant $C$ and any $s\in\mathbb{R}$. From Lemma 7 and Chernoff bound, we have

$$\boldsymbol{\alpha}^T\Big(\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}-\frac{\partial^2f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\Big)\boldsymbol{\alpha}\le C\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\tag{139}$$

with probability at least $1-N^{-d}$. By selecting $\xi=\frac{1}{2}$ in Lemmas 8 and 9, we have

$$\bigg\|\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}-\frac{\partial^2f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\bigg\|_2\le C\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}}\tag{140}$$

with probability at least $1-\left(\frac{5}{N}\right)^d$.

In conclusion, we have

$$\begin{aligned}\|\nabla^2 f_{\Omega_t}(\boldsymbol{W})-\nabla^2\hat{f}_{\Omega_t}(\boldsymbol{W})\|_2 \leq & \sum_{j=1}^{K}\sum_{k=1}^{K}\left\|\frac{\partial^2\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}-\frac{\partial^2 f_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_k}\right\|_2 \\ \leq & CK^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log d}{|\Omega_t|}}\end{aligned} \tag{141}$$

with probability at least $1-\left(\frac{5}{d}\right)^d$. □

### B.2.4. Proof of Lemma 11

*Proof of Lemma 11.* Recall that the first-order derivative of $\hat{f}_{\Omega_t}(\boldsymbol{W})$ is calculated from

$$\frac{\partial\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j}=-\frac{1}{K|\Omega_t|}\sum_{n\in\Omega}\frac{y_n-g(\boldsymbol{W};\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)\big(1-g(\boldsymbol{W};\boldsymbol{a}_n)\big)}\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)\boldsymbol{X}^T\boldsymbol{a}_n. \tag{142}$$

Similar to (134), we have

$$\left|\frac{\phi'(\boldsymbol{w}_j^T\boldsymbol{X}^T\boldsymbol{a}_n)}{g(\boldsymbol{W};\boldsymbol{a}_n)}\right|=\frac{\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n)(1-\phi(\boldsymbol{w}_k^T\boldsymbol{X}^T\boldsymbol{a}_n))}{\frac{1}{K}\sum_{l=1}^{K}\phi(\boldsymbol{w}_l^T\boldsymbol{X}^T\boldsymbol{a}_n)}\leq K. \tag{143}$$

Similar to (137), for any fixed $\boldsymbol{\alpha}\in\mathbb{R}^{dK}$, we can show that random variable $\boldsymbol{\alpha}^T\frac{\partial\hat{f}_{\Omega_t}(\boldsymbol{W})}{\partial\boldsymbol{w}_j}$ belongs to sub-exponential distribution with the same bounded norm up to a constant. Hence, by applying Lemma 7 and the Chernoff bound, we have

$$\left\|\nabla f_{\Omega_t}(\boldsymbol{W})-\nabla\hat{f}_{\Omega_t}(\boldsymbol{W})\right\|_2\lesssim K^2\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega_t|}} \tag{144}$$

with probability at least $1-\left(\frac{5}{N}\right)^d$. □

## C. Proof of Lemma 1

*Proof of Lemma 1.* Let $\widetilde{\boldsymbol{A}}$ denote the adjacency matrix, then we have

$$\sigma_1(\widetilde{\boldsymbol{A}})=\max_{\boldsymbol{z}}\frac{\boldsymbol{z}^T\widetilde{\boldsymbol{A}}\boldsymbol{z}}{\boldsymbol{z}^T\boldsymbol{z}}\geq\frac{\mathbf{1}^T\widetilde{\boldsymbol{A}}\mathbf{1}}{\mathbf{1}^T\mathbf{1}}=1+\frac{\sum_{n=1}^{N}\delta_n}{N}, \tag{145}$$

where $\delta_n$ denotes the degree of node $v_n$. Let $\boldsymbol{z}$ be the eigenvetor of the maximum eigenvalue $\sigma_1(\boldsymbol{A})$. Since $\sigma_1(\boldsymbol{A})=\boldsymbol{D}^{-1/2}\widetilde{\boldsymbol{A}}\boldsymbol{D}^{-1/2}$ and $\boldsymbol{D}$ is diagonal matrix, then $\boldsymbol{z}$ is the eigenvector to $\sigma_1(\widetilde{\boldsymbol{A}})$ as well. Then, let $n\in[N]$ be the index of the largest value of vector $\boldsymbol{z}_n$ as $z_n=\|\boldsymbol{z}\|_\infty$, we have

$$\sigma_1(\widetilde{\boldsymbol{A}})=\frac{(\widetilde{\boldsymbol{A}}\boldsymbol{z})_n}{z_n}=\frac{\widetilde{\boldsymbol{a}}_n^T\boldsymbol{z}}{z_n}\leq\frac{\|\boldsymbol{a}_n\|_1\|\boldsymbol{z}\|_\infty}{z_n}=1+\delta. \tag{146}$$

where $\widetilde{\boldsymbol{a}}_n$ is the $n$-th row of $\widetilde{\boldsymbol{A}}$.

Since $\boldsymbol{D}$ is a diagonal matrix with $\|\boldsymbol{D}\|_2\leq 1+\delta$, then we can conclude the inequality in this lemma. □

## D. Proof of Lemma 4

The proof of Lemma 4 is divided into three major parts to bound $I_1$, $I_2$ and $I_3$ in (153). Lemmas 15, 16 and 17 provide the error bounds for $I_1$, $I_2$ and $I_3$, respectively. The proofs of these preliminary lemmas are similar to those of Theorem 5.6 in (Zhong et al., 2017b), the difference is to apply Lemma 7 plus Chernoff inequality instead of standard Hoeffding inequality, and we skip the details of the proofs of Lemmas 15, 16 and 17 here.

**Lemma 15.** *Suppose $\boldsymbol{M}_2$ is defined as in (7) and $\widehat{\boldsymbol{M}}_2$ is the estimation of $\boldsymbol{M}_2$ by samples. Then, with probability $1-N^{-10}$, we have*

$$\|\widehat{\boldsymbol{M}}_2-\boldsymbol{M}_2\|\lesssim\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega|}}, \tag{147}$$

*provided that $|\Omega|\gtrsim(1+\delta^2)d\log^4 N$.*

**Lemma 16.** *Let $\widehat{\boldsymbol{V}}$ be generated by step 4 in Subroutine 1. Suppose $\boldsymbol{M}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})$ is defined as in (9) and $\widehat{\boldsymbol{M}}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})$ is the estimation of $\boldsymbol{M}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})$ by samples. Further, we assume $\boldsymbol{V}\in\mathbb{R}^{d\times K}$ is an orthogonal basis of $\boldsymbol{W}^*$ and satisfies $\|\boldsymbol{V}\boldsymbol{V}^T-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\|\le 1/4$. Then, provided that $N\gtrsim K^5\log^6 d$, with probability at least $1-N^{-10}$, we have*

$$\|\widehat{\boldsymbol{M}}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})-\boldsymbol{M}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})\|\lesssim\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)K^3\log N}{|\Omega|}}. \tag{148}$$

**Lemma 17.** *Suppose $\boldsymbol{M}_1$ is defined as in (6) and $\widehat{\boldsymbol{M}}_1$ is the estimation of $\boldsymbol{M}_1$ by samples. Then, with probability $1-N^{-10}$, we have*

$$\|\widehat{\boldsymbol{M}}_1-\boldsymbol{M}_1\|\lesssim\sigma_1^2(\boldsymbol{A})\sqrt{\frac{(1+\delta^2)d\log N}{|\Omega|}} \tag{149}$$

*provided that $|\Omega|\gtrsim(1+\delta^2)d\log^4 N$.*

**Lemma 18** ((Zhong et al., 2017b), Lemma E.6)**.** *Let $\boldsymbol{V}\in\mathbb{R}^{d\times K}$ be an orthogonal basis of $\boldsymbol{W}^*$ and $\widehat{\boldsymbol{V}}$ be generated by step 4 in Subroutine 1. Assume $\|\widehat{\boldsymbol{M}}_2-\boldsymbol{M}_2\|_2\le\sigma_K(\boldsymbol{M}_2)/10$. Then, for some small $\varepsilon_0$, we have*

$$\|\boldsymbol{V}\boldsymbol{V}^T-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\|_2\le\frac{\|\boldsymbol{M}_2-\widehat{\boldsymbol{M}}_2\|}{\sigma_K(\boldsymbol{M}_2)}. \tag{150}$$

**Lemma 19** ((Zhong et al., 2017b), Lemma E.13)**.** *Let $\boldsymbol{V}\in\mathbb{R}^{d\times K}$ be an orthogonal basis of $\boldsymbol{W}^*$ and $\widehat{\boldsymbol{V}}$ be generated by step 4 in Subroutine 1. Assume $\boldsymbol{M}_1$ can be written in the form of (6) with some homogeneous function $\phi_1$, and let $\widehat{\boldsymbol{M}}_1$ be the estimation of $\boldsymbol{M}_1$ by samples. Let $\widehat{\alpha}$ be the optimal solution of (11) with $\widehat{\overline{\boldsymbol{w}}}_j=\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j$. Then, for each $j\in[K]$, if*

$$\begin{aligned}
T_1&:=\|\boldsymbol{V}\boldsymbol{V}^T-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\|_2\le\frac{1}{\kappa^2\sqrt{K}},\\
T_2&:=\|\widehat{\boldsymbol{u}}_j-\widehat{\boldsymbol{V}}^T\overline{\boldsymbol{w}}_j\|_2\le\frac{1}{\kappa^2\sqrt{K}},\\
T_3&:=\|\widehat{\boldsymbol{M}}_1-\boldsymbol{M}_1\|_2\le\frac{1}{4}\|\boldsymbol{M}_1\|_2,
\end{aligned} \tag{151}$$

*then we have*

$$\Big|\|\boldsymbol{w}_j\|_2-\widehat{\alpha}_j\Big|\le\Big(\kappa^4K^{\frac{3}{2}}\big(T_1+T_2\big)+\kappa^2K^{\frac{1}{2}}T_3\Big)\|\boldsymbol{W}^*\|_2. \tag{152}$$

*Proof of Lemma 4.* we have

$$\begin{aligned}
\|\boldsymbol{w}_j^*-\widehat{\alpha}_j\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2\le&\Big\|\boldsymbol{w}_j^*-\|\boldsymbol{w}_j\|_2\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j+\|\boldsymbol{w}_j\|_2\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j-\widehat{\alpha}_j\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\Big\|_2\\
\le&\Big\|\boldsymbol{w}_j^*-\|\boldsymbol{w}_j\|_2\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2\Big\|_2\Big\|\|\boldsymbol{w}_j\|_2\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j-\widehat{\alpha}_j\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\Big\|_2\\
\le&\|\boldsymbol{w}_j^*\|_2\|\overline{\boldsymbol{w}}_j^*-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2+\Big|\|\boldsymbol{w}_j\|_2-\widehat{\alpha}_j\Big\|_2\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2\\
\le&\sigma_1\big(\|\overline{\boldsymbol{w}}_j^*-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\overline{\boldsymbol{w}}_j^*\|_2+\|\widehat{\boldsymbol{V}}^T\overline{\boldsymbol{w}}_j^*-\widehat{\boldsymbol{u}}_j\|_2\big)+\Big|\|\boldsymbol{w}_j\|_2-\widehat{\alpha}_j\Big|\\
:=&\sigma_1\big(I_1+I_2\big)+I_3.
\end{aligned} \tag{153}$$

From Lemma 18, we have

$$I_1=\|\overline{\boldsymbol{w}}_j^*-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\overline{\boldsymbol{w}}_j^*\|_2\le\|\boldsymbol{V}\boldsymbol{V}^T-\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^T\|_2\le\frac{\|\widehat{\boldsymbol{M}}_2-\boldsymbol{M}_2\|_2}{\sigma_K(\boldsymbol{M}_2)}, \tag{154}$$

where the last inequality comes from Lemma 15. Then, from (7), we know that

$$\sigma_K(\boldsymbol{M}_2) \lesssim \min_{1\le j\le K} \|\boldsymbol{w}_j^*\|_2 \lesssim \sigma_K(\boldsymbol{W}^*). \tag{155}$$

From Theorem 3 in (Kuleshov et al., 2015), we have

$$I_2 = \|\widehat{\boldsymbol{V}}^T \overline{\boldsymbol{w}}_j^* - \widehat{\boldsymbol{u}}_j\|_2 \lesssim \frac{\kappa}{\sigma_K(\boldsymbol{W}^*)} \|\widehat{\boldsymbol{M}}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}}) - \boldsymbol{M}_3(\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}},\widehat{\boldsymbol{V}})\|_2. \tag{156}$$

To guarantee the condition (151) in Lemma 19 hold, according to Lemmas 15 and 16, we need $|\Omega| \gtrsim \kappa^3(1+\delta^2)Kd\log N$. Then, from Lemma 19, we have

$$I_3 = \Big(\kappa^4 K^{3/2}(I_1+I_2) + \kappa^2 K^{1/2}\|\widehat{\boldsymbol{M}}_1 - \boldsymbol{M}_1\|\Big)\|\boldsymbol{W}^*\|_2. \tag{157}$$

Since $d \gg K$, according to Lemmas 15, 16 and 17, we have

$$\|\boldsymbol{w}_j^* - \widehat{\alpha}_j \widehat{\boldsymbol{V}}\widehat{\boldsymbol{u}}_j\|_2 \lesssim \kappa^6 \sigma_1^2(\boldsymbol{A}) \sqrt{\frac{K^3(1+\delta^2)d\log N}{|\Omega|}} \|\boldsymbol{W}^*\|_2 \tag{158}$$

provided $|\Omega| \gtrsim (1+\delta^2)d\log^4 N$. □