

A INITIALIZATION VIA TENSOR METHOD

Following Zhong et al. (2017), we define a special outer product, denoted by $\tilde{\otimes}$. For any vector $\mathbf{v} \in \mathbb{R}^{d_1}$ and $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$,

$$\mathbf{v} \tilde{\otimes} \mathbf{Z} = \sum_{i=1}^{d_2} (\mathbf{v} \otimes \mathbf{z}_i \otimes \mathbf{z}_i + \mathbf{z}_i \otimes \mathbf{v} \otimes \mathbf{z}_i + \mathbf{z}_i \otimes \mathbf{z}_i \otimes \mathbf{v}), \quad (12)$$

where \otimes is the outer product and \mathbf{z}_i is the i -th column of \mathbf{Z} . Recall that $\tilde{\mathbf{x}} = \frac{1}{\sqrt{K}} \sum_{j=1}^K \mathbf{x}_{\Omega_j}$. Next, we define the high order momentum in the following way:

$$\mathbf{M}_1 = \mathbb{E}_{\mathbf{x}} \{y \tilde{\mathbf{x}}\} \in \mathbb{R}^d, \quad (13)$$

$$\mathbf{M}_2 = \mathbb{E}_{\mathbf{x}} \left[y (\tilde{\mathbf{x}} \otimes \tilde{\mathbf{x}} - \mathbb{E}_{\mathbf{x}} \tilde{\mathbf{x}} \tilde{\mathbf{x}}^T) \right] \in \mathbb{R}^{d \times d}, \quad (14)$$

$$\mathbf{M}_3 = \mathbb{E}_{\mathbf{x}} \left[y (\tilde{\mathbf{x}}^{\otimes 3} - \tilde{\mathbf{x}} \tilde{\otimes} \mathbb{E}_{\mathbf{x}} \tilde{\mathbf{x}} \tilde{\mathbf{x}}^T) \right] \in \mathbb{R}^{d \times d \times d}, \quad (15)$$

where $\mathbf{z}^{\otimes 3} := \mathbf{z} \otimes \mathbf{z} \otimes \mathbf{z}$, and $\mathbb{E}_{\mathbf{x}}$ is the expectation over \mathbf{x} .

Following the same calculate formulas in the Claim 5.2 (Zhong et al., 2017), there exist some known constants $\psi_i, i = 1, 2, 3$, such that

$$\mathbf{M}_1 = \sum_{j=1}^K \psi_1 \cdot \|\mathbf{w}_j^*\|_2 \cdot \bar{\mathbf{w}}_j^*, \quad (16)$$

$$\mathbf{M}_2 = \sum_{j=1}^K \psi_2 \cdot \|\mathbf{w}_j^*\|_2 \cdot \bar{\mathbf{w}}_j^* \bar{\mathbf{w}}_j^{*T}, \quad (17)$$

$$\mathbf{M}_3 = \sum_{j=1}^K \psi_3 \cdot \|\mathbf{w}_j^*\|_2 \cdot \bar{\mathbf{w}}_j^{*\otimes 3}, \quad (18)$$

where $\bar{\mathbf{w}}_j^* = \mathbf{w}_j^* / \|\mathbf{w}_j^*\|_2$ in (13)-(15) is the normalization of \mathbf{w}_j^* .

$\mathbf{M}_1, \mathbf{M}_2$ and \mathbf{M}_3 can be estimated through the samples $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, and let $\widehat{\mathbf{M}}_1, \widehat{\mathbf{M}}_2, \widehat{\mathbf{M}}_3$ denote the corresponding estimates. First, we will decompose the rank- k tensor $\widehat{\mathbf{M}}_3$ and obtain the $\{\bar{\mathbf{w}}_j^*\}_{j=1}^K$. By applying the tensor decomposition method Kuleshov et al. (2015) to $\widehat{\mathbf{M}}_3$, the outputs, denoted by $\widehat{\bar{\mathbf{w}}}_j^*$, are the estimations of $\{\bar{\mathbf{w}}_j^*\}_{j=1}^K$. Next, we will estimate $\|\mathbf{w}_j^*\|_2$ through solving the following optimization problem:

$$\widehat{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} : \left| \widehat{\mathbf{M}}_1 - \sum_{j=1}^K \psi_1 \alpha_j \widehat{\bar{\mathbf{w}}}_j^* \right|, \quad (19)$$

From (16) and (19), we know that $|\widehat{\alpha}_j|$ is the estimation of $\|\mathbf{w}_j^*\|_2$. Thus, $\mathbf{W}^{(0)}$ is given as $[\widehat{\alpha}_1 \widehat{\bar{\mathbf{w}}}_1^*, \dots, \widehat{\alpha}_j \widehat{\bar{\mathbf{w}}}_j^*, \dots, \widehat{\alpha}_K \widehat{\bar{\mathbf{w}}}_K^*]$.

To reduce the computational complexity of tensor decomposition, one can project $\widehat{\mathbf{M}}_3$ to a lower-dimensional tensor (Zhong et al., 2017). The idea is to first estimate the subspace spanned by $\{\mathbf{w}_j^*\}_{j=1}^K$, and let $\widehat{\mathbf{V}}$ denote the estimated subspace.

Moreover, we have

$$\mathbf{M}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}}) = \mathbb{E}_{\mathbf{x}} \left[y ((\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^{\otimes 3} - (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}) \tilde{\otimes} \mathbb{E}_{\mathbf{x}} (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}) (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^T) \right] \in \mathbb{R}^{K \times K \times K}, \quad (20)$$

Subroutine 1 Tensor Initialization Method

- 1: **Input:** training data $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$;
- 2: Partition \mathcal{D} into three disjoint subsets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$;
- 3: Calculate $\widehat{\mathbf{M}}_1, \widehat{\mathbf{M}}_2$ following (13), (14) using $\mathcal{D}_1, \mathcal{D}_2$, respectively;
- 4: Obtain the estimate subspace $\widehat{\mathbf{V}}$ of $\widehat{\mathbf{M}}_2$;
- 5: Calculate $\widehat{\mathbf{M}}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}})$ through \mathcal{D}_3 ;
- 6: Obtain $\{\widehat{\mathbf{u}}_j\}_{j=1}^K$ via tensor decomposition method Kuleshov et al. (2015);
- 7: Obtain $\widehat{\alpha}$ by solving optimization problem (19);
- 8: **Return:** $\mathbf{w}_j^{(0)} = \widehat{\alpha}_j \widehat{\mathbf{V}} \widehat{\mathbf{u}}_j, j = 1, \dots, K$.

Then, one can decompose the estimate $\widehat{\mathbf{M}}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}})$ to obtain unit vectors $\{\widehat{\mathbf{u}}_j\}_{j=1}^K \in \mathbb{R}^K$. Since $\overline{\mathbf{w}}^*$ lies in the subspace \mathbf{V} , we have $\mathbf{V}\mathbf{V}^T\overline{\mathbf{w}}^* = \overline{\mathbf{w}}^*$. Then, $\widehat{\mathbf{V}}\widehat{\mathbf{u}}_j$ is an estimate of $\overline{\mathbf{w}}_j^*$. The initialization process is summarized in Subroutine 1.

B NOTATIONS

In this section, we summarize some important notations that will be used in the following proofs. First, recall that the empirical risk function over data $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ in (2) is defined as

$$\hat{f}_{\mathcal{D}}(\mathbf{W}) = \frac{1}{2N} \sum_{n=1}^N \left(\frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^T \mathbf{x}_{n, \Omega_j}) - y_n \right)^2. \quad (21)$$

In addition, the population risk function, which is the expectation of the empirical risk function over the data \mathcal{D} , is defined as

$$\begin{aligned} f(\mathbf{W}) &= \mathbb{E}_{\mathcal{D}} \hat{f}_{\mathcal{D}}(\mathbf{W}) = \mathbb{E}_{\mathcal{D}} \frac{1}{2N} \sum_{n=1}^N \left(\frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^T \mathbf{x}_{n, \Omega_j}) - y_n \right)^2 \\ &= \mathbb{E}_{\mathbf{x}} \frac{1}{2} \left(\frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^T \mathbf{x}_{\Omega_j}) - y \right)^2, \end{aligned} \quad (22)$$

where $\mathbf{x} \in \mathbb{R}^d$ belongs to standard Gaussian distribution, and $y = g(\mathbf{W}^*; \mathbf{x})$. Besides these, we use σ_i to denote the i -th largest singular value of \mathbf{W}^* . Then, κ is defined as σ_1/σ_K , and $\gamma = \prod_{i=1}^K \sigma_i/\sigma_K$. ρ is defined in Property 3.2 (Zhong et al., 2017) and a fixed constant for the ReLU activation function.

Next, to avoid high dimensional tensor in analyzing the second order derivative of the objective function. The proofs will be based on $\text{Vec}(\mathbf{W}) \in \mathbb{R}^{Kr}$, which is the vectorized \mathbf{W} , instead. For notational convenience, we will still use \mathbf{W} in the proofs, but \mathbf{W} is a vector instead of a matrix. Hence, the first order derivative of the empirical risk function $\nabla \hat{f}_{\mathcal{D}} \in \mathbb{R}^{Kr}$, and the second order derivative $\nabla^2 \hat{f}_{\mathcal{D}} \in \mathbb{R}^{Kr \times Kr}$.

Moreover, without special descriptions, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \dots, \boldsymbol{\alpha}_K^T]^T$ stands for any unit vectors that in \mathbb{R}^{Kr} with $\boldsymbol{\alpha}_j \in \mathbb{R}^r$. Therefore, we have

$$\|\nabla^2 \hat{f}_{\mathcal{D}}\|_2 = \max_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}^T \nabla^2 \hat{f}_{\mathcal{D}} \boldsymbol{\alpha}\|_2 = \max_{\boldsymbol{\alpha}} \left(\sum_{j=1}^K \boldsymbol{\alpha}_j^T \frac{\partial \hat{f}_{\mathcal{D}}}{\partial \mathbf{w}_j} \right)^2. \quad (23)$$

Finally, since we focus on order-wise analysis, some constant number will be ignored in majority of the steps. In particular, we use $h_1(z) \gtrsim$ (or \lesssim, \asymp) $h_2(z)$ to denote there exists some positive constant C such that $h_1(z) \geq$ (or $\leq, =$) $C \cdot h_2(z)$ when z is sufficiently large.

C PROOF OF THEOREM 1

The main idea in proving Theorem 1 is to use triangle inequality as shown in (27) by bounding the second order derivative of the population risk function and the distance between the empirical risk and population risk functions. Lemma 3 provides the lower and upper bound for the population risk function, while Lemma 4 provides the error bound between the second order derivation of empirical risk and population risk functions.

Lemma 2 (Weyl's inequality, Bhatia (2013)). *Suppose $\mathbf{B} = \mathbf{A} + \mathbf{E}$ be a matrix with dimension $m \times m$. Let $\lambda_i(\mathbf{B})$ and $\lambda_i(\mathbf{A})$ be the i -th largest eigenvalues of \mathbf{B} and \mathbf{A} , respectively. Then, we have*

$$|\lambda_i(\mathbf{B}) - \lambda_i(\mathbf{A})| \leq \|\mathbf{E}\|_2, \quad \forall i \in [m]. \quad (24)$$

Lemma 3. *Let f be the population risk function in (22). Assume \mathbf{W} satisfies (6), then the second-order derivative of f over \mathbf{W} is bounded as*

$$\frac{(1 - \varepsilon_0)\rho}{11\kappa^2\gamma K^2} \mathbf{I} \leq \nabla^2 f(\mathbf{W}) \leq \frac{7}{K} \mathbf{I}. \quad (25)$$

Lemma 4. *Let $\hat{f}_{\mathcal{D}}$ and f be the empirical and population risk function in (21) and (22), respectively, then the second-order derivative of $\hat{f}_{\mathcal{D}}$ is close to its expectation f with an upper bound as:*

$$\|\nabla^2 \hat{f}_{\mathcal{D}} - \nabla^2 f\|_2 \lesssim \sqrt{\frac{r \log d}{N}}. \quad (26)$$

Proof of Theorem 1 . Let $\hat{\lambda}_{\max}$ and $\hat{\lambda}_{\min}$ denote the largest and smallest eigenvalues of $\nabla^2 \hat{f}_{\mathcal{D}}$, respectively. Also, Let λ_{\max} and λ_{\min} denote the largest and smallest eigenvalues of $\nabla^2 f_{\mathcal{D}}$, respectively.

Then, from Lemma 2, we have

$$\hat{\lambda}_{\max} \leq \lambda_{\max} + \|\nabla^2 \hat{f}_{\mathcal{D}} - \nabla^2 f\|_2 \quad (27)$$

and

$$\hat{\lambda}_{\min} \geq \lambda_{\min} - \|\nabla^2 \hat{f}_{\mathcal{D}} - \nabla^2 f\|_2. \quad (28)$$

When the sample complexity satisfies $N \gtrsim \varepsilon_1^{-2} \rho^{-2} \kappa^4 \gamma^2 K^4 r \log d$, then from Lemma 4, we have

$$\|\nabla^2 \hat{f}_{\mathcal{D}} - \nabla^2 f\|_2 \leq \frac{\varepsilon_1 \rho}{11\kappa^2\gamma K^2}. \quad (29)$$

Then, from (27), (28) and (29), we have

$$\hat{\lambda}_{\max} \leq \frac{8}{K}, \quad (30)$$

and

$$\hat{\lambda}_{\min} \geq \frac{(1 - \varepsilon_0 - \varepsilon_1)\rho}{11\kappa^2\gamma K^2}, \quad (31)$$

which completes the proof. \square

D PROOF OF THEOREM 2

The major idea in proving Theorem 2 is to first characterize the gradient descent term as

$$\begin{aligned}\nabla \hat{f}_{\Omega_t}(\mathbf{W}^{(t)}) &= f_{\Omega_t}(\mathbf{W}^{(t)}) + (\hat{f}_{\Omega_t}(\mathbf{W}^{(t)}) - f_{\Omega_t}(\mathbf{W}^{(t)})) \\ &= \langle \nabla^2 f_{\Omega_t}(\widehat{\mathbf{W}}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \rangle + (\hat{f}_{\Omega_t}(\mathbf{W}^{(t)}) - f_{\Omega_t}(\mathbf{W}^{(t)})),\end{aligned}\quad (32)$$

where the last inequality is obtained through intermediate value theorem, and $\widehat{\mathbf{W}}^{(t)}$ lies in the convex hull of $\mathbf{W}^{(t)}$ and \mathbf{W}^* . The reason that intermediate value theorem is applied on population risk function instead of empirical risk function is the non-smoothness of the empirical risk functions. Due to the non-smoothness of ReLU activation function at zero point, the empirical risk function is not smooth, either. However, the expectation of the empirical risk function over the Gaussian input \mathbf{x} is smooth. Hence, compared with smooth empirical risk function, i.e., neural networks equipped with sigmoid activation function, we have an additional lemma to bound $\nabla \hat{f}_{\mathcal{D}_t}$ to its expectation ∇f , which is summarized in Lemma 5.

The momentum term $\beta(\mathbf{W}^{(t)} - \mathbf{W}^{(t-1)})$ plays an important role in determining the convergence rate, and the recursive rule is obtained in the following way:

$$\begin{bmatrix} \mathbf{W}^{(t+1)} - \mathbf{W}^* \\ \mathbf{W}^{(t)} - \mathbf{W}^* \end{bmatrix} = \mathbf{A}(\beta) \begin{bmatrix} \mathbf{W}^{(t)} - \mathbf{W}^* \\ \mathbf{W}^{(t-1)} - \mathbf{W}^* \end{bmatrix}, \quad (33)$$

where $\mathbf{A}(\beta)$ is a matrix with respect to the value of β and defined in (38). Then, we know iterates $\mathbf{W}^{(t)}$ converge to the ground-truth with a linear rate which is the largest singular value of matrix $\mathbf{A}(\beta)$. 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 singular value of $\mathbf{A}(\beta)$ is smaller than $\mathbf{A}(0)$ for some $\beta > 0$.

Lemma 5. *Let $\hat{f}_{\mathcal{D}}$ and f be the empirical and population risk function in (21) and (22), respectively, then the first-order derivative of $\hat{f}_{\mathcal{D}}$ is close to its expectation f with an upper bound as:*

$$\|\nabla \hat{f}_{\mathcal{D}}(\mathbf{W}) - \nabla f(\mathbf{W})\|_2 \lesssim \sqrt{\frac{r \log d}{N}} (\|\mathbf{W} - \mathbf{W}^*\|_2 + \xi). \quad (34)$$

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

$$\begin{aligned}\mathbf{W}^{(t+1)} &= \mathbf{W}^{(t)} - \eta \nabla \hat{f}_{\mathcal{D}_t}(\mathbf{W}^{(t)}) + \beta(\mathbf{W}^{(t)} - \mathbf{W}^{(t-1)}) \\ &= \mathbf{W}^{(t)} - \eta \nabla f(\mathbf{W}^{(t)}) + \beta(\mathbf{W}^{(t)} - \mathbf{W}^{(t-1)}) + \eta(\nabla f(\mathbf{W}^{(t)}) - \nabla \hat{f}_{\mathcal{D}_t}(\mathbf{W}^{(t)}))\end{aligned}\quad (35)$$

Since $\nabla^2 f$ is a smooth function, by the intermediate value theorem, we have

$$\begin{aligned}\mathbf{W}^{(t+1)} &= \mathbf{W}^{(t)} - \eta \nabla^2 f(\widehat{\mathbf{W}}^{(t)})(\mathbf{W}^{(t)} - \mathbf{W}^*) + \beta(\mathbf{W}^{(t)} - \mathbf{W}^{(t-1)}) \\ &\quad + \eta(\nabla f(\mathbf{W}^{(t)}) - \nabla \hat{f}_{\mathcal{D}_t}(\mathbf{W}^{(t)})),\end{aligned}\quad (36)$$

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

Next, we have

$$\begin{aligned}\begin{bmatrix} \mathbf{W}^{(t+1)} - \mathbf{W}^* \\ \mathbf{W}^{(t)} - \mathbf{W}^* \end{bmatrix} &= \begin{bmatrix} \mathbf{I} - \eta \nabla^2 f(\widehat{\mathbf{W}}^{(t)}) + \beta \mathbf{I} & \beta \mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{W}^{(t)} - \mathbf{W}^* \\ \mathbf{W}^{(t-1)} - \mathbf{W}^* \end{bmatrix} \\ &\quad + \eta \begin{bmatrix} \nabla f(\mathbf{W}^{(t)}) - \nabla \hat{f}_{\mathcal{D}_t}(\mathbf{W}^{(t)}) \\ 0 \end{bmatrix}\end{aligned}\quad (37)$$

Let

$$\mathbf{A}(\beta) = \begin{bmatrix} \mathbf{I} - \eta \nabla^2 f(\widehat{\mathbf{W}}^{(t)}) + \beta \mathbf{I} & \beta \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}, \quad (38)$$

so we have

$$\left\| \begin{bmatrix} \mathbf{W}^{(t+1)} - \mathbf{W}^* \\ \mathbf{W}^{(t)} - \mathbf{W}^* \end{bmatrix} \right\|_2 = \|\mathbf{A}(\beta)\|_2 \left\| \begin{bmatrix} \mathbf{W}^{(t)} - \mathbf{W}^* \\ \mathbf{W}^{(t-1)} - \mathbf{W}^* \end{bmatrix} \right\|_2 + \eta \left\| \begin{bmatrix} \nabla f(\mathbf{W}^{(t)}) - \nabla \hat{f}_{\mathcal{D}_t}(\mathbf{W}^{(t)}) \\ \mathbf{0} \end{bmatrix} \right\|_2. \quad (39)$$

From Lemma 5, we know that

$$\eta \left\| \nabla f(\mathbf{W}^{(t)}) - \nabla \hat{f}_{\mathcal{D}_t}(\mathbf{W}^{(t)}) \right\|_2 \leq C_5 \eta \sqrt{\frac{r \log d}{N_t}} (\|\mathbf{W} - \mathbf{W}^*\|_2 + |\xi|) \quad (40)$$

for some constant $C_5 > 0$. Then, we have

$$\begin{aligned} \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_2 &\leq \left(\|\mathbf{A}(\beta)\|_2 + C_5 \eta \sqrt{\frac{r \log d}{N_t}} \right) \|\mathbf{W}^{(t)} - \mathbf{W}^*\|_2 + C_5 \eta \sqrt{\frac{r \log d}{N_t}} |\xi| \\ &:= \nu(\beta) \|\mathbf{W}^{(t)} - \mathbf{W}^*\|_2 + C_5 \eta \sqrt{\frac{r \log d}{N_t}} |\xi|. \end{aligned} \quad (41)$$

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

$$\mathbf{A}(\beta) := \begin{bmatrix} \mathbf{S}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^T \end{bmatrix} \mathbf{A}(\beta) \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \eta \mathbf{\Lambda} + \beta \mathbf{I} & \beta \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \quad (42)$$

Since $\begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{bmatrix} \begin{bmatrix} \mathbf{S}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^T \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$, we know $\mathbf{A}(\beta)$ and $\mathbf{A}(\beta)$ share the same eigenvalues.

Let λ_i be the i -th eigenvalue of $\nabla^2 f(\widehat{\mathbf{W}}^{(t)})$, then the corresponding i -th eigenvalue of (42), denoted by $\delta_i(\beta)$, satisfies

$$\nu_i^2 - (1 - \eta \lambda_i + \beta) \delta_i + \beta = 0. \quad (43)$$

Then, we have

$$\delta_i(\beta) = \frac{(1 - \eta \lambda_i + \beta) + \sqrt{(1 - \eta \lambda_i + \beta)^2 - 4\beta}}{2}, \quad (44)$$

and

$$|\delta_i(\beta)| = \begin{cases} \sqrt{\beta}, & \text{if } \beta \geq (1 - \sqrt{\eta \lambda_i})^2, \\ \frac{1}{2} \left| (1 - \eta \lambda_i + \beta) + \sqrt{(1 - \eta \lambda_i + \beta)^2 - 4\beta} \right|, & \text{otherwise.} \end{cases} \quad (45)$$

Note that the other root of (43) is abandoned because the root in (44) is always larger than or at least equal to the other root with $|1 - \eta \lambda_i| < 1$. By simple calculation, we have

$$\delta_i(0) > \delta_i(\beta), \quad \text{for } \forall \beta \in (0, (1 - \eta \lambda_i)^2), \quad (46)$$

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

Let us first assume $\mathbf{W}^{(t)}$ satisfies (6), then from Lemma 3, we know that

$$0 < \frac{(1 - \varepsilon_0)}{11\kappa^2\gamma K^2} \leq \lambda_i \leq \frac{7}{K}.$$

Let $\gamma_1 = \frac{\rho(1 - \varepsilon_0)}{11\kappa^2\gamma K^2}$ and $\gamma_2 = \frac{7}{K}$. If we choose β such that

$$\beta^* = \max \left\{ (1 - \sqrt{\eta \gamma_1})^2, (1 - \sqrt{\eta \gamma_2})^2 \right\}, \quad (47)$$

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

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

$$\begin{aligned} \|\mathbf{A}(\beta^*)\|_2 &= \max_i \delta_i(\beta^*) = 1 - \sqrt{\frac{\gamma_1}{2\gamma_2}} = 1 - \sqrt{\frac{1 - \varepsilon_0}{154\rho^{-1}\kappa^2\gamma K}} \\ &\leq 1 - \frac{1 - 3/4 \cdot \varepsilon_0}{\sqrt{154\rho^{-1}\kappa^2\gamma K}}. \end{aligned} \quad (48)$$

Then, let

$$C_5\eta\sqrt{\frac{r \log d}{N_t}} \leq \frac{\varepsilon_0}{4\sqrt{154\rho^{-1}\kappa^2\gamma K}}, \quad (49)$$

we need $N_t \gtrsim \varepsilon_0^{-2}\rho^{-1}\kappa^2\gamma K^3 r \log d$.

Combine (48) and (49), we have

$$\nu(\beta^*) \leq 1 - \frac{1 - \varepsilon_0}{\sqrt{154\rho^{-1}\kappa^2\gamma K}}. \quad (50)$$

While let $\beta = 0$, we have

$$\nu(0) \geq \|\mathbf{A}(0)\|_2 = 1 - \frac{1 - \varepsilon_0}{154\rho^{-1}\kappa^2\gamma K} \quad (51)$$

and

$$\nu(0) \leq \|\mathbf{A}(0)\|_2 + C_5\eta\sqrt{\frac{r \log d}{N_t}} \leq 1 - \frac{1 - 2\varepsilon_0}{154\rho^{-1}\kappa^2\gamma K} \quad (52)$$

if $N_t \gtrsim \varepsilon_0^{-2}\rho^{-1}\kappa^2\gamma K^4 r \log d$.

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

$$\|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_2 \leq \left(1 - \frac{1 - \varepsilon_0}{\sqrt{154\rho^{-1}\kappa^2\gamma K}}\right) \|\mathbf{W}^{(t)} - \mathbf{W}^*\|_2 + 2C\eta\sqrt{\frac{r \log d}{N_t}}|\xi|. \quad (53)$$

if $\mathbf{W}^{(t+1)}$ satisfies (6) and $N_t \gtrsim \varepsilon_0^{-2}\rho^{-1}\kappa^2\gamma K^3 r \log d$.

Then, we can start mathematical induction of (53) over t .

Base case: (6) holds for $\mathbf{W}^{(0)}$ naturally from the assumption in Theorem 2. Since (6) holds and the number of samples exceeds the required bound in (53), we have (53) holds for $t = 0$.

Induction step: Assume (53) holds for t , to make sure the mathematical induction of (53) holds, we need $\mathbf{W}^{(t+1)}$ satisfies (6). That is

$$\eta\sqrt{\frac{d \log d}{N_t}} \lesssim \frac{1 - \varepsilon_0}{\sqrt{132\rho^{-1}\kappa^2\gamma K}} \cdot \frac{\varepsilon_0\sigma_K}{44\rho^{-1}\kappa^2\gamma K^2}. \quad (54)$$

Hence, we need

$$N_t \gtrsim \varepsilon_0^{-2}\rho^{-1}\kappa^8\gamma^3 K^6 d \log d. \quad (55)$$

In addition, with (6) and (53) hold for all $t \leq T$, the following equation

$$\left\| \begin{bmatrix} \mathbf{W}^{(t+1)} - \mathbf{W}^* \\ \mathbf{W}^{(t)} - \mathbf{W}^* \end{bmatrix} \right\|_\infty = \|\mathbf{A}(\beta)\|_2 \left\| \begin{bmatrix} \mathbf{W}^{(t)} - \mathbf{W}^* \\ \mathbf{W}^{(t-1)} - \mathbf{W}^* \end{bmatrix} \right\|_\infty + \eta \left\| \begin{bmatrix} \nabla f(\mathbf{W}^{(t)}) - \nabla \hat{f}_{\mathcal{D}_t}(\mathbf{W}^{(t)}) \\ 0 \end{bmatrix} \right\|_\infty \quad (56)$$

holds as well, and $\|\mathbf{A}(\beta)\|_2$ is bounded by $\nu(\beta)$. Hence, (53) also holds in infinity norm as

$$\|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_\infty \leq \left(1 - \frac{1 - \varepsilon_0}{\sqrt{154\kappa^2\gamma K}}\right) \|\mathbf{W}^{(t)} - \mathbf{W}^*\|_\infty + 2C\eta\sqrt{\frac{r \log d}{N_t}}|\xi|. \quad (57)$$

In conclusion, when $N_t \gtrsim \varepsilon_0^{-2}\kappa^9\gamma^3 K^8 d \log^4 d$, we know that (53) holds for all $1 \leq t \leq T$ with probability at least $1 - K^2 T \cdot d^{-10}$. By simple calculation, we can obtain

$$\|\mathbf{W}^{(T)} - \mathbf{W}^*\|_2 \leq \left(1 - \frac{1 - \varepsilon_0}{\sqrt{132\kappa^2\gamma K}}\right)^T \|\mathbf{W}^{(0)} - \mathbf{W}^*\|_2 + C\sqrt{\frac{\kappa^2\gamma K^2 r \log d}{N_t}} \cdot |\xi|. \quad (58)$$

for some constant $C > 0$. □

E SUPPLEMENTARY PROOF FOR THE STATEMENT IN SECTION 4.1

Suppose $M^{(t)}$ to denote the mask matrix by truncating the smallest $(1 - r^*/d)$ fraction of entries in iterate $\mathbf{W}^{(t)}$. Let M^* denote the ground-truth mask matrix for the teacher network, the following corollary holds from Theorem 2.

Corollary 1. *Suppose the noise $|\xi| \leq \widehat{W}_{\min}^*$ and the number of samples satisfies $N = \Omega(K^8 d \log d \log(1/\varepsilon))$. Let $\{\mathbf{W}^{(t_1)}\}_{t_1=1}^{T_1}$ be the iterates generated from Algorithm 1 by setting $r = d$. Then, for any $T_1 \geq \log(\widehat{W}_{\max}^*/\widehat{W}_{\min}^*)$, we have*

$$M^{(T_1)} = M^*. \quad (59)$$

Proof of Corollary 1. From (57), we know that

$$\|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_\infty \leq \left(1 - \frac{1 - \varepsilon_0}{\sqrt{154\kappa^2\gamma K}}\right) \|\mathbf{W}^{(t)} - \mathbf{W}^*\|_\infty + 2C\eta\sqrt{\frac{d \log d}{N_t}}|\xi|. \quad (60)$$

Hence, we have

$$\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_\infty \leq \left(1 - \frac{1 - \varepsilon_0}{\sqrt{154\kappa^2\gamma K}}\right)^{T_1} \|\mathbf{W}^{(0)} - \mathbf{W}^*\|_\infty + 2C\eta\sqrt{\frac{d \log d}{N_t}}|\xi|. \quad (61)$$

With $T_1 \geq \log(2\widehat{W}_{\max}^*/\widehat{W}_{\min}^*)$, we have

$$\left(1 - \frac{1 - \varepsilon_0}{\sqrt{154\kappa^2\gamma K}}\right)^{T_1} \|\mathbf{W}^{(0)} - \mathbf{W}^*\|_\infty \leq \frac{1}{4}\widehat{W}_{\min}^* \cdot \frac{\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_\infty}{\|\mathbf{W}^*\|_\infty} \leq \frac{1}{4}\widehat{W}_{\min}^*. \quad (62)$$

Since $N = \Omega(K^8 d \log d \log(1/\varepsilon))$ and $|\xi| \leq \widehat{W}_{\min}^*$, we have

$$2C\eta\sqrt{\frac{d \log d}{N_t}}|\xi| \leq \frac{1}{4}\widehat{W}_{\min}^*. \quad (63)$$

From (62) and (63), we know that

$$\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_\infty \leq \frac{1}{2}\widehat{W}_{\min}^*. \quad (64)$$

Therefore, for any entry in $W_{i,j}^{(T_1)}$, if the corresponding entry in augmented ground-truth weights $\widetilde{\mathbf{W}}^*$ is zero, we have

$$|W_{i,j}^{(T_1)}| \leq \frac{1}{2} \widehat{W}_{\min}^*; \quad (65)$$

if the corresponding entry in $\widetilde{\mathbf{W}}^*$ is non-zero, we have

$$|W_{i,j}^{(T_1)}| \geq |\widehat{W}_{i,j}^*| - \frac{1}{2} \widehat{W}_{\min}^* \geq \frac{1}{2} \widehat{W}_{\min}^*. \quad (66)$$

As we know that there are only r^*/d fraction of non-zero weights in the ground-truth model, $\mathbf{M}^{(T_1)} = \mathbf{M}^*$ holds. \square

F PROOF OF LEMMA 1

Instead of providing the proof for Lemma 1, we turn to prove a more general bound for the performance of tensor initialization method as shown in Lemma 6. One easily verify that Lemma 1 holds naturally from Lemma 6. Also, to guarantee the independence among $\hat{f}_{\mathcal{D}}$, the data used in the tensor initialization need to be independent with the data used in AGD.

Lemma 6. Assume the noise level $|\xi| \leq K\sigma_1$ and the number of samples $N \gtrsim \kappa^8 K^8 r \log^6 d$, the tensor initialization method in Subroutine 1 outputs $\mathbf{W}^{(0)}$ such that

$$\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_2 \lesssim \kappa^6 \sqrt{\frac{K^4 r \log d}{N}} (\sigma_1 + |\xi|) \quad (67)$$

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

F.1 PROOF OF LEMMA 6

Lemma 7. Suppose \mathbf{M}_2 is defined as in (14) and $\widehat{\mathbf{M}}_2$ is the estimation of \mathbf{M}_2 by samples $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$. Then, with probability $1 - d^{-10}$, we have

$$\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\| \lesssim \sqrt{\frac{r \log d}{N}} (\sigma_1 + |\xi|), \quad (68)$$

provided that $N \gtrsim r \log^4 d$.

Lemma 8. Let $\widehat{\mathbf{V}}$ be generated by step 4 in Subroutine 1. Suppose $\mathbf{M}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}})$ is defined as in (20) and $\widehat{\mathbf{M}}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}})$ is the estimation of $\mathbf{M}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}})$ by samples $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$. Further, we assume $\mathbf{V} \in \mathbb{R}^{r \times K}$ is an orthogonal basis of \mathbf{W}^* and satisfies $\|\mathbf{V}\mathbf{V}^T - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^T\| \leq 1/4$. Then, provided that $N \gtrsim K^5 \log^6 d$, with probability at least $1 - d^{-10}$, we have

$$\|\widehat{\mathbf{M}}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}}) - \mathbf{M}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}})\| \lesssim \sqrt{\frac{K^3 \log d}{N}} (\sigma_1 + |\xi|). \quad (69)$$

Lemma 9. Suppose \mathbf{M}_1 is defined as in (13) and $\widehat{\mathbf{M}}_1$ is the estimation of \mathbf{M}_1 by samples $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$. Then, with probability $1 - d^{-10}$, we have

$$\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\| \lesssim \sqrt{\frac{r \log d}{N}} (\sigma_1 + |\xi|) \quad (70)$$

provided that $N \gtrsim r \log^4 d$.

Lemma 10 (Tropp (2012), Theorem 1.6). *Consider a finite sequence $\{\mathbf{Z}_k\}$ of independent, random matrices with dimensions $d_1 \times d_2$. Assume that such random matrix satisfies*

$$\mathbb{E}(\mathbf{Z}_k) = 0 \quad \text{and} \quad \|\mathbf{Z}_k\| \leq R \quad \text{almost surely.}$$

Define

$$\delta^2 := \max \left\{ \left\| \sum_k \mathbb{E}(\mathbf{Z}_k \mathbf{Z}_k^*) \right\|, \left\| \sum_k \mathbb{E}(\mathbf{Z}_k^* \mathbf{Z}_k) \right\| \right\}.$$

Then for all $t \geq 0$, we have

$$\text{Prob} \left\{ \left\| \sum_k \mathbf{Z}_k \right\| \geq t \right\} \leq (d_1 + d_2) \exp \left(\frac{-t^2/2}{\delta^2 + Rt/3} \right).$$

Lemma 11 (Zhong et al. (2017), Lemma E.6). *Let $\mathbf{V} \in \mathbb{R}^{r \times K}$ be an orthogonal basis of \mathbf{W}^* and $\widehat{\mathbf{V}}$ be generated by step 4 in Subroutine 1. Assume $\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_2 \leq \sigma_K(\mathbf{M}_2)/10$. Then, we have*

$$\|\mathbf{V}\mathbf{V}^T - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^T\|_2 \leq \frac{\|\mathbf{M}_2 - \widehat{\mathbf{M}}_2\|}{\sigma_K(\mathbf{M}_2)}. \quad (71)$$

Lemma 12 (Zhong et al. (2017), Lemmas E.13 and E.14). *Let $\mathbf{V} \in \mathbb{R}^{r \times K}$ be an orthogonal basis of \mathbf{W}^* and $\widehat{\mathbf{V}}$ be generated by step 4 in Subroutine 1. Assume \mathbf{M}_1 can be written in the form of (16) with some constant ψ_1 , and let $\widehat{\mathbf{M}}_1$ be the estimation of \mathbf{M}_1 by samples $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$. Let $\widehat{\alpha}$ be the optimal solutions of (19) with $\widehat{\mathbf{w}}_j = \widehat{\mathbf{V}}\widehat{\mathbf{u}}_j$. Then, for each $j \in \{1, 2, \dots, K\}$, if*

$$\begin{aligned} T_1 &:= \|\mathbf{V}\mathbf{V}^T - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^T\|_2 \leq \frac{1}{\kappa^2 \sqrt{K}}, \\ T_2 &:= \|\widehat{\mathbf{u}}_j - \widehat{\mathbf{V}}^T \overline{\mathbf{w}}_j\|_2 \leq \frac{1}{\kappa^2 \sqrt{K}}, \\ T_3 &:= \|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2 \leq \frac{1}{4} \|\mathbf{M}_1\|_2, \end{aligned} \quad (72)$$

then we have

$$\left| \alpha_j^* - \widehat{\alpha}_j \right| \leq \left(\kappa^4 K^{\frac{3}{2}} (T_1 + T_2) + \kappa^2 K^{\frac{1}{2}} T_3 \right) |\alpha_j^*|, \quad (73)$$

where $\alpha_j^* = \|\mathbf{w}_j^*\|_2$.

Proof of Lemma 1. By simple calculation, we have

$$\begin{aligned} & \|\mathbf{w}_j^* - |\widehat{\alpha}_j| \widehat{\mathbf{V}}\widehat{\mathbf{u}}_j\|_2 \\ & \leq \left\| \mathbf{w}_j^* - \|\mathbf{w}_j\|_2 \widehat{\mathbf{V}}\widehat{\mathbf{u}}_j + \|\mathbf{w}_j\|_2 \widehat{\mathbf{V}}\widehat{\mathbf{u}}_j - |\widehat{\alpha}_j| \widehat{\mathbf{V}}\widehat{\mathbf{u}}_j \right\|_2 \\ & \leq \left\| \mathbf{w}_j^* - \|\mathbf{w}_j\|_2 \widehat{\mathbf{V}}\widehat{\mathbf{u}}_j \right\|_2 + \left\| \|\mathbf{w}_j\|_2 \widehat{\mathbf{V}}\widehat{\mathbf{u}}_j - |\widehat{\alpha}_j| \widehat{\mathbf{V}}\widehat{\mathbf{u}}_j \right\|_2 \\ & \leq \|\mathbf{w}_j^*\|_2 \|\overline{\mathbf{w}}_j^* - \widehat{\mathbf{V}}\widehat{\mathbf{u}}_j\|_2 + \left| \|\mathbf{w}_j\|_2 - |\widehat{\alpha}_j| \right| \|\widehat{\mathbf{V}}\widehat{\mathbf{u}}_j\|_2 \\ & \leq \sigma_1 \left(\|\overline{\mathbf{w}}_j^* - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^T \overline{\mathbf{w}}_j^*\|_2 + \|\widehat{\mathbf{V}}^T \overline{\mathbf{w}}_j^* - \widehat{\mathbf{u}}_j\|_2 \right) + \left| \|\mathbf{w}_j\|_2 - |\widehat{\alpha}_j| \right| \\ & := \sigma_1 (I_1 + I_2) + I_3. \end{aligned} \quad (74)$$

From Lemma 11, we have

$$I_1 = \|\overline{\mathbf{w}}_j^* - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^T \overline{\mathbf{w}}_j^*\|_2 \leq \|\mathbf{V}\mathbf{V}^T - \widehat{\mathbf{V}}\widehat{\mathbf{V}}^T\|_2 \leq \frac{\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_2}{\sigma_K(\mathbf{M}_2)}, \quad (75)$$

where the last inequality comes from Lemma 7. Then, from (17), we know that

$$\sigma_K(\mathbf{M}_2) \lesssim \min_{1 \leq j \leq K} \|\mathbf{w}_j\|_2 \lesssim \sigma_K. \quad (76)$$

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

$$I_2 = \|\widehat{\mathbf{V}}^T \overline{\mathbf{w}}_j^* - \widehat{\mathbf{u}}_j\|_2 \lesssim \frac{\kappa}{\sigma_K} \|\widehat{\mathbf{M}}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}}) - \mathbf{M}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}})\|_2. \quad (77)$$

To guarantee the condition (72) in Lemma 12 hold, according to Lemmas 7 and 8, we need $N \gtrsim \kappa^3 K r \log d$. Then, from Lemma 12, we have

$$I_3 = \left(\kappa^4 K^{3/2} (I_1 + I_2) + \kappa^2 K^{1/2} \|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\| \right) \sigma_1. \quad (78)$$

When $r \gg K$, according to Lemmas 7, 8 and 9, we have

$$\|\mathbf{w}_j^* - |\widehat{\alpha}_j| \widehat{\mathbf{V}} \widehat{\mathbf{u}}_j\|_2 \lesssim \kappa^6 \sqrt{\frac{K^3 r \log d}{N}} (\sigma_1 + |\xi|) \quad (79)$$

provided that $N \gtrsim r \log^4 d$.

In conclusion, we have

$$\|\mathbf{W}^* - \mathbf{W}^{(0)}\|_2 \leq \sqrt{K} \cdot \|\mathbf{w}_j^* - |\widehat{\alpha}_j| \widehat{\mathbf{V}} \widehat{\mathbf{u}}_j\|_2 \lesssim \kappa^6 \sqrt{\frac{K^4 r \log d}{N}} (\sigma_1 + |\xi|). \quad (80)$$

□

G ADDITIONAL PROOF OF THE LEMMAS IN APPENDIX C

G.1 PROOF OF LEMMA 3

The eigenvalues of $\nabla^2 f$ at any fixed point \mathbf{W} is bounded through the ones at the ground truth \mathbf{W}^* by using Lemma 2. The eigenvalues of $\nabla^2 f$ at ground truth \mathbf{W}^* is bounded in (83) and (84).

Lemma 13. *Let f be the population risk function in (22) and \mathbf{W} satisfy (6), then we have*

$$\|\nabla^2 f(\mathbf{W}) - \nabla^2 f(\mathbf{W}^*)\|_2 \leq \frac{4\|\mathbf{W}^* - \mathbf{W}\|_2}{\sigma_K}. \quad (81)$$

Proof of Lemma 3. Let $\lambda_{\max}(\mathbf{W})$ and $\lambda_{\min}(\mathbf{W})$ denote the largest and smallest eigenvalues of $\nabla^2 f_{\mathcal{D}}$ at point \mathbf{W} , respectively. Then, from Lemma 2, we have

$$\begin{aligned} \lambda_{\max}(\mathbf{W}) &\leq \lambda_{\max}(\mathbf{W}^*) + \|\nabla^2 f(\mathbf{W}) - \nabla^2 f(\mathbf{W}^*)\|_2, \\ \text{and } \lambda_{\min}(\mathbf{W}) &\geq \lambda_{\min}(\mathbf{W}^*) - \|\nabla^2 f(\mathbf{W}) - \nabla^2 f(\mathbf{W}^*)\|_2. \end{aligned} \quad (82)$$

Next, we provide the lower bound of Hessian of population function at ground truth \mathbf{W}^* . For any $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \dots, \boldsymbol{\alpha}_K^T]^T$ with $\boldsymbol{\alpha}_j \in \mathbb{R}^r$, we have

$$\begin{aligned}
\min_{\|\boldsymbol{\alpha}\|_2=1} \boldsymbol{\alpha}^T \nabla^2 f(\mathbf{W}^*) \boldsymbol{\alpha} &= \frac{1}{K^2} \min_{\|\boldsymbol{\alpha}\|_2=1} \mathbb{E}_{\mathbf{x}} \left(\sum_{j=1}^K \boldsymbol{\alpha}_j^T \mathbf{x}_j \phi'(\mathbf{w}_j^{*T} \mathbf{x}_j) \right)^2 \\
&= \frac{1}{K^2} \min_{\|\tilde{\boldsymbol{\alpha}}\|_2=1, \text{supp}(\tilde{\boldsymbol{\alpha}}_j) = \text{supp}(\tilde{\mathbf{w}}_j^*)} \mathbb{E}_{\mathbf{x}} \left(\sum_{j=1}^K \tilde{\boldsymbol{\alpha}}_j^T \mathbf{x}_j \phi'(\tilde{\mathbf{w}}_j^{*T} \mathbf{x}_j) \right)^2 \\
&\geq \frac{1}{K^2} \min_{\|\tilde{\boldsymbol{\alpha}}\|_2=1} \mathbb{E}_{\mathbf{x}} \left(\sum_{j=1}^K \tilde{\boldsymbol{\alpha}}_j^T \mathbf{x}_j \phi'(\tilde{\mathbf{w}}_j^{*T} \mathbf{x}_j) \right)^2 \\
&\geq \frac{\rho}{11\kappa^2\lambda K^2},
\end{aligned} \tag{83}$$

where $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^{Kd}$ with $\tilde{\boldsymbol{\alpha}}_j \in \mathbb{R}^d$, and the last inequality comes from Lemma D.6 (Zhong et al., 2017).

Next, the upper bound of Hessian of population function at ground truth \mathbf{W}^* can be bounded in the following way. For any $\boldsymbol{\alpha}$, we have

$$\begin{aligned}
\boldsymbol{\alpha}^T \nabla^2 f(\mathbf{W}^*) \boldsymbol{\alpha} &= \frac{1}{K^2} \mathbb{E}_{\mathbf{x}} \left(\sum_{j=1}^K \boldsymbol{\alpha}_j^T \mathbf{x}_j \phi'(\mathbf{w}_j^{*T} \mathbf{x}_j) \right)^2 \leq \frac{2}{K^2} \cdot \mathbb{E}_{\mathbf{x}} \sum_{j=1}^K \left(\boldsymbol{\alpha}_j^T \mathbf{x}_j \phi'(\mathbf{w}_j^{*T} \mathbf{x}_j) \right)^2 \\
&= \frac{2}{K^2} \sum_{j=1}^K \mathbb{E}_{\mathbf{x}} \left(\boldsymbol{\alpha}_j^T \mathbf{x}_j \phi'(\mathbf{w}_j^{*T} \mathbf{x}_j) \right)^2 \\
&\leq \frac{2}{K^2} \sum_{j=1}^K \left(\mathbb{E}_{\mathbf{x}} (\boldsymbol{\alpha}_j^T \mathbf{x}_j)^4 \mathbb{E}_{\mathbf{x}} |\phi'|^4 \right)^{\frac{1}{2}} \\
&\leq \frac{2}{K^2} \cdot K \cdot 3 = \frac{6}{K}.
\end{aligned} \tag{84}$$

Then, from Lemma 13, when \mathbf{W} satisfies (6), we have that

$$\|\nabla^2 f(\mathbf{W}) - \nabla^2 f(\mathbf{W}^*)\|_2 \leq \frac{\varepsilon_0 \rho}{11\kappa^2 \gamma}. \tag{85}$$

Hence, from (82) and (85), we have that

$$\frac{(1 - \varepsilon_0)\rho}{11\kappa^2 \gamma K^2} \mathbf{I} \leq \nabla^2 f(\mathbf{W}) \leq \frac{7}{K} \mathbf{I}. \tag{86}$$

□

G.2 PROOF OF LEMMA 4

We first show that the second order derivative of $\hat{f}_{\mathcal{D}}$ is a sum of several random sub-exponential variables as shown in (93). Then, by concentration theory, i.e., Chernoff bound, we can show that the error bound of $\nabla^2 \hat{f}_{\mathcal{D}}$ to its expectation.

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

$$(\mathbb{E}|X|^p)^{1/p} \leq c_1 \sqrt{p} \tag{87}$$

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

$$\mathbb{E}e^{s(X-\mathbb{E}X)} \leq e^{c_2\|X\|_{\psi_2}^2 s^2} \quad (88)$$

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

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

$$(\mathbb{E}|X|^p)^{1/p} \leq c_3 p \quad (89)$$

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

$$\mathbb{E}e^{s(X-\mathbb{E}X)} \leq e^{c_4\|X\|_{\psi_1}^2 s^2} \quad (90)$$

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 $\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$.

Proof of Lemma 4. Recall the definition of f and \hat{f} in (22) and (21), we have

$$\begin{aligned} & \frac{\partial^2 f}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} - \frac{\partial^2 \hat{f}_{\mathcal{D}}}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} \\ &= \mathbb{E}_{\mathbf{x}} \left[\phi'(\mathbf{w}_{j_1}^T \mathbf{x}_{\Omega_{j_1}}) \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{\Omega_{j_2}}) \mathbf{x}_{\Omega_{j_1}} \mathbf{x}_{\Omega_{j_2}}^T - \frac{1}{N} \sum_{n=1}^N \phi(\mathbf{w}_{j_1}^T \mathbf{x}_{n, \Omega_{j_1}}) \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{n, \Omega_{j_2}}) \mathbf{x}_{n, \Omega_{j_1}} \mathbf{x}_{n, \Omega_{j_2}}^T \right]. \end{aligned} \quad (91)$$

For any $\boldsymbol{\alpha}$, we have

$$\begin{aligned} \|\nabla^2 f - \nabla^2 \hat{f}_{\mathcal{D}}\|_2 &= \max_{\|\boldsymbol{\alpha}\|_2=1} \left| \boldsymbol{\alpha}^T (\nabla^2 f - \nabla^2 \hat{f}_{\mathcal{D}}) \boldsymbol{\alpha} \right| \\ &= \sum_{j_1=1}^K \sum_{j_2=1}^K \max_{\|\boldsymbol{\alpha}\|_2=1} \left| \boldsymbol{\alpha}_{j_1}^T \left(\frac{\partial^2 f}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} - \frac{\partial^2 \hat{f}_{\mathcal{D}}}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} \right) \boldsymbol{\alpha}_{j_2} \right| \\ &= \frac{1}{K^2} \sum_{j_1=1}^K \sum_{j_2=1}^K \max_{\|\boldsymbol{\alpha}\|_2=1} \mathbb{E}_{\mathbf{x}} \left[\phi'(\mathbf{w}_{j_1}^T \mathbf{x}_{\Omega_{j_1}}) \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{\Omega_{j_2}}) \boldsymbol{\alpha}_{j_1}^T \mathbf{x}_{\Omega_{j_1}} \boldsymbol{\alpha}_{j_2}^T \mathbf{x}_{\Omega_{j_2}} \right. \\ & \quad \left. - \frac{1}{N} \sum_{n=1}^N \phi(\mathbf{w}_{j_1}^T \mathbf{x}_{n, \Omega_{j_1}}) \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{n, \Omega_{j_2}}) \boldsymbol{\alpha}_{j_1}^T \mathbf{x}_{n, \Omega_{j_1}} \boldsymbol{\alpha}_{j_2}^T \mathbf{x}_{n, \Omega_{j_2}} \right]. \end{aligned} \quad (92)$$

Then, define $Z_n(j_1, j_2) = \phi(\mathbf{w}_{j_1}^T \mathbf{x}_{n, \Omega_{j_1}}) \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{n, \Omega_{j_2}}) \boldsymbol{\alpha}_{j_1}^T \mathbf{x}_{n, \Omega_{j_1}} \boldsymbol{\alpha}_{j_2}^T \mathbf{x}_{n, \Omega_{j_2}}$, and we say Z belongs to sub-Exponential distribution. According to Definition 2, we have

$$\begin{aligned} (\mathbb{E}|Z_n|^p)^{1/p} &\leq \left(\mathbb{E} \left| (\boldsymbol{\alpha}_{j_1}^T \mathbf{x}_{n, \Omega_{j_1}}) \cdot (\boldsymbol{\alpha}_{j_2}^T \mathbf{x}_{n, \Omega_{j_2}}) \right|^p \right)^{1/p} \\ &\leq \left(\mathbb{E} \left| (\boldsymbol{\alpha}_{j_1}^T \mathbf{x}_{n, \Omega_{j_1}}) \right|^{2p} \right)^{1/(2p)} \cdot \left(\mathbb{E} \left| (\boldsymbol{\alpha}_{j_2}^T \mathbf{x}_{n, \Omega_{j_2}}) \right|^{2p} \right)^{1/(2p)} \\ &\leq C_{\mathbf{x}} \cdot \sqrt{2p} \cdot C_{\mathbf{x}} \sqrt{2p} \\ &= 2C_{\mathbf{x}}^2 \cdot p. \end{aligned} \quad (93)$$

Then, we have

$$\mathbb{E}_{Z_n} e^{s(Z_n - \mathbb{E}Z_n)} \leq e^{-Cs^2} \quad (94)$$

for some constant $C > 0$ and any $s \in \mathbb{R}$. From Chernoff bound, we have

$$\text{Prob}\left\{\left|\frac{1}{N} \sum_{n=1}^N (Z_n - \mathbb{E}Z_n)\right| < t\right\} \leq 1 - \frac{e^{-Cs^2}}{e^{st}}. \quad (95)$$

Let us select $t = \sqrt{\frac{r \log d}{N}}$ and $s = \frac{C}{2} \cdot t$, then we have

$$\left|\frac{1}{N} \sum_{n=1}^N (Z_n - \mathbb{E}Z_n)\right| \lesssim \sqrt{\frac{r \log d}{N}} \quad (96)$$

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

Hence, we have

$$\max_{\|\boldsymbol{\alpha}\|_2=1} \left| \boldsymbol{\alpha}_{j_1}^T \left(\frac{\partial^2 f}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} - \frac{\partial^2 \hat{f}_\Omega}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} \right) \boldsymbol{\alpha}_{j_2} \right| \lesssim \sqrt{\frac{r \log d}{N}}, \quad (97)$$

and

$$\|\nabla^2 f(\mathbf{W}) - \nabla^2 \hat{f}_\Omega(\mathbf{W})\|_2 \lesssim \sqrt{\frac{r \log d}{N}} \quad (98)$$

with probability at least $1 - d^{-r}$. \square

H PROOF OF LEMMA 5

Proof of Lemma 5. The first-order derivative of the empirical risk function is written as

$$\begin{aligned} \frac{\partial \hat{f}_\mathcal{D}}{\partial \mathbf{w}_k} &= \frac{1}{K \cdot N} \sum_{n=1}^N \left(y_n - \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^T \mathbf{x}_{n, \Omega_j}) \right) \mathbf{x}_{n, \Omega_j} \phi'(\mathbf{w}_k^T \mathbf{x}_{n, \Omega_j}) \\ &= \frac{1}{K^2 \cdot N} \sum_{n=1}^N \sum_{j=1}^K (\phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) - \phi(\mathbf{w}_j^T \mathbf{x}_{n, \Omega_j})) \mathbf{x}_{n, \Omega_j} \phi'(\mathbf{w}_k^T \mathbf{x}_{n, \Omega_j}) \\ &\quad + \frac{1}{K^2 \cdot N} \sum_{j=1}^K \xi_n \mathbf{x}_{n, \Omega_j} \phi'(\mathbf{w}_k^T \mathbf{x}_{n, \Omega_j}) \end{aligned} \quad (99)$$

Define $\mathbf{z}_n(j, k) = (\phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) - \phi(\mathbf{w}_j^T \mathbf{x}_{n, \Omega_j})) \phi'(\mathbf{w}_k^T \mathbf{x}_{n, \Omega_j}) \mathbf{x}_{n, \Omega_j}$. Then, for any $\boldsymbol{\alpha}_j \in \mathbb{R}^r$, we have

$$\begin{aligned} p^{-1} \left(\mathbb{E}_{\mathbf{x}} |\boldsymbol{\alpha}_j^T \mathbf{z}_n|^p \right)^{\frac{1}{p}} &= p^{-1} \left(\mathbb{E}_{\mathbf{x}} |(\boldsymbol{\alpha}_j^T \mathbf{x}_{n, \Omega_j}) (\phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) - \phi(\mathbf{w}_j^T \mathbf{x}_{n, \Omega_j})) \phi'(\mathbf{w}_k^T \mathbf{x}_{n, \Omega_j})|^p \right)^{\frac{1}{p}} \\ &\leq p^{-1} \left(\mathbb{E}_{\mathbf{x}} |(\boldsymbol{\alpha}_j^T \mathbf{x}_{n, \Omega_j}) (\phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) - \phi(\mathbf{w}_j^T \mathbf{x}_{n, \Omega_j}))|^p \right)^{\frac{1}{p}} \\ &\leq p^{-1} \left(\mathbb{E}_{\mathbf{x}} |\boldsymbol{\alpha}_j^T \mathbf{x}_{n, \Omega_j}|^{2p} \right)^{\frac{1}{2p}} \cdot \left(\mathbb{E}_{\mathbf{x}} |\phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) - \phi(\mathbf{w}_j^T \mathbf{x}_{n, \Omega_j})|^{2p} \right)^{\frac{1}{2p}} \\ &\leq p^{-1} \left(\mathbb{E}_{\mathbf{x}} |\boldsymbol{\alpha}_j^T \mathbf{x}_{n, \Omega_j}|^{2p} \right)^{\frac{1}{2p}} \cdot \left(\mathbb{E}_{\mathbf{x}} |(\mathbf{w}_j^* - \mathbf{w}_j)^T \mathbf{x}_{n, \Omega_j}|^{2p} \right)^{\frac{1}{2p}} \\ &\leq 2 \|\mathbf{w}_j^* - \mathbf{w}_j\|_2. \end{aligned} \quad (100)$$

Following similar steps in (95), by Chernoff bound, we have

$$\left\| \frac{1}{N} \sum_{n=1}^N (\mathbf{z}_n - \mathbb{E}_{\mathbf{x}} \mathbf{z}_n) \right\|_2 \lesssim \sqrt{\frac{r \log d}{N}} \cdot \|\mathbf{w}_j^* - \mathbf{w}_j\|_2 \quad (101)$$

with probability at least $1 - d^{-r}$. Also, we know that $\mathbf{x}_{n,\Omega_j} \phi'(W_k^T \mathbf{x}_{n,\Omega_j})$ belongs to sub-Gaussian distribution as well. Then, by Chernoff bound, we have

$$\begin{aligned} \left\| \frac{1}{N} \sum_{n=1}^N \xi_n \mathbf{x}_{n,\Omega_j} \phi'(\mathbf{w}_j^T \mathbf{x}_{n,\Omega_j}) \right\|_2 &\lesssim |\xi| \cdot \left\| \frac{1}{N} \sum_{n=1}^N \mathbf{x}_{n,\Omega_j} \phi'(\mathbf{w}_j^T \mathbf{x}_{n,\Omega_j}) \right\|_2 \\ &\lesssim |\xi| \cdot \sqrt{\frac{r \log d}{N}} \end{aligned} \quad (102)$$

with probability at least $1 - d^{-r}$. \square

I PROOF OF LEMMA 13

Proof of Lemma 13. Recall the definition of population risk function, we have

$$\frac{\partial^2 f(\mathbf{W}^*)}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} = \frac{1}{K^2} \mathbb{E}_{\mathbf{x}} \phi'(\mathbf{w}_{j_1}^{*T} \mathbf{x}_{\Omega_{j_1}}) \phi'(\mathbf{w}_{j_2}^{*T} \mathbf{x}_{\Omega_{j_2}}) \mathbf{x}_{\Omega_{j_1}} \mathbf{x}_{\Omega_{j_2}}^T \quad (103)$$

and

$$\frac{\partial^2 f(\mathbf{W})}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} = \frac{1}{K^2} \mathbb{E}_{\mathbf{x}} \phi'(\mathbf{w}_{j_1}^T \mathbf{x}_{\Omega_{j_1}}) \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{\Omega_{j_2}}) \mathbf{x}_{\Omega_{j_1}} \mathbf{x}_{\Omega_{j_2}}^T \quad (104)$$

Then, we have

$$\begin{aligned} &\frac{\partial^2 f(\mathbf{W}^*)}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} - \frac{\partial^2 f(\mathbf{W})}{\partial \mathbf{w}_{j_1} \partial \mathbf{w}_{j_2}} \\ &= \frac{1}{K^2} \mathbb{E}_{\mathbf{x}} \left[\phi'(\mathbf{w}_{j_1}^{*T} \mathbf{x}_{\Omega_{j_1}}) \phi'(\mathbf{w}_{j_2}^{*T} \mathbf{x}_{\Omega_{j_2}}) - \phi'(\mathbf{w}_{j_1}^T \mathbf{x}_{\Omega_{j_1}}) \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{\Omega_{j_2}}) \right] \mathbf{x}_{\Omega_{j_1}} \mathbf{x}_{\Omega_{j_2}}^T \\ &= \frac{1}{K^2} \mathbb{E}_{\mathbf{x}} \left[\phi'(\mathbf{w}_{j_1}^{*T} \mathbf{x}_{\Omega_{j_1}}) (\phi'(\mathbf{w}_{j_2}^{*T} \mathbf{x}_{\Omega_{j_2}}) - \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{\Omega_{j_2}})) + \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{\Omega_{j_2}}) (\phi'(\mathbf{w}_{j_1}^{*T} \mathbf{x}_{\Omega_{j_1}}) - \phi'(\mathbf{w}_{j_1}^T \mathbf{x}_{\Omega_{j_1}})) \right] \mathbf{x}_{\Omega_{j_1}} \mathbf{x}_{\Omega_{j_2}}^T \\ &= \frac{1}{K^2} \left[\mathbb{E}_{\mathbf{x}} \phi'(\mathbf{w}_{j_1}^{*T} \mathbf{x}_{\Omega_{j_1}}) (\phi'(\mathbf{w}_{j_2}^{*T} \mathbf{x}_{\Omega_{j_2}}) - \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{\Omega_{j_2}})) \mathbf{x}_{\Omega_{j_1}} \mathbf{x}_{\Omega_{j_2}}^T \right. \\ &\quad \left. + \mathbb{E}_{\mathbf{x}} \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{\Omega_{j_2}}) (\phi'(\mathbf{w}_{j_1}^{*T} \mathbf{x}_{\Omega_{j_1}}) - \phi'(\mathbf{w}_{j_1}^T \mathbf{x}_{\Omega_{j_1}})) \mathbf{x}_{\Omega_{j_1}} \mathbf{x}_{\Omega_{j_2}}^T \right] \\ &:= \frac{1}{K^2} (\mathbf{I}_1 + \mathbf{I}_2). \end{aligned} \quad (105)$$

For any $\boldsymbol{\alpha}_{j_1}$ and $\boldsymbol{\alpha}_{j_2} \in \mathbb{R}^r$, we have

$$\begin{aligned} &\max_{\|\boldsymbol{\alpha}_{j_1}\|_2, \|\boldsymbol{\alpha}_{j_2}\|_2=1} \boldsymbol{\alpha}_{j_1}^T \mathbf{I}_1 \boldsymbol{\alpha}_{j_2} \\ &= \max_{\|\boldsymbol{\alpha}_{j_1}\|_2, \|\boldsymbol{\alpha}_{j_2}\|_2=1} \mathbb{E}_{\mathbf{x}} \phi'(\mathbf{w}_{j_1}^{*T} \mathbf{x}_{\Omega_{j_1}}) (\phi'(\mathbf{w}_{j_2}^{*T} \mathbf{x}_{\Omega_{j_2}}) - \phi'(\mathbf{w}_{j_2}^T \mathbf{x}_{\Omega_{j_2}})) \cdot (\boldsymbol{\alpha}_{j_1}^T \mathbf{x}_{\Omega_{j_1}}) \cdot (\boldsymbol{\alpha}_{j_2}^T \mathbf{x}_{\Omega_{j_2}}) \\ &\leq \max_{\|\boldsymbol{\alpha}\|_2=1} \mathbb{E}_{\mathbf{x}} \phi'(\tilde{\mathbf{w}}_{j_1}^{*T} \mathbf{x}) (\phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x})) \cdot (\boldsymbol{\alpha}^T \mathbf{x})^2, \end{aligned} \quad (106)$$

where $\mathbf{a} \in \mathbb{R}^d$. Let $I = \phi'(\tilde{\mathbf{w}}_{j_1}^{*T} \mathbf{x})(\phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x})) \cdot (\mathbf{a}^T \mathbf{x})^2$. It is easy to verify there exists a basis such that $\mathcal{B} = \{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{a}_4^\perp, \dots, \mathbf{a}_d^\perp\}$ with $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ spans a subspace that contains $\mathbf{a}, \mathbf{w}_{j_2}$ and $\mathbf{w}_{j_2}^*$. Then, for any \mathbf{x} , we have a unique $\mathbf{z} = [z_1 \ z_2 \ \dots \ z_d]^T$ such that

$$\mathbf{x} = z_1 \mathbf{a} + z_2 \mathbf{b} + z_3 \mathbf{c} + \dots + z_d \mathbf{a}_d^\perp.$$

Also, since $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, we have $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Then, we have

$$\begin{aligned} I &= \mathbb{E}_{z_1, z_2, z_3} |\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})| \cdot |\mathbf{a}^T \mathbf{x}|^2 \\ &= \int |\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})| \cdot |\mathbf{a}^T \mathbf{x}|^2 \cdot f_Z(z_1, z_2, z_3) dz_1 dz_2 dz_3, \end{aligned}$$

where $\mathbf{x} = z_1 \mathbf{a} + z_2 \mathbf{b} + z_3 \mathbf{c}$ and $f_Z(z_1, z_2, z_3)$ is probability density function of (z_1, z_2, z_3) . Next, we consider spherical coordinates with $z_1 = r \cos \phi_1, z_2 = r \sin \phi_1 \sin \phi_2, z_3 = r \sin \phi_1 \cos \phi_2$. Hence,

$$I = \int |\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})| \cdot |r \cos \phi_1|^2 \cdot f_Z(r, \phi_1, \phi_2) r^2 \sin \phi_1 dr d\phi_1 d\phi_2. \quad (107)$$

It is easy to verify that $\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x})$ only depends on the direction of \mathbf{x} and

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

only depends on r . Then, we have

$$\begin{aligned} I(i_2, j_2) &= \int |\phi'(\tilde{\mathbf{w}}_{j_2}^T(\mathbf{x}/r)) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T}(\mathbf{x}/r))| \cdot |r \cos \phi_1|^2 \cdot f_Z(r) r^2 \sin \phi_1 dr d\phi_1 d\phi_2 \\ &= \int_0^\infty r^4 f_Z(r) dr \int_0^\pi \int_0^{2\pi} |\cos \phi_1|^2 \cdot \sin \phi_1 \cdot |\phi'(\tilde{\mathbf{w}}_{j_2}^T(\mathbf{x}/r)) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T}(\mathbf{x}/r))| d\phi_1 d\phi_2 \\ &\leq \sqrt{\frac{8}{\pi}} \int_0^\infty r^2 f_Z(r) dr \int_0^\pi \int_0^{2\pi} \sin \phi_1 \cdot |\phi'(\tilde{\mathbf{w}}_{j_2}^T(\mathbf{x}/r)) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T}(\mathbf{x}/r))| d\phi_1 d\phi_2 \\ &= \sqrt{\frac{8}{\pi}} \mathbb{E}_{z_1, z_2, z_3} |\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})| \\ &= \sqrt{\frac{8}{\pi}} \mathbb{E}_{\mathbf{x}} |\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})|. \end{aligned} \quad (108)$$

Define a set $\mathcal{A}_1 = \{\mathbf{x} | (\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}) < 0\}$. If $\mathbf{x} \in \mathcal{A}_1$, then $\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x}$ and $\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}$ have different signs, which means the value of $\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x})$ and $\phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})$ are different. This is equivalent to say that

$$|\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})| = \begin{cases} 1, & \text{if } \mathbf{x} \in \mathcal{A}_1 \\ 0, & \text{if } \mathbf{x} \in \mathcal{A}_1^c \end{cases}. \quad (109)$$

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

$$|\mathbf{w}_{j_2}^{*T} \mathbf{x}| \leq |\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x} - \tilde{\mathbf{w}}_{j_2}^T \mathbf{x}| \leq \|\tilde{\mathbf{w}}_{j_2}^* - \tilde{\mathbf{w}}_{j_2}\| \cdot \|\mathbf{x}\|. \quad (110)$$

Define a set \mathcal{A}_2 such that

$$\mathcal{A}_2 = \left\{ \mathbf{x} \mid \frac{|\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x}|}{\|\tilde{\mathbf{w}}_{j_2}^*\| \|\mathbf{x}\|} \leq \frac{\|\tilde{\mathbf{w}}_{j_2}^* - \tilde{\mathbf{w}}_{j_2}\|}{\|\tilde{\mathbf{w}}_{j_2}^*\|} \right\} = \left\{ \theta_{\mathbf{x}, \tilde{\mathbf{w}}_{j_2}^*} \mid \left| \cos \theta_{\mathbf{x}, \tilde{\mathbf{w}}_{j_2}^*} \right| \leq \frac{\|\tilde{\mathbf{w}}_{j_2}^* - \tilde{\mathbf{w}}_{j_2}\|}{\|\tilde{\mathbf{w}}_{j_2}^*\|} \right\}. \quad (111)$$

Hence, we have that

$$\begin{aligned} \mathbb{E}_{\mathbf{x}} |\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})|^2 &= \mathbb{E}_{\mathbf{x}} |\phi'(\tilde{\mathbf{w}}_{j_2}^T \mathbf{x}) - \phi'(\tilde{\mathbf{w}}_{j_2}^{*T} \mathbf{x})| \\ &= \text{Prob}(\mathbf{x} \in \mathcal{A}_1) \\ &\leq \text{Prob}(\mathbf{x} \in \mathcal{A}_2). \end{aligned} \quad (112)$$

Since $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\theta_{\mathbf{x}, \mathbf{w}_{j_2}^*}$ belongs to the uniform distribution on $[-\pi, \pi]$, we have

$$\begin{aligned} \text{Prob}(\mathbf{x} \in \mathcal{A}_2) &= \frac{\pi - \arccos \frac{\|\tilde{\mathbf{w}}_{j_2}^* - \tilde{\mathbf{w}}_{j_2}\|}{\|\tilde{\mathbf{w}}_{j_2}^*\|}}{\pi} \leq \frac{1}{\pi} \tan(\pi - \arccos \frac{\|\tilde{\mathbf{w}}_{j_2}^* - \tilde{\mathbf{w}}_{j_2}\|}{\|\tilde{\mathbf{w}}_{j_2}^*\|}) \\ &= \frac{1}{\pi} \cot(\arccos \frac{\|\tilde{\mathbf{w}}_{j_2}^* - \tilde{\mathbf{w}}_{j_2}\|}{\|\tilde{\mathbf{w}}_{j_2}^*\|}) \\ &\leq \frac{2}{\pi} \frac{\|\tilde{\mathbf{w}}_{j_2}^* - \tilde{\mathbf{w}}_{j_2}\|}{\|\tilde{\mathbf{w}}_{j_2}^*\|} \\ &= \frac{2}{\pi} \frac{\|\mathbf{w}_{j_2}^* - \mathbf{w}_{j_2}\|}{\|\mathbf{w}_{j_2}^*\|}. \end{aligned} \quad (113)$$

Hence, (108) and (113) suggest that

$$I \leq \frac{6}{\pi} \frac{\|\mathbf{w}_{j_2}^* - \mathbf{w}_{j_2}\|_2}{\|\mathbf{w}_{j_2}^*\|_2}. \quad (114)$$

The same bound that shown in (114) holds for \mathbf{I}_2 as well. \square

J ADDITIONAL PROOFS OF LEMMAS IN APPENDIX F

J.1 ERROR BOUND FOR THE SECOND-ORDER MOMENT

Proof of Lemma 7. Let us define

$$\tilde{\mathbf{x}}_n = \frac{1}{\sqrt{K}} \sum_{j=1}^K \mathbf{x}_{n, \Omega_j}. \quad (115)$$

Then, for $\widehat{M}_2 - M_2$, we have

$$\begin{aligned}
& \widehat{M}_2 - M_2 \\
&= \frac{1}{N} \sum_{n=1}^N y_n (\tilde{\mathbf{x}}_n \otimes \tilde{\mathbf{x}}_n - \mathbb{E} \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n^T) - \mathbb{E}_{\mathbf{x}} y (\tilde{\mathbf{x}} \otimes \tilde{\mathbf{x}} - \mathbb{E} \tilde{\mathbf{x}} \tilde{\mathbf{x}}^T) \\
&= \frac{1}{N} \sum_{n=1}^N \left(\frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j} + \xi_n) (\tilde{\mathbf{x}}_n \otimes \tilde{\mathbf{x}}_n - \mathbb{E} \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n^T) \right. \\
&\quad \left. - \mathbb{E}_{\mathbf{x}} \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*T} \mathbf{x}_{\Omega_j}) (\tilde{\mathbf{x}} \otimes \tilde{\mathbf{x}} - \mathbb{E} \tilde{\mathbf{x}} \tilde{\mathbf{x}}^T) \right) \\
&= \frac{1}{K \cdot N} \sum_{n=1}^N \sum_{j=1}^K \left(\phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) (\tilde{\mathbf{x}}_n \otimes \tilde{\mathbf{x}}_n - \mathbb{E} \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n^T) - \mathbb{E}_{\mathbf{x}} \phi(\mathbf{w}_j^{*T} \mathbf{x}_{\Omega_j}) (\tilde{\mathbf{x}} \otimes \tilde{\mathbf{x}} - \mathbb{E} \tilde{\mathbf{x}} \tilde{\mathbf{x}}^T) \right) \\
&\quad + \frac{1}{N} \sum_{n=1}^N \xi_n (\tilde{\mathbf{x}}_n \otimes \tilde{\mathbf{x}}_n - \mathbb{E} \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n^T)
\end{aligned} \tag{116}$$

Following the notations in Lemma E.2 of [40], we denote

$$\mathbf{B}_2(\mathbf{x}_n) := \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) (\tilde{\mathbf{x}}_n \otimes \tilde{\mathbf{x}}_n - \mathbb{E} \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n^T). \tag{117}$$

Following the similar calculations of (I) - (III) in Lemma E.2 [40], we know that

$$\begin{aligned}
\|\mathbf{B}_2(\mathbf{x})\|_2 &\lesssim \sigma_1 r \log^{\frac{3}{2}} d, \\
\|\mathbb{E}_{\mathbf{x}} \mathbf{B}_2(\mathbf{x})\|_2 &\lesssim \sigma_1, \\
\|\mathbb{E}_{\mathbf{x}} \mathbf{B}_2^2(\mathbf{x})\|_2 &\lesssim \sigma_1^2 r
\end{aligned} \tag{118}$$

hold with probability at least $1 - d^{-10}$.

Define $\mathbf{Z}_{2,n} = \frac{1}{N} (\mathbf{B}_2(\mathbf{x}_n) - \mathbb{E}_{\mathbf{x}} \mathbf{B}_2(\mathbf{x}))$ for \mathbf{x}_n with $n \in [N]$, and it is obvious \mathbf{Z}_n is zero mean. Also, we have

$$R_2 = \|\mathbf{Z}_{2,n}\|_2 \leq \frac{1}{N} (\|\mathbf{B}_2(\mathbf{x}_n)\|_2 + \|\mathbb{E}_{\mathbf{x}} \mathbf{B}_2(\mathbf{x})\|_2) \lesssim N^{-1} \sigma_1 r \log^{\frac{3}{2}} d, \tag{119}$$

and

$$\begin{aligned}
\delta_2^2 &= \left\| \sum_{n=1}^N \mathbb{E} \mathbf{Z}_{2,n}^2 \right\|_2^2 \leq \left\| \sum_{n=1}^N \frac{1}{N^2} \left(\mathbb{E} \mathbf{B}_2^2(\mathbf{x}_n) - (\mathbb{E} \mathbf{B}_2(\mathbf{x}_n))^2 \right) \right\|_2 \\
&\leq \frac{1}{N} \left(\|\mathbb{E} \mathbf{B}_2^2(\mathbf{x}_n)\|_2 + \|\mathbb{E} \mathbf{B}_2(\mathbf{x}_n)\|_2^2 \right) \\
&\lesssim N^{-1} \sigma_1^2 r.
\end{aligned} \tag{120}$$

Next, let $t = \Theta(\sigma_1 \sqrt{\frac{r \log d}{N}})$. To make sure $\delta_2^2 \geq R_2 t / 3$, we need $N \gtrsim r \log^4 d$. Then, by Lemma 10, we have

$$\text{Prob} \left\{ \left\| \sum_n \mathbf{Z}_{2,n} \right\|_2 \geq t \right\} \leq 2r \exp \left(\frac{-t^2/2}{\delta^2 + Rt/3} \right) \leq 2r \exp \left(\frac{-t^2}{4\delta^2} \right). \tag{121}$$

That is

$$\left\| \sum_{n=1}^N \mathbf{z}_{2,n} \right\|_2 \lesssim \sigma_1 \sqrt{\frac{r \log d}{N}} \quad (122)$$

with probability at least $1 - d^{-10}$. Because $\tilde{\mathbf{x}}_n$ belongs to the sub-Gaussian distribution, we know that

$$\left\| \frac{1}{N} \sum_{n=1}^N (\tilde{\mathbf{x}}_n \otimes \tilde{\mathbf{x}}_n - \mathbb{E} \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n^T) \right\|_2 \lesssim \sqrt{\frac{r \log d}{N}} \quad (123)$$

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

In conclusion, we have

$$\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\| \lesssim (\sigma_1 + |\xi|) \sqrt{\frac{r \log d}{N}} \quad (124)$$

with probability at least $1 - d^{-C}$ provided that $N \gtrsim r \log^4 d$. \square

J.2 ERROR BOUND FOR THE THIRD-ORDER MOMENT

Proof of Lemma 8. For $\widehat{\mathbf{M}}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}}) - \mathbf{M}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}})$, we have

$$\begin{aligned} & \widehat{\mathbf{M}}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}}) - \mathbf{M}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}}) \\ &= \frac{1}{N} \sum_{n=1}^N y_n [(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)^{\otimes 3} - (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n) \otimes (\mathbb{E}(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)^T)] \\ & \quad - \mathbb{E}_x y [(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^{\otimes 3} - (\widehat{\mathbf{V}}^T \mathbf{x}^T) \otimes \mathbb{E}(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)^T] \\ &= \frac{1}{N} \sum_{n=1}^N \left(\frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*T} \mathbf{x}_{n,\Omega_j}) + \xi_n \right) \cdot [(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)^{\otimes 3} - (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n) \otimes (\mathbb{E}(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)^T)] \\ & \quad - \mathbb{E}_x \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*T} \mathbf{x}_{\Omega_j}) [(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^{\otimes 3} - (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}) \otimes (\mathbb{E}(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^T)] \\ &= \frac{1}{K \cdot N} \sum_{n=1}^N \sum_{j=1}^K \left[\phi(\mathbf{w}_j^{*T} \mathbf{x}_{n,\Omega_j}) \cdot [(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)^{\otimes 3} - (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n) \otimes (\mathbb{E}(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n)^T)] \right. \\ & \quad \left. - \mathbb{E}_x \phi(\mathbf{w}_j^{*T} \mathbf{x}_{\Omega_j}) [(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^{\otimes 3} - (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}) \otimes (\mathbb{E}(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^T)] \right] \\ & \quad + \frac{1}{N} \sum_{n=1}^N \xi_n [(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^{\otimes 3} - (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}) \otimes (\mathbb{E}(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^T)] \end{aligned} \quad (125)$$

Following the notations in Lemma E.8 of [40], we define

$$\mathbf{T}(\mathbf{x}) := \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*T} \mathbf{x}_{n,\Omega_j}) \cdot [(\widehat{\mathbf{V}}^T \mathbf{x}_n)^{\otimes 3} - (\widehat{\mathbf{V}}^T \mathbf{x}_n) \otimes (\mathbb{E}(\widehat{\mathbf{V}}^T \mathbf{x}_n)(\widehat{\mathbf{V}}^T \mathbf{x}_n)^T)]. \quad (126)$$

Then, $\mathbf{B}_3(\mathbf{x}) \in \mathbb{R}^{K \times K^2}$ is defined as flattening the tensor $\mathbf{T}(\mathbf{x})$ along the first dimension. Hence, we have

$$\begin{aligned} \|\mathbf{B}_3(\mathbf{x})\|_2 &\lesssim \max_j |\mathbf{w}_j^{*T} \mathbf{x}_{\Omega_j}| \cdot \left(\|\widehat{\mathbf{V}}^T \mathbf{x}_n\|_2^3 + 3K \|\widehat{\mathbf{V}}^T \mathbf{x}_n\|_2 \right) \\ &\lesssim \sigma_1 K^{\frac{3}{2}} \log^{\frac{5}{2}} d \end{aligned} \quad (127)$$

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

Following the similar calculations of (II) and (III) in Lemma E.8 of [40], we know that

$$\begin{aligned} & \|\mathbb{E}_{\mathbf{x}} \mathbf{B}_3(\mathbf{x})\|_2 \lesssim \sigma_1, \\ \max \left\{ \|\mathbb{E}_{\mathbf{x}}[\mathbf{B}_3(\mathbf{x})^T \mathbf{B}_3(\mathbf{x})]\|_2, \|\mathbb{E}_{\mathbf{x}}[\mathbf{B}_3(\mathbf{x})^T \mathbf{B}_3(\mathbf{x})]\|_2 \right\} & \lesssim K^2 \sigma_1^2. \end{aligned} \quad (128)$$

Define $\mathbf{Z}_{3,n} = \frac{1}{N}(\mathbf{B}_3(\mathbf{x}_n) - \mathbb{E}_{\mathbf{x}} \mathbf{B}_3(\mathbf{x}))$ for $(\mathbf{x}_n, y_n) \in \mathcal{D}$, and it is obvious $\mathbf{Z}_{3,n}$ is zero mean. Also, we have

$$\begin{aligned} R_3 = \|\mathbf{Z}_{3,n}\|_2 & \leq \frac{1}{N}(\|\mathbf{B}_3(\mathbf{x}_n)\|_2 + \|\mathbb{E}_{\mathbf{x}} \mathbf{B}_3(\mathbf{x})\|_2) \\ & \lesssim N^{-1} \sigma_1 K^{\frac{3}{2}} \log^{\frac{5}{2}} d, \end{aligned} \quad (129)$$

and

$$\begin{aligned} \delta_3^2 & = \left\{ \left\| \sum_{n=1}^N \mathbb{E} \mathbf{Z}_{3,n} \mathbf{Z}_{3,n}^T \right\|_2, \left\| \sum_{n=1}^N \mathbb{E} \mathbf{Z}_{3,n} \mathbf{Z}_{3,n}^T \right\|_2 \right\} \leq \frac{1}{N} (\|\mathbb{E} \mathbf{B}_3^2(\mathbf{x}_n)\|_2 + \|\mathbb{E} \mathbf{B}_3(\mathbf{x}_n)\|_2^2) \\ & \lesssim N^{-1} K^2 \sigma_1^2. \end{aligned} \quad (130)$$

Similar to (121), by applying Lemma 10, we have

$$\left\| \sum_{n=1}^N \mathbf{Z}_{3,n} \right\|_2 \lesssim \sigma_1 \sqrt{\frac{K^2 \log d}{N}} \quad (131)$$

with probability at least $1 - d^{-10}$ provided that $N \gtrsim K^3 \log^6 d$.

Similar to (127), we define \mathbf{B} by flattening the tensor $\sum_{n=1}^N [(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^{\otimes 3} - (\widehat{\mathbf{V}}^T \tilde{\mathbf{x}}) \otimes (\mathbb{E}(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})(\widehat{\mathbf{V}}^T \tilde{\mathbf{x}})^T)]$ along the first dimension. Then, we know that

$$\begin{aligned} \|\mathbf{B}\|_2 & \leq \left\| \sum_{n=1}^N \widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n \right\|_2^3 + 3K \left\| \sum_{n=1}^N \widehat{\mathbf{V}}^T \tilde{\mathbf{x}}_n \right\|_2 \lesssim \left(\frac{K \log d}{N} \right)^{\frac{3}{2}} + 3K \left(\frac{K \log d}{N} \right)^{\frac{1}{2}} \\ & \lesssim \left(\frac{K \log d}{N} \right)^{\frac{1}{2}} + \left(\frac{K^3 \log d}{N} \right)^{\frac{1}{2}} \\ & \lesssim \sqrt{\frac{K^3 \log d}{N}}, \end{aligned} \quad (132)$$

provided that $N \gtrsim K \log d$.

In conclusion, we have

$$\left\| \widehat{\mathbf{M}}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}}) - \mathbf{M}_3(\widehat{\mathbf{V}}, \widehat{\mathbf{V}}, \widehat{\mathbf{V}}) \right\| \lesssim (\sigma_1 + |\xi|) \sqrt{\frac{K^3 \log d}{N}} \quad (133)$$

with probability at least $1 - d^{-C}$ provided that $N \gtrsim K^3 \log^6 d$. □

J.3 ERROR BOUND FOR THE FIRST-ORDER MOMENT

Proof of Lemma 9. For $\widehat{\mathbf{M}}_1 - \mathbf{M}_1$, we have

$$\begin{aligned}
\widehat{\mathbf{M}}_1 - \mathbf{M}_1 &= \frac{1}{N} \sum_{n=1}^N y_n \tilde{\mathbf{x}}_n - \mathbb{E}_{\mathbf{x}} y \tilde{\mathbf{x}} \\
&= \frac{1}{N} \sum_{n=1}^N \left(\frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) + \xi_n \right) \tilde{\mathbf{x}}_n - \mathbb{E}_{\mathbf{x}} \sum_{j=1}^K \frac{1}{K} \phi(\mathbf{w}_j^{*T} \mathbf{x}_{\Omega_j}) \tilde{\mathbf{x}} \\
&= \frac{1}{K \cdot N} \sum_{j=1}^K \sum_{n=1}^N \left(\phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) \tilde{\mathbf{x}}_n - \mathbb{E}_{\mathbf{x}} \phi(\mathbf{w}_j^{*T} \mathbf{x}_{\Omega_j}) \tilde{\mathbf{x}} \right) + \frac{1}{N} \sum_{n=1}^N \xi_n \cdot \tilde{\mathbf{x}}_n.
\end{aligned} \tag{134}$$

Define $\mathbf{B}_1(\mathbf{x}) := \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*T} \mathbf{x}_{n, \Omega_j}) \tilde{\mathbf{x}}_n$, then we have

$$\begin{aligned}
\|\mathbf{B}_1(\mathbf{x})\|_2 &\lesssim \sigma_1 r \log^{\frac{3}{2}} d; \\
\|\mathbb{E}_{\mathbf{x}} \mathbf{B}_1(\mathbf{x})\|_2 &\lesssim \sigma_1; \\
\left\{ \|\mathbb{E}_{\mathbf{x}} [\mathbf{B}_1(\mathbf{x}) \mathbf{B}_1(\mathbf{x})^T]\|_2, \|\mathbb{E}_{\mathbf{x}} [\mathbf{B}_{1,j}(\mathbf{x})^T \mathbf{B}_1(\mathbf{x})]\|_2 \right\} &\lesssim \sigma_1^2.
\end{aligned} \tag{135}$$

Next, define $\mathbf{Z}_{1,n} = \frac{1}{N} (\mathbf{B}_{1,j}(\mathbf{x}_n) - \mathbb{E}_{\mathbf{x}} \mathbf{B}_2(\mathbf{x}))$ for $(\mathbf{x}_n, y_n) \in \mathcal{D}$, by calculation, we can obtain

$$R_1 = \|\mathbf{Z}_{1,n}\|_2 \lesssim N^{-1} \sigma_1 r \log^{\frac{3}{2}} d, \tag{136}$$

and

$$\delta_1^2 = \max \left\{ \left\| \sum_{n=1}^N \mathbb{E} \mathbf{Z}_{1,n} \mathbf{Z}_{1,n}^T \right\|_2^2, \left| \sum_{n=1}^N \mathbf{Z}_{1,n}^T \mathbf{Z}_{1,n} \right| \right\} \lesssim N^{-1} \sigma_1^2 r. \tag{137}$$

By applying Lemma 10, we have

$$\left\| \sum_{n=1}^N \mathbf{Z}_{1,n} \right\|_2 \lesssim \sigma_1 \sqrt{\frac{r \log d}{N}} \tag{138}$$

with probability at least $1 - d^{-10}$ provided that $N \gtrsim r \log^4 d$. Since $\mathbf{x} \in \mathbb{R}^r$ belongs to the Gaussian distribution, we have

$$\left\| \frac{1}{N} \sum_{n=1}^N \tilde{\mathbf{x}} \right\|_2 \lesssim \sqrt{\frac{r \log d}{N}} \tag{139}$$

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

In conclusion, we have

$$\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\| \lesssim (\sigma_1 + |\xi|) \sqrt{\frac{r \log d}{N}} \tag{140}$$

with probability at least $1 - d^{-C}$, provided that $N \gtrsim r \log^4 d$. \square