Linear Convergence Analysis of Neural Collapse under the MSE Loss

July 5, 2024

1 Preliminaries and Main Results

Without loss of generality, let $\boldsymbol{H}_i := [\boldsymbol{h}_{1,i},\ \dots,\ \boldsymbol{h}_{K,i}] \in \mathbb{R}^{d\times K}$ for $i \in [n]$ and $\boldsymbol{H} := [\boldsymbol{H}_1,\ \cdots,\ \boldsymbol{H}_n] \in \mathbb{R}^{d\times N}$ be the matrix with the columns being organized unconstrained features, associated with the label matrix $\boldsymbol{Y} = \boldsymbol{1}_n^T \otimes \boldsymbol{I}_K \in \mathbb{R}^{K\times N}$. Let

$$F(\boldsymbol{W},\boldsymbol{H}) := \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{W}^T\boldsymbol{h}_{k,i}+\boldsymbol{b},\boldsymbol{y}_k) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2 + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2. \tag{1}$$

where $\boldsymbol{y}_k \in \mathbb{R}^K$ is a membership matrix with all the entries being 0 but the k-th one being 1, $\lambda > 0$ is the regularized parameter, and $\mathcal{L}: \mathbb{R}^d \times \mathbb{R}^K \to \mathbb{R}_+$ is a loss function. Let

$$\min_{\boldsymbol{W},\boldsymbol{H}\in\mathbb{R}^{d\times K}} f(\boldsymbol{W},\boldsymbol{H}) := \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}(\boldsymbol{W}^T\boldsymbol{h}_k+\boldsymbol{b},\boldsymbol{y}_k) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2 + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2. \tag{2}$$

We define $\boldsymbol{Z} := \boldsymbol{W}^T\boldsymbol{H} + \boldsymbol{b}\boldsymbol{1}_K^T$. This implies $\boldsymbol{z}_k = \boldsymbol{W}^T\boldsymbol{h}_k$ for all $k \in [K]$. To simplify our development, let

$$g_k(\boldsymbol{z}) = \frac{1}{K}\mathcal{L}(\boldsymbol{z},\boldsymbol{y}_k), \quad g(\boldsymbol{Z}) = \sum_{k=1}^{K} g_k(\boldsymbol{z}_k). \tag{3}$$

Consider a optimization problem

$$v^* = \min_{\boldsymbol{x}\in\mathcal{E}} F(\boldsymbol{x}), \tag{4}$$

where $\mathcal{E}$ is a finite-dimensional Euclidean space and $f: \mathcal{E} \to (\infty,\infty)$ is a continuously differentiable function. Let $\mathcal{X} \subseteq \mathcal{E}$ denote the set of optimal solutions of Problem (4).

Definition 1 (Error bound condition)**.** *We say that an error bound condition holds for Problem* (4) *if there exists a constant* $\kappa > 0$ *such that for all* $\boldsymbol{x} \in \mathbb{R}^n$ *with* $\operatorname{dist}(\boldsymbol{x},\mathcal{X}) \le \delta$*,*

$$\operatorname{dist}(\boldsymbol{x},\mathcal{X}) \le \kappa\|\nabla F(\boldsymbol{x})\|. \tag{5}$$

It follows from [1, 2, 3] that the error bound condition can be used to analyze the convergence rate of first-order methods.

Fact 1 (cf. [1, 2, 3])**.** *Suppose that the optimal solution of Problem (4) is non-empty, i.e., $\mathcal{X} \neq \emptyset$, and the error bound holds for Problem (4). Suppose in addition that the sequence $\{\boldsymbol{x}^k\}_{k \geq k_1}$ for an index $k_1 \geq 0$ satisfies the following properties:*
(A1). (Sufficient Decrease) *There exists a constant $\kappa_1 > 0$ such that*

$$F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^k) \leq -\kappa_1 \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2.$$

(A2). (Cost-to-Go Estimate) *There exists a constant $\kappa_2 > 0$ such that*

$$F(\boldsymbol{x}^{k+1}) - v^* \leq \kappa_2 \left(d^2(\boldsymbol{x}^k, \mathcal{X}) + \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 \right).$$

(A3). (Safeguard) *There exists a constant $\kappa_3 > 0$ such that*

$$\|\nabla F(\boldsymbol{x}^k)\| \leq \kappa_3 \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|.$$

Then, the sequence $\{F(\boldsymbol{x}^k)\}_{k \geq 0}$ converges Q-linearly to v^* and $\{\boldsymbol{x}^k\}_{k \geq 0}$ converges R-linearly to some $\boldsymbol{x}^* \in \mathcal{X}$.

Despite the fact that F in Problem (4) is non-convex, we can verify that the sequence generated by the gradient descent method for solving Problem (4) satisfies (A1)-(A3) in Fact (1).

Notation. For a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times K}$, we denote by $\boldsymbol{a}_k$ by its k-th column and by $\boldsymbol{a}^i$ by its i-th row. Let $\boldsymbol{P} = \boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T$ be a projection matrix and $\boldsymbol{P}^\perp = \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T$ be its complement. Let $\lambda_{\min} = \min\{\lambda_W, \lambda_H\}$ and $\lambda_{\max} = \max\{\lambda_W, \lambda_H\}$.

2 Neural Collapse with Regularized MSE Loss

Suppose that $\mathcal{L}(\cdot, \cdot)$ is the meas squared error (MSE) loss:

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{y}_k) = \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{y}_k\|^2. \tag{6}$$

Substituting this into Problem (1) and Problem (2) respectively yields

$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times K}, \boldsymbol{H} \in \mathbb{R}^{d \times N}} F(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{2N}\|\boldsymbol{W}^T\boldsymbol{H} - \boldsymbol{Y}\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2, \tag{7}$$

$$\min_{\boldsymbol{W}, \boldsymbol{\Theta} \in \mathbb{R}^{d \times K}} f(\boldsymbol{W}, \boldsymbol{\Theta}) = \frac{1}{2K}\|\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{I}_K\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}\|_F^2. \tag{8}$$

Note that

$$F(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{W}, \boldsymbol{H}_i). \tag{9}$$

We first study the following optimization problems:

$$\min_{\boldsymbol{W}, \boldsymbol{\Theta} \in \mathbb{R}^{d \times K}} f(\boldsymbol{W}, \boldsymbol{\Theta}) = \frac{1}{2K}\|\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{\Sigma}\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}\|_F^2, \tag{10}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_K)$ is a diagonal matrix with $\sigma_1 \geq \cdots \geq \sigma_K \geq 0$. It is worth noting that Problem (8) is a special case of this problem by taking $\sigma_k = 1$ for all $k \in [K]$.

Lemma 1. *The optimal solution set of Problem* (10) *takes the form of*

$$\left\{(\boldsymbol{W},\boldsymbol{\Theta}) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}\boldsymbol{U}\left(\max\left\{\boldsymbol{\Sigma} - K\sqrt{\lambda_W\lambda_H}\boldsymbol{I}_K, \boldsymbol{0}\right\}\right)^{\frac{1}{2}}, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}, \boldsymbol{U} \in \mathcal{O}^{d\times K}\right\}. \tag{11}$$

Proof. According to the first-order optimality condition, we have

$$\begin{cases} \nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta}) = \boldsymbol{\Theta}\left(\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{\Sigma}\right)^T + \lambda_W K\boldsymbol{W} = \boldsymbol{0}, \\ \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta}) = \boldsymbol{W}\left(\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{\Sigma}\right) + n\lambda_H K\boldsymbol{\Theta} = \boldsymbol{0}. \end{cases} \tag{12}$$

Using $\boldsymbol{W}\nabla_{\boldsymbol{W}}^T f(\boldsymbol{W},\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{\Theta}^T = \boldsymbol{0}$ yields

$$\boldsymbol{\Theta}\boldsymbol{\Theta}^T = \frac{\lambda_W}{n\lambda_H}\boldsymbol{W}\boldsymbol{W}^T. \tag{13}$$

This, together with (12), implies

$$\begin{cases} \dfrac{\lambda_W}{n\lambda_H}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{W} - \boldsymbol{\Theta}\boldsymbol{\Sigma} + \lambda_W K\boldsymbol{W} = \boldsymbol{0}, & \text{(14a)} \\ \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{W}\boldsymbol{\Sigma} + n\lambda_H K\boldsymbol{\Theta} = \boldsymbol{0}. & \text{(14b)} \end{cases}$$

According to $\sqrt{n\lambda_H} \times$ (14a) $- \sqrt{\lambda_W} \times$ (14b) $= \boldsymbol{0}$, we have

$$\left(\frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}\boldsymbol{W}^T + K\sqrt{n\lambda_W\lambda_H}\boldsymbol{I}_d\right)\left(\sqrt{\lambda_W}\boldsymbol{W} - \sqrt{n\lambda_H}\boldsymbol{\Theta}\right) + \left(\sqrt{\lambda_W}\boldsymbol{W} - \sqrt{n\lambda_H}\boldsymbol{\Theta}\right)\boldsymbol{\Sigma} = \boldsymbol{0}.$$

Since $\sqrt{\lambda_W}/\sqrt{n\lambda_H}\boldsymbol{W}\boldsymbol{W}^T + K\sqrt{n\lambda_W\lambda_H}\boldsymbol{I}_d$ is positive definite and $\boldsymbol{\Theta}$ is diagonal with non-negative diagonal entries, we have $\sqrt{\lambda_W}\boldsymbol{W} = \sqrt{n\lambda_H}\boldsymbol{\Theta}$. For all $(\boldsymbol{W},\boldsymbol{\Theta})$ satisfying the first-order optimality condition, we compute

$$\begin{aligned} f(\boldsymbol{W},\boldsymbol{\Theta}) &= \frac{1}{2K}\|\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{\Theta}\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}\|_F^2 \\ &= \frac{1}{2K}\left\|\frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}^T\boldsymbol{W} - \boldsymbol{\Theta}\right\|_F^2 + \lambda_W\|\boldsymbol{W}\|_F^2 \\ &= \frac{1}{2K}\left\|\frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}^T\boldsymbol{W} - \left(\boldsymbol{\Theta} - K\sqrt{n\lambda_W\lambda_H}\boldsymbol{I}_K\right)\right\|_F^2 + \sqrt{n\lambda_W\lambda_H}\operatorname{tr}(\boldsymbol{\Sigma}) - \frac{nK^2}{2}\lambda_W\lambda_H. \end{aligned}$$

To obtain the global minimum of $f(\boldsymbol{W},\boldsymbol{\Theta})$, we have

$$\boldsymbol{W}^T\boldsymbol{W} = \frac{\sqrt{n\lambda_H}}{\sqrt{\lambda_W}}\max\left\{\boldsymbol{\Sigma} - K\sqrt{n\lambda_W\lambda_H}\boldsymbol{I}_K, \boldsymbol{0}\right\},$$

which implies

$$\boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}\boldsymbol{U}\left(\max\left\{\boldsymbol{\Sigma} - K\sqrt{n\lambda_W\lambda_H}\boldsymbol{I}_K, \boldsymbol{0}\right\}\right)^{1/2}, \text{ for all } \boldsymbol{U} \in \mathcal{O}^{d\times K}.$$

This, together with $\sqrt{\lambda_W}\boldsymbol{W} = \sqrt{n\lambda_H}\boldsymbol{\Theta}$, gives (11). □

Proposition 1. *The optimal solution set of Problem* (8) *takes the form of*

$$\mathcal{X}_f = \left\{(\boldsymbol{W},\boldsymbol{\Theta}) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}\left(\max\{c, 0\}\right)^{\frac{1}{2}}\boldsymbol{U}, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}, \boldsymbol{U} \in \mathcal{O}^{d\times K}\right\}, \tag{15}$$

where $c := 1 - K\sqrt{n\lambda_W\lambda_H}$.

Corollary 1. *The optimal solution set of Problem* (7) *takes the form of*

$$\mathcal{X}_F = \left\{(\boldsymbol{W}, \boldsymbol{H}) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}} \left(\max\{c, 0\}\right)^{\frac{1}{2}} \boldsymbol{U}, \boldsymbol{H}_i = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}} \boldsymbol{W}, \forall i \in [n], \boldsymbol{U} \in \mathcal{O}^{d\times K}\right\}, \tag{16}$$

where $c := 1 - K\sqrt{n\lambda_W\lambda_H}$.

Proof. Suppose that $(\boldsymbol{W}^*, \boldsymbol{\Theta}^*) \in \mathcal{X}_f$ is an optimal solution of Problem (8). According to (9), we have

$$\min F(\boldsymbol{W}, \boldsymbol{H}) \geq \frac{1}{n}\sum_{i=1}^{n} \min f(\boldsymbol{W}, \boldsymbol{\Theta}) = f(\boldsymbol{W}^*, \boldsymbol{\Theta}^*),$$

where the equality holds if $\boldsymbol{W} = \boldsymbol{W}^*$ and $\boldsymbol{H}_i = \boldsymbol{\Theta}^*$ for all $i \in [N]$. This, together with (15), implies (16). □

Note that when $\lambda_W\lambda_H \geq 1/(nK^2)$, we have $c \leq 0$. Then the optimal solution set of Problem (7) is $\{(\boldsymbol{0}, \boldsymbol{0})\}$. To avoid this trivial case, it suffices to consider $\lambda_W\lambda_H < 1/(nK^2)$.

Lemma 2. *Suppose that* $\lambda_W\lambda_H < 1/(nK^2)$ *and the error bound holds for Problem* (2)*, i.e., there exist constants* $\delta, \kappa > 0$ *such that*

$$\operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{\Theta}), \mathcal{X}_f\right) \leq \kappa \|\nabla f(\boldsymbol{W}, \boldsymbol{\Theta})\|_F \tag{17}$$

for all $(\boldsymbol{W}, \boldsymbol{\Theta}) \in \mathbb{R}^{d\times K}$ *satisfying* $\operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{\Theta}), \mathcal{X}_f\right) \leq \delta$. *Then for all* $(\boldsymbol{W}, \boldsymbol{H})$ *satisfying*

$$\operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{H}), \mathcal{X}_F\right) \leq \delta_F := \min\left\{\frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}, \frac{\sqrt[4]{\lambda_W}}{\sqrt[4]{n\lambda_H}}, \frac{\sqrt{2}}{8} nK^2\lambda_H\sqrt{n\lambda_W\lambda_H}\delta, \frac{\sqrt{2}}{2}\delta\right\}, \tag{18}$$

it holds that

$$\operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{H}), \mathcal{X}_F\right) \leq \kappa_F \|\nabla F(\boldsymbol{W}, \boldsymbol{H})\|_F, \tag{19}$$

where

$$\kappa_F := \frac{\sqrt{2n}\kappa \max\left\{1, \frac{n\lambda_H}{\kappa} + \frac{9}{nK\lambda_H\sqrt{\lambda_W\lambda_H}}\right\}}{\min\left\{\sqrt{\frac{n}{2}}\left(1 + \frac{2\lambda_W}{n\lambda_H}\right)^{-1/2}, \frac{\sqrt{2}}{4}\min\left\{1, \frac{\sqrt{n\lambda_H}}{\sqrt{\lambda_W}}\right\}\right\}}. \tag{20}$$

Proof. For ease of exposition, let $c := 1 - K\sqrt{n\lambda_W\lambda_H}$ and $\rho := \sqrt[4]{n\lambda_H}/\sqrt[4]{\lambda_W}$. According to (16) in Corollary 1 and $\lambda_W\lambda_H < 1/(nK^2)$, we compute

$$\begin{aligned}
\operatorname{dist}^2\left((\boldsymbol{W}, \boldsymbol{H}), \mathcal{X}_F\right) &= \min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}} \left\{\left\|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\right\|_F^2 + \sum_{i=1}^{n}\left\|\boldsymbol{H}_i - \frac{\sqrt{c}}{\rho}\boldsymbol{U}\right\|_F^2\right\} \\
&\leq \left(1 + \frac{2\lambda_W}{n\lambda_H}\right)\min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}} \left\|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\right\|_F^2 + 2\sum_{i=1}^{n}\left\|\boldsymbol{H}_i - \frac{1}{\rho^2}\boldsymbol{W}\right\|_F^2,
\end{aligned} \tag{21}$$

where the inequality follows from $\|\boldsymbol{A}+\boldsymbol{B}\|_F^2 \le 2\|\boldsymbol{A}\|_F^2 + 2\|\boldsymbol{B}\|_F^2$. Moreover, given some $(\boldsymbol{W}, \boldsymbol{\Theta})$, it follows from (15) in Proposition 1 that

$$\begin{aligned}
\operatorname{dist}^2\left((\boldsymbol{W},\boldsymbol{\Theta}),\mathcal{X}_f\right) &= \min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}} \left\{ \|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\|_F^2 + \left\|\boldsymbol{\Theta} - \frac{\sqrt{c}}{\rho}\boldsymbol{U}\right\|_F^2 \right\} \\
&\ge \min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}} \left\{ \|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\|_F^2 + \alpha\left\|\boldsymbol{\Theta} - \frac{\sqrt{c}}{\rho}\boldsymbol{U}\right\|_F^2 \right\} \\
&\ge \frac{1}{2}\min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}} \|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\|_F^2 + \frac{\alpha}{2}\left\|\boldsymbol{\Theta} - \frac{1}{\rho^2}\boldsymbol{W}\right\|_F^2, \qquad (22)
\end{aligned}$$

where the first inequality is due to $\alpha := \min\{1, n\lambda_H/\lambda_W\}/2 \le 1$, and the second inequality follows from $\|\boldsymbol{A}+\boldsymbol{B}\|_F^2 \ge \|\boldsymbol{A}\|_F^2/2 - \|\boldsymbol{B}\|_F^2$ and $\alpha\lambda_W/(n\lambda_H) \le 1/2$. This, together with (21), yields

$$\begin{aligned}
&\min\left\{\frac{n}{2}\left(1+\frac{2\lambda_W}{n\lambda_H}\right)^{-1}, \frac{\alpha}{4}\right\}\operatorname{dist}^2\left((\boldsymbol{W},\boldsymbol{H}),\mathcal{X}_F\right) \le \frac{n}{2}\min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}} \|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\|_F^2 \\
&+ \frac{\alpha}{2}\sum_{i=1}^n \left\|\boldsymbol{H}_i - \frac{1}{\rho^2}\boldsymbol{W}\right\|_F^2 \le \sum_{i=1}^n \operatorname{dist}^2\left((\boldsymbol{W},\boldsymbol{H}_i),\mathcal{X}_f\right). \qquad (23)
\end{aligned}$$

Noting that $f(\boldsymbol{W},\boldsymbol{\Theta})$ is strongly convex w.r.t. $\boldsymbol{\Theta}$ with constant $n\lambda_H$, let $\theta(\boldsymbol{W}) := \arg\min_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})$ denote the unique minimizer. Using the first-order optimality condition, we obtain

$$\theta(\boldsymbol{W}) = \left(\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H\boldsymbol{I}\right)^{-1}\boldsymbol{W}. \qquad (24)$$

Using the strongly convexity again, we have for arbitrary $\boldsymbol{\Theta} \in \mathbb{R}^{d\times K}$,

$$n\lambda_H\|\boldsymbol{\Theta} - \theta(\boldsymbol{W})\|_F^2 \le \langle \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta}) - \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\theta(\boldsymbol{W})), \boldsymbol{\Theta} - \theta(\boldsymbol{W})\rangle.$$

This, together with the Cauchy–Schwarz inequality and $\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\theta(\boldsymbol{W})) = \boldsymbol{0}$, implies $\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F \ge n\lambda_H\|\boldsymbol{\Theta} - \theta(\boldsymbol{W})\|_F$. Then, it holds for arbitrary $\boldsymbol{H}$ that for all $i \in [n]$,

$$\|\nabla_{\boldsymbol{H}_i} f(\boldsymbol{W},\boldsymbol{H}_i)\|_F \ge n\lambda_H\|\boldsymbol{H}_i - \theta(\boldsymbol{W})\|_F. \qquad (25)$$

Let $(\boldsymbol{W}^*,\boldsymbol{H}^*) \in \mathcal{X}_F$ be such that $\operatorname{dist}\left((\boldsymbol{W},\boldsymbol{H}),\mathcal{X}_F\right) = \|(\boldsymbol{W},\boldsymbol{H}) - (\boldsymbol{W}^*,\boldsymbol{H}^*)\|_F \le \delta_F$. It follows from $\operatorname{dist}\left((\boldsymbol{W},\boldsymbol{H}),\mathcal{X}_F\right) \le \delta_F$ that

$$\|\boldsymbol{W} - \boldsymbol{W}^*\|_F^2 + \sum_{i=1}^n \|\boldsymbol{H}_i - \boldsymbol{H}_i^*\|_F^2 \le \delta_F^2. \qquad (26)$$

This, together with (16) and (24), gives

$$\|\boldsymbol{W}\| \le \|\boldsymbol{W} - \boldsymbol{W}^*\|_F + \|\boldsymbol{W}^*\| \le \delta_F + \rho\sqrt{c},\ \|\boldsymbol{H}_i\| \le \delta_F + \frac{\sqrt{c}}{\rho},\ \forall i \in [n], \qquad (27)$$

$$\|\theta(\boldsymbol{W})\| \le \left\|\left(\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H\boldsymbol{I}\right)^{-1}\right\|\|\boldsymbol{W}\| \le \frac{1}{nK\lambda_H}\left(\delta_F + \rho\sqrt{c}\right). \qquad (28)$$

Using (12), we compute

$$\begin{aligned}
\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{H}_i) - \nabla_{\boldsymbol{W}} f(\boldsymbol{W},\theta(\boldsymbol{W}))\|_F &= \left\|\boldsymbol{H}_i\boldsymbol{H}_i^T\boldsymbol{W} - \theta(\boldsymbol{W})\theta(\boldsymbol{W})^T\boldsymbol{W} + \theta(\boldsymbol{W}) - \boldsymbol{H}_i\right\|_F \\
&\le \left(\left(\|\boldsymbol{H}_i\| + \|\theta(\boldsymbol{W})\|\right)\|\boldsymbol{W}\| + 1\right)\|\boldsymbol{H}_i - \theta(\boldsymbol{W})\|_F \\
&\le \frac{9}{K\sqrt{n\lambda_W\lambda_H}}\|\boldsymbol{H}_i - \theta(\boldsymbol{W})\|_F, \qquad (29)
\end{aligned}$$

where the last inequality follows from (18), (27), and (28). Next, we bound

$$\begin{aligned}
\left\|\sum_{i=1}^{n} \nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}_i)\right\|_F &= \left\|\sum_{i=1}^{n} \nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \theta(\boldsymbol{W})) + \nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}_i) - \nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \theta(\boldsymbol{W}))\right\|_F \\
&\geq n \left\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \theta(\boldsymbol{W}))\right\|_F - \sum_{i=1}^{n} \left\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}_i) - \nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \theta(\boldsymbol{W}))\right\|_F \\
&\geq n \left\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \theta(\boldsymbol{W}))\right\|_F - \frac{9}{K\sqrt{n\lambda_W \lambda_H}} \sum_{i=1}^{n} \left\|\boldsymbol{H}_i - \theta(\boldsymbol{W})\right\|_F,
\end{aligned}$$

where the last inequality follows from (29). Substituting (25) into this inequality yields

$$\left\|\sum_{i=1}^{n} \nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}_i)\right\|_F + \frac{9}{nK\lambda_H\sqrt{n\lambda_W\lambda_H}} \sum_{i=1}^{n} \left\|\nabla_{\boldsymbol{H}_i} f(\boldsymbol{W}, \boldsymbol{H}_i)\right\|_F \geq n \left\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \theta(\boldsymbol{W}))\right\|_F. \tag{30}$$

According to (24), we compute

$$\begin{aligned}
\left\|\theta(\boldsymbol{W}) - \theta(\boldsymbol{W}^*)\right\|_F &= \Big\| \left((\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H \boldsymbol{I})^{-1} - (\boldsymbol{W}^*\boldsymbol{W}^{*^T} + nK\lambda_H \boldsymbol{I})^{-1} \right) \boldsymbol{W}^* \\
&\quad + \left(\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H \boldsymbol{I}\right)^{-1} (\boldsymbol{W} - \boldsymbol{W}^*) \Big\|_F \\
&\leq \left(\frac{1}{(nK\lambda_H)^2} \left(\|\boldsymbol{W}^*\| + \|\boldsymbol{W}\|\right) \|\boldsymbol{W}^*\| + \frac{1}{nK\lambda_H} \right) \|\boldsymbol{W}^* - \boldsymbol{W}\|_F \\
&\leq \left(\frac{1}{(nK\lambda_H)^2} \left(\delta_F + 2\rho\sqrt{c}\right) \rho\sqrt{c} + \frac{1}{nK\lambda_H} \right) \delta_F \leq \frac{\sqrt{2}}{2}\delta,
\end{aligned} \tag{31}$$

where the first inequality follows from $(\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H\boldsymbol{I})^{-1} - (\boldsymbol{W}^*\boldsymbol{W}^{*^T} + nK\lambda_H\boldsymbol{I})^{-1} = (\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H\boldsymbol{I})^{-1}(\boldsymbol{W}^*\boldsymbol{W}^{*^T} + nK\lambda_H\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T - nK\lambda_H\boldsymbol{I})(\boldsymbol{W}^*\boldsymbol{W}^{*^T} + nK\lambda_H\boldsymbol{I})^{-1} = (\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H\boldsymbol{I})^{-1} \left(\boldsymbol{W}^*(\boldsymbol{W}^* - \boldsymbol{W})^T - (\boldsymbol{W}^* - \boldsymbol{W})\boldsymbol{W}^T\right)(\boldsymbol{W}^*\boldsymbol{W}^{*^T} + nK\lambda_H\boldsymbol{I})^{-1}$, $\|(\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H\boldsymbol{I})^{-1}\| \leq 1/(nK\lambda_H)$, and $\|(\boldsymbol{W}^*\boldsymbol{W}^{*T} + nK\lambda_H\boldsymbol{I})^{-1}\| \leq 1/(nK\lambda_H)$, the second inequality uses (26) and (27), and the last inequality is due to (18) and $c \leq 1$. This, together with $\theta(\boldsymbol{W}^*) = \boldsymbol{H}^*$ implies $\|\theta(\boldsymbol{W}) - \boldsymbol{H}^*\|_F = \|\theta(\boldsymbol{W}) - \theta(\boldsymbol{W}^*)\|_F \leq \sqrt{2}\delta/2$. Using this and (18), we have

$$\operatorname{dist}^2\left((\boldsymbol{W}, \theta(\boldsymbol{W})), \mathcal{X}_f\right) \leq \|\boldsymbol{W} - \boldsymbol{W}^*\|_F^2 + \|\theta(\boldsymbol{W}) - \boldsymbol{H}^*\|_F^2 \leq \delta^2. \tag{32}$$

Using $\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W}, \theta(\boldsymbol{W})) = \boldsymbol{0}$, we have

$$\begin{aligned}
\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \theta(\boldsymbol{W}))\|_F = \|\nabla f(\boldsymbol{W}, \theta(\boldsymbol{W}))\|_F &\geq \kappa^{-1} \operatorname{dist}\left((\boldsymbol{W}, \theta(\boldsymbol{W})), \mathcal{X}_f\right) \\
&\geq \kappa^{-1} \operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{H}_i), \mathcal{X}_f\right) - \kappa^{-1}\|\boldsymbol{H}_i - \theta(\boldsymbol{W})\|_F, \ \forall i \in [n],
\end{aligned}$$

where the first inequality follows from (17) and (32). Substituting (25) into this inequality yields

$$\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \theta(\boldsymbol{W}))\|_F + \frac{n\lambda_H}{\kappa} \|\nabla_{\boldsymbol{H}_i} f(\boldsymbol{W}, \boldsymbol{H}_i)\|_F \geq \frac{1}{\kappa} \operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{H}_i), \mathcal{X}_f\right).$$

Summing up this inequality from $i = 1$ to $i = n$ and using (30) gives

$$\left\|\sum_{i=1}^{n} \nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}_i)\right\|_F + \left(\frac{n\lambda_H}{\kappa} + \frac{9}{nK\lambda_H\sqrt{n\lambda_W\lambda_H}}\right) \sum_{i=1}^{n} \|\nabla_{\boldsymbol{H}_i} f(\boldsymbol{W}, \boldsymbol{H}_i)\|_F \geq \frac{1}{\kappa} \sum_{i=1}^{n} \operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{H}_i), \mathcal{X}_f\right).$$

This, together with (9), implies

$$\left(\sum_{i=1}^{n} \operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{H}_i), \mathcal{X}_f\right)\right)^2 \le 2n\kappa^2 \max\left\{1, \left(\frac{n\lambda_H}{\kappa} + \frac{9}{nK\lambda_H\sqrt{\lambda_W\lambda_H}}\right)^2\right\} \|\nabla F(\boldsymbol{W}, \boldsymbol{H})\|_F^2$$

Combining this with (23) yields

$$\operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{H}), \mathcal{X}_F\right) \le \frac{\sqrt{2n}\kappa \max\left\{1, \frac{n\lambda_H}{\kappa} + \frac{9}{nK\lambda_H\sqrt{\lambda_W\lambda_H}}\right\}}{\min\left\{\sqrt{\frac{n}{2}}\left(1+\frac{2\lambda_W}{n\lambda_H}\right)^{-1/2}, \frac{\sqrt{2}}{4}\min\left\{1, \frac{\sqrt{n\lambda_H}}{\sqrt{\lambda_W}}\right\}\right\}} \|\nabla F(\boldsymbol{W}, \boldsymbol{H})\|_F.$$

□

According to this lemma, it suffices to consider the error bound of Problem (8).

Theorem 1. *Suppose that $\lambda_W\lambda_H < 1/(nK^2)$. For all $(\boldsymbol{W}, \boldsymbol{\Theta}) \in \mathbb{R}^{d\times K} \times \mathbb{R}^{d\times K}$ satisfying*

$$\operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{\Theta}), \mathcal{X}_f\right) \le \delta := \frac{1}{2}\min\left\{\frac{\sqrt[4]{\lambda_W}}{\sqrt[4]{n\lambda_H}}, \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}\right\}\left(1 - K\sqrt{n\lambda_W\lambda_H}\right)^{1/2}, \tag{33}$$

it holds that

$$\operatorname{dist}\left((\boldsymbol{W}, \boldsymbol{\Theta}), \mathcal{X}_f\right) \le \kappa \|\nabla f(\boldsymbol{W}, \boldsymbol{H})\|_F, \tag{34}$$

where κ is a constant that depends on λ_W and λ_H.

Proof. For ease of exposition, let

$$c := 1 - K\sqrt{n\lambda_W\lambda_H},\ \rho := \sqrt[4]{n\lambda_H}/\sqrt[4]{\lambda_W},\ \lambda_{\min} := \min\{n\lambda_H, \lambda_W\},\ \lambda_{\max} := \max\{n\lambda_H, \lambda_W\}.$$

By Proposition 1 and the condition $\lambda_W\lambda_H < 1/(nK^2)$, we obtain

$$\mathcal{X}_f = \left\{(\boldsymbol{W}, \boldsymbol{\Theta}) : \boldsymbol{W} = \rho\sqrt{c}\boldsymbol{U}, \boldsymbol{\Theta} = \frac{\sqrt{c}}{\rho}\boldsymbol{U}, \boldsymbol{U} \in \mathcal{O}^{d\times K}\right\}, \tag{35}$$

Next, we calculate

$$\begin{aligned}\operatorname{dist}^2\left((\boldsymbol{W}, \boldsymbol{\Theta}), \mathcal{X}_f\right) &= \min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}}\left\{\|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\|_F^2 + \left\|\boldsymbol{\Theta} - \frac{\sqrt{c}}{\rho}\boldsymbol{U}\right\|_F^2\right\} \\ &\le 2\left\|\boldsymbol{\Theta} - \frac{1}{\rho^2}\boldsymbol{W}\right\|_F^2 + \left(1 + \frac{2\lambda_W}{n\lambda_H}\right)\min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}}\|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\|_F^2, \end{aligned} \tag{36}$$

where the inequality follows from $\|\boldsymbol{A}+\boldsymbol{B}\|_F^2 \le 2\|\boldsymbol{A}\|_F^2 + 2\|\boldsymbol{B}\|_F^2$ for any $\boldsymbol{A}, \boldsymbol{B}$ of the same size. Then, we bound each term above in turn. Let $(\boldsymbol{W}^*, \boldsymbol{\Theta}^*) \in \mathcal{X}_f$ be such that $\operatorname{dist}((\boldsymbol{W}, \boldsymbol{\Theta}), \mathcal{X}_f) = \|(\boldsymbol{W}, \boldsymbol{\Theta}) - (\boldsymbol{W}^*, \boldsymbol{\Theta}^*)\|_F \le \delta$. According to (33) and (36), we have

$$\|\boldsymbol{W}\| \le \|\boldsymbol{W} - \boldsymbol{W}^*\| + \|\boldsymbol{W}^*\| \le \delta + \rho\sqrt{c} \le 2\rho\sqrt{c},\ \|\boldsymbol{\Theta}\| \le \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\| + \|\boldsymbol{\Theta}^*\| \le \frac{2\sqrt{c}}{\rho}. \tag{37}$$

It directly follows from (12) that

$$\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{\Theta})\|_F = \|\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T + \lambda_W K\boldsymbol{I}\right)\boldsymbol{W} - \boldsymbol{\Theta}\|_F, \tag{38}$$

$$\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W}, \boldsymbol{\Theta})\|_F = \|\left(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K\boldsymbol{I}\right)\boldsymbol{\Theta} - \boldsymbol{W}\|_F. \tag{39}$$

Summing up $\sqrt{n\lambda_H}\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \sqrt{\lambda_W}\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F$ yields

$$\begin{aligned}
&\sqrt{n\lambda_H}\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \sqrt{\lambda_W}\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F \\
&= \sqrt{n\lambda_H}\left\| \left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T + \lambda_W K \boldsymbol{I}\right)\boldsymbol{W} - \boldsymbol{\Theta}\right\|_F + \sqrt{\lambda_W}\left\| \left(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K \boldsymbol{I}\right)\boldsymbol{\Theta} - \boldsymbol{W}\right\|_F.
\end{aligned} \tag{40}$$

This, together with (37), yields

$$\begin{aligned}
&2\max\left\{\rho, \frac{1}{\rho}\right\}\sqrt{c\lambda_{\max}}\left(\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right) \\
&\geq 2\max\left\{\rho, \frac{1}{\rho}\right\}\sqrt{c}\left(\sqrt{n\lambda_H}\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \sqrt{\lambda_W}\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right) \\
&\geq \sqrt{n\lambda_H}\left\|\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T + \lambda_W K\boldsymbol{I}\right)\boldsymbol{W} - \boldsymbol{\Theta}\right\|_F \|\boldsymbol{W}\| + \sqrt{\lambda_W}\left\|\left(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K\boldsymbol{I}\right)\boldsymbol{\Theta} - \boldsymbol{W}\right\|_F \|\boldsymbol{\Theta}\| \\
&\geq \sqrt{\lambda_{\min}}\left(\left\|\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T + \lambda_W K\boldsymbol{I}\right)\boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{\Theta}\boldsymbol{W}^T\right\|_F + \left\|\left(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K\boldsymbol{I}\right)\boldsymbol{\Theta}\boldsymbol{\Theta}^T - \boldsymbol{W}\boldsymbol{\Theta}^T\right\|_F\right) \\
&\geq K\sqrt{\lambda_{\min}}\left\|\lambda_W \boldsymbol{W}\boldsymbol{W}^T - n\lambda_H\boldsymbol{\Theta}\boldsymbol{\Theta}^T\right\|_F,
\end{aligned}$$

where the last inequality uses the triangle inequality. This implies

$$\left\|\lambda_W \boldsymbol{W}\boldsymbol{W}^T - n\lambda_H\boldsymbol{\Theta}\boldsymbol{\Theta}^T\right\|_F \leq \kappa_1\left(\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right), \tag{41}$$

where

$$\kappa_1 := \frac{2\max\{\rho, 1/\rho\}\sqrt{c\lambda_{\max}}}{K\sqrt{\lambda_{\min}}}. \tag{42}$$

By letting $\boldsymbol{\Phi} := \sqrt{\lambda_W}\boldsymbol{W} - \sqrt{n\lambda_H}\boldsymbol{\Theta}$, we obtain

$$\begin{aligned}
&\sqrt{\lambda_{\max}}\left(\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right) \\
&\geq \sqrt{n\lambda_H}\left\| \left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T + \lambda_W K \boldsymbol{I}\right)\boldsymbol{W} - \boldsymbol{\Theta}\right\|_F + \sqrt{\lambda_W}\left\| \left(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K \boldsymbol{I}\right)\boldsymbol{\Theta} - \boldsymbol{W}\right\|_F \\
&\geq \left\|\sqrt{n\lambda_H}\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T + \lambda_W K\boldsymbol{I}\right)\boldsymbol{W} - \sqrt{\lambda_W}\left(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K\boldsymbol{I}\right)\boldsymbol{\Theta} + \boldsymbol{\Phi}\right\|_F \\
&= \left\|\sqrt{n\lambda_H}\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^T - \frac{\lambda_W}{n\lambda_H}\boldsymbol{W}\boldsymbol{W}^T\right)\boldsymbol{W} + \left(\frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}\boldsymbol{W}^T + (K\sqrt{n\lambda_H\lambda_W} + 1)\boldsymbol{I}\right)\boldsymbol{\Phi}\right\|_F \\
&\geq \left\|\left(\frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}\boldsymbol{W}^T + (K\sqrt{n\lambda_H\lambda_W} + 1)\boldsymbol{I}\right)\boldsymbol{\Phi}\right\|_F - \sqrt{n\lambda_H}\|\boldsymbol{W}\|\left\|\boldsymbol{\Theta}\boldsymbol{\Theta}^T - \frac{\lambda_W}{n\lambda_H}\boldsymbol{W}\boldsymbol{W}^T\right\|_F \\
&\geq \left(1 + K\sqrt{n\lambda_H\lambda_W}\right)\|\boldsymbol{\Phi}\|_F - \frac{2\rho\sqrt{c}}{\sqrt{n\lambda_H}}\left\|\lambda_W\boldsymbol{W}\boldsymbol{W}^T - n\lambda_H\boldsymbol{\Theta}\boldsymbol{\Theta}^T\right\|_F,
\end{aligned}$$

where the first inequality uses (40), and the last inequality is due to (37). This, together with (41), implies

$$\left\|\boldsymbol{\Theta} - \frac{1}{\rho^2}\boldsymbol{W}\right\|_F = \frac{1}{\sqrt{n\lambda_H}}\left\|\sqrt{\lambda_W}\boldsymbol{W} - \sqrt{n\lambda_H}\boldsymbol{\Theta}\right\|_F \leq \kappa_2\left(\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right), \tag{43}$$

where

$$\kappa_2 := \frac{\sqrt{\lambda_{\max}} + 2\rho\kappa_1\sqrt{c}/\sqrt{n\lambda_H}}{\sqrt{n\lambda_H}\left(1 + K\sqrt{n\lambda_H\lambda_W}\right)}. \tag{44}$$

Using (39), we compute

$$\begin{aligned}
\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F &= \left\|(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K\boldsymbol{I})\left(\boldsymbol{\Theta} - \frac{1}{\rho^2}\boldsymbol{W} + \frac{1}{\rho^2}\boldsymbol{W}\right) - \boldsymbol{W}\right\|_F \\
&\geq \frac{1}{\rho^2}\left\|(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K\boldsymbol{I})\,\boldsymbol{W} - \rho^2\boldsymbol{W}\right\|_F - \left\|(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K\boldsymbol{I})\left(\boldsymbol{\Theta} - \frac{1}{\rho^2}\boldsymbol{W}\right)\right\|_F \\
&= \frac{1}{\rho^2}\left\|\boldsymbol{W}\left(\boldsymbol{W}^T\boldsymbol{W} - \rho^2 c\boldsymbol{I}\right)\right\|_F - \left\|(\boldsymbol{W}\boldsymbol{W}^T + n\lambda_H K\boldsymbol{I})\left(\boldsymbol{\Theta} - \frac{1}{\rho^2}\boldsymbol{W}\right)\right\|_F \\
&\geq \frac{\sigma_{\min}(\boldsymbol{W})}{\rho^2}\|\boldsymbol{W}^T\boldsymbol{W} - \rho^2 c\boldsymbol{I}\|_F - \left(\|\boldsymbol{W}\|^2 + n\lambda_H K\right)\left\|\boldsymbol{\Theta} - \frac{1}{\rho^2}\boldsymbol{W}\right\|_F,
\end{aligned}$$

where the second equality follows from $\rho^2 - n\lambda_H K = \sqrt{n\lambda_H}/\sqrt{\lambda_W} - n\lambda_H K = \rho^2 c$. This, together with (37) and (43), yields

$$\frac{\sigma_{\min}(\boldsymbol{W})}{\rho^2}\|\boldsymbol{W}^T\boldsymbol{W} - \rho^2 c\boldsymbol{I}\|_F \leq \left(\kappa_2(4\rho^2 c + n\lambda_H K) + 1\right)\left(\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right). \quad (45)$$

Let $\boldsymbol{W} = \boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{V}_1^T$ be the thin singular value decomposition of $\boldsymbol{W}$, where $\boldsymbol{U}_1 \in \mathcal{O}^{d\times K}$, $\boldsymbol{V}_1 \in \mathcal{O}^K$, and $\boldsymbol{\Lambda}_1 \in \mathbb{R}^{K\times K}$ is a diagonal matrix. Now, we compute

$$\begin{aligned}
\|\boldsymbol{W}^T\boldsymbol{W} - \rho^2 c\boldsymbol{I}\|_F &= \|\boldsymbol{\Lambda}_1^2 - \rho^2 c\boldsymbol{I}\|_F = \|(\boldsymbol{\Lambda}_1 - \rho\sqrt{c}\boldsymbol{I})(\boldsymbol{\Lambda}_1 + \rho\sqrt{c}\boldsymbol{I})\|_F \\
&\geq \rho\sqrt{c}\|\boldsymbol{\Lambda}_1 - \rho\sqrt{c}\boldsymbol{I}\|_F = \rho\sqrt{c}\|\boldsymbol{U}_1(\boldsymbol{\Lambda}_1 - \rho\sqrt{c}\boldsymbol{I})\boldsymbol{V}_1^T\|_F \\
&= \rho\sqrt{c}\|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}_1\boldsymbol{V}_1^T\|_F \geq \rho\sqrt{c}\min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}}\left\|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\right\|_F. \quad (46)
\end{aligned}$$

Let $\boldsymbol{U}^* \in \mathcal{O}^{d\times K}$ be such that

$$\|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}^*\|_F = \min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}}\left\|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\right\|_F.$$

Using Weyl's inequality, we obtain

$$\sigma_{\min}(\boldsymbol{W}) \geq \sigma_{\min}(\rho\sqrt{c}\boldsymbol{U}^*) - \|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}^*\| \geq \rho\sqrt{c} - \delta \geq \frac{1}{2}\rho\sqrt{c},$$

where the second inequality follows from the equality in (36) and (33), and the last inequality is due to $\delta \leq \rho\sqrt{c}/2$ by (33). Substituting this and (46) into (45) yields

$$\min_{\boldsymbol{U}\in\mathcal{O}^{d\times K}}\left\|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\right\|_F \leq \frac{2}{c}\left(\kappa_2(4\rho^2 c + n\lambda_H K) + 1\right)\left(\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right)$$

This, together with (43) and (36), yields

$$\begin{aligned}
\mathrm{dist}^2\left((\boldsymbol{W},\boldsymbol{\Theta}),\mathcal{X}_f\right) &\leq 2\kappa_2^2\left(\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right)^2 + \left(1 + \frac{2\lambda_W}{n\lambda_H}\right) \\
&\quad \frac{4}{c^2}\left(\kappa_2(4\rho^2 c + n\lambda_H K) + 1\right)^2\left(\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right)^2 \\
&\leq \left(4\kappa_2^2 + \frac{8}{c^2}\left(\kappa_2(4\rho^2 c + n\lambda_H K) + 1\right)^2\right)\|\nabla f(\boldsymbol{W},\boldsymbol{\Theta})\|_F^2.
\end{aligned}$$

Then, we complete the proof. □

3 MSE Loss with a Bias Term

We consider the following problems:

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} F(\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}) = \frac{1}{2N}\|\boldsymbol{W}^T\boldsymbol{H} + \boldsymbol{b}\mathbf{1}_N^T - \boldsymbol{Y}\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2 + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2, \tag{47}$$

$$\min_{\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}) = \frac{1}{2K}\|\boldsymbol{W}^T\boldsymbol{\Theta} + \boldsymbol{b}\mathbf{1}_K^T - \boldsymbol{I}_K\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}\|_F^2 + \frac{\lambda_b}{2}\|\boldsymbol{b}\|^2, \tag{48}$$

where $\lambda_W, \lambda_H, \lambda_b > 0$ are the penalties for $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{b}$, respectively. One can easily verify

$$F(\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{W},\boldsymbol{H}_i,\boldsymbol{b}). \tag{49}$$

To proceed, let

$$\bar{\boldsymbol{V}} := \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{(K-1)K}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{(K-1)K}} \\ 0 & -\frac{\sqrt{2}}{\sqrt{3}} & \cdots & \frac{1}{\sqrt{(K-1)K}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{\sqrt{K-1}}{\sqrt{K}} \end{bmatrix} \in \mathcal{O}^{K\times(K-1)}. \tag{50}$$

Proposition 2. *The optimal solution set of Problem* (48) *can be characterized as follows:*
(i) If $\lambda_W\lambda_H \geq 1/(nK^2)$, *we have*

$$\mathcal{X}_f = \left\{\left(\mathbf{0},\mathbf{0},\frac{1}{K(1+\lambda_b)}\mathbf{1}_K\right)\right\}.$$

(ii) If $\lambda_b^2/(nK^2(1+\lambda_b)^2) \leq \lambda_W\lambda_H < 1/(nK^2)$, *we have*

$$\mathcal{X}_f = \left\{\left(\boldsymbol{W},\boldsymbol{\Theta},\frac{1}{K(1+\lambda_b)}\mathbf{1}_K\right) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}\sqrt{c}\boldsymbol{U}\bar{\boldsymbol{V}}^T, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}, \boldsymbol{U} \in \mathcal{O}^{d\times(K-1)}\right\}, \tag{51}$$

where $c := 1 - K\sqrt{n\lambda_W\lambda_H}$.
(iii) If $\lambda_W\lambda_H < \lambda_b^2/(nK^2(1+\lambda_b)^2)$, *we have*

$$\mathcal{X} = \left\{\left(\boldsymbol{W},\boldsymbol{\Theta},\frac{\sqrt{n\lambda_W\lambda_H}}{\lambda_b}\mathbf{1}_K\right) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}\boldsymbol{U}\begin{bmatrix}\sqrt{c}\bar{\boldsymbol{V}}^T \\ \sqrt{c'/K}\mathbf{1}^T\end{bmatrix}, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}, \boldsymbol{U} \in \mathcal{O}^{d\times K}\right\},$$

where $c' := 1 - K(1+\lambda_b)\sqrt{n\lambda_W\lambda_H}/\lambda_b$.

Proof. Using the first-order optimality of $\boldsymbol{b}$, we obtain

$$\boldsymbol{b}^* = \frac{1}{K(1+\lambda_b)}(\boldsymbol{I} - \boldsymbol{W}^T\boldsymbol{\Theta})\mathbf{1}. \tag{52}$$

Substituting this back into Problem (48) yields

$$\begin{aligned} &\frac{1}{2K}\|\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{I}\|_F^2 - \frac{1}{2K^2(1+\lambda_b)}\|(\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{I})\mathbf{1}\|^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{H}\|_F^2 \\ &= \frac{1}{2K}\|(\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{I})(\boldsymbol{I} - \alpha\mathbf{1}\mathbf{1}^T)\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}\|_F^2, \end{aligned}$$

where $\alpha := \frac{1+\sqrt{1-1/(1+\lambda_b)}}{K}$. Then, it suffices to consider

$$\min_{\boldsymbol{W},\boldsymbol{\Theta}\in\mathbb{R}^{d\times K}} h(\boldsymbol{W},\boldsymbol{\Theta}) := \frac{1}{2K}\|(\boldsymbol{W}^T\boldsymbol{\Theta}-\boldsymbol{I})(\boldsymbol{I}-\alpha\mathbf{1}\mathbf{1}^T)\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}\|_F^2. \tag{53}$$

Noting that $\boldsymbol{I}-\alpha\mathbf{1}\mathbf{1}^T = \boldsymbol{P} + (1-K\alpha)\boldsymbol{P}^\perp$, we have

$$\begin{aligned} h(\boldsymbol{W},\boldsymbol{\Theta}) &= \frac{1}{2K}\|(\boldsymbol{W}^T\boldsymbol{\Theta}-\boldsymbol{I})(\boldsymbol{P}+(1-K\alpha)\boldsymbol{P}^\perp)\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}\|_F^2 \\ &= \frac{1}{2K}\|(\boldsymbol{W}^T\boldsymbol{\Theta}-\boldsymbol{I})\boldsymbol{P}\|_F^2 + \frac{\lambda_b}{2K(1+\lambda_b)}\|(\boldsymbol{W}^T\boldsymbol{\Theta}-\boldsymbol{I})\boldsymbol{P}^\perp\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}\|_F^2, \end{aligned}$$

where the second equality follows from $(1-K\alpha)^2 = 1/(1+\lambda_b)$. For ease of exposition, let

$$\boldsymbol{W}_1 := \boldsymbol{W}\boldsymbol{P},\ \boldsymbol{W}_2 := \boldsymbol{W}\boldsymbol{P}^\perp,\ \boldsymbol{\Theta}_1 := \boldsymbol{\Theta}\boldsymbol{P},\ \boldsymbol{\Theta}_2 := \boldsymbol{\Theta}\boldsymbol{P}^\perp. \tag{54}$$

Since $\boldsymbol{P}$ and $\boldsymbol{P}^\perp$ are both projection matrices, we have

$$\begin{aligned} h(\boldsymbol{W},\boldsymbol{\Theta}) \geq \frac{1}{2K}&\|\boldsymbol{P}(\boldsymbol{W}^T\boldsymbol{\Theta}-\boldsymbol{I})\boldsymbol{P}\|_F^2 + \frac{\lambda_b}{2K(1+\lambda_b)}\|\boldsymbol{P}^\perp(\boldsymbol{W}^T\boldsymbol{\Theta}-\boldsymbol{I})\boldsymbol{P}^\perp\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 \\ &+ \frac{n\lambda_H}{2}\|\boldsymbol{H}\|_F^2 = \frac{1}{2K}\|\boldsymbol{W}_1^T\boldsymbol{\Theta}_1-\boldsymbol{P}\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}_1\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}_1\|_F^2 + \\ &\frac{\lambda_b}{2K(1+\lambda_b)}\|\boldsymbol{W}_2^T\boldsymbol{\Theta}_2-\boldsymbol{P}^\perp\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}_2\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}_2\|_F^2 \end{aligned} \tag{55}$$

where the first inequality becomes equality if and only if $\boldsymbol{W}_2^T\boldsymbol{\Theta}_1 = \mathbf{0}$ and $\boldsymbol{W}_1^T\boldsymbol{\Theta}_2 = \mathbf{0}$. This yields

$$\begin{aligned} h(\boldsymbol{W},\boldsymbol{\Theta}) \geq &\min_{\boldsymbol{W}_1,\boldsymbol{\Theta}_1}\left\{\frac{1}{2K}\|\boldsymbol{W}_1^T\boldsymbol{\Theta}_1-\boldsymbol{P}\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}_1\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}_1\|_F^2\right\} + \\ &\frac{\lambda_b}{1+\lambda_b}\min_{\boldsymbol{W}_2,\boldsymbol{\Theta}_2}\left\{\frac{1}{2K}\|\boldsymbol{W}_2^T\boldsymbol{\Theta}_2-\boldsymbol{P}^\perp\|_F^2 + \frac{\lambda_W(1+\lambda_b)}{2\lambda_b}\|\boldsymbol{W}_2\|_F^2 + \frac{n\lambda_H(1+\lambda_b)}{2\lambda_b}\|\boldsymbol{\Theta}_2\|_F^2\right\}, \end{aligned} \tag{56}$$

where the inequality becomes equality if and only if there exists $\boldsymbol{W}^*,\boldsymbol{\Theta}^*$ such that the optimal solutions $(\boldsymbol{W}_1^*,\boldsymbol{\Theta}_1^*)$ and $(\boldsymbol{W}_2^*,\boldsymbol{\Theta}_2^*)$ of the above optimization problems satisfy $\boldsymbol{W}_1^* = \boldsymbol{W}^*\boldsymbol{P},\ \boldsymbol{W}_2^* = \boldsymbol{W}^*\boldsymbol{P}^\perp,\ \boldsymbol{\Theta}_1^* = \boldsymbol{\Theta}^*\boldsymbol{P},\ \boldsymbol{\Theta}_2 = \boldsymbol{\Theta}^*\boldsymbol{P}^\perp$. Now, we can optimize the above two optimization problems, respectively. One can verify that the matrix $\boldsymbol{P}$ admits the eigenvalue decomposition $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{I}_{K-1} & \mathbf{0}\\ \mathbf{0} & 0\end{bmatrix}$, and $\boldsymbol{V} = [\bar{\boldsymbol{V}}\ \mathbf{1}_K/\sqrt{K}] \in \mathcal{O}^K$ with $\bar{\boldsymbol{V}}\bar{\boldsymbol{V}}^T = \boldsymbol{P}$ and $\boldsymbol{v}\boldsymbol{v}^T = \boldsymbol{P}^\perp$. Then, we can verify that $(\boldsymbol{W}_1^*,\boldsymbol{\Theta}_1^*)$ is an optimal solution to the first optimization problem in (56) if and only if $(\boldsymbol{W}_1^*\boldsymbol{V},\boldsymbol{\Theta}_1^*\boldsymbol{V})$ is an optimal solution to

$$\min_{\boldsymbol{W}_1,\boldsymbol{\Theta}_1}\left\{\frac{1}{2K}\|\boldsymbol{W}_1^T\boldsymbol{\Theta}_1-\boldsymbol{\Sigma}\|_F^2 + \frac{\lambda_W}{2}\|\boldsymbol{W}_1\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}_1\|_F^2\right\}$$

This, together with Lemma 1, implies that the solution set of the first optimization problem in (56) with variables $\boldsymbol{W}_1,\boldsymbol{\Theta}_1$ is as follows:

$$\mathcal{X}_1 := \left\{(\boldsymbol{W}\boldsymbol{V}^T,\boldsymbol{\Theta}\boldsymbol{V}^T) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}\boldsymbol{U}\begin{bmatrix}(\max\{c,0\})^{1/2}\,\boldsymbol{I}_{K-1} & \mathbf{0}\\ \mathbf{0} & 0\end{bmatrix}, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}, \boldsymbol{U}\in\mathcal{O}^{d\times K}\right\}.$$

where $c := 1 - K\sqrt{n\lambda_W \lambda_H}$. Note that $\boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{P} = \boldsymbol{V}(\boldsymbol{I} - \boldsymbol{\Sigma})\boldsymbol{V}^T$. By the same argument, we show that the solution set of the second optimization problem in (56) with variables $\boldsymbol{W}_2, \boldsymbol{\Theta}_2$ is as follows:

$$\mathcal{X}_2 := \left\{ (\boldsymbol{W}\boldsymbol{V}^T, \boldsymbol{\Theta}\boldsymbol{V}^T) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}} \boldsymbol{U} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & (\max\{c', 0\}, 0)^{1/2} \end{bmatrix}, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}} \boldsymbol{W},\ \boldsymbol{U} \in \mathcal{O}^{d\times K} \right\},$$

where $c' := 1 - K(1+\lambda_b)\sqrt{n\lambda_W\lambda_H}/\lambda_b$. Now, we discuss the optimal solution set of Problem (53) denoted by $\mathcal{X}$ case by case.

Case 1. If $\lambda_W\lambda_H \ge \frac{1}{nK^2}$, then $c = c' = 0$. Therefore, we obtain $\mathcal{X}_1 = \{(\boldsymbol{0}, \boldsymbol{0})\}$ and $\mathcal{X}_2 = \{(\boldsymbol{0}, \boldsymbol{0})\}$. Then, one can verify that the inequalities become equalities in (55) and (56) for any $(\boldsymbol{W}_1^*, \boldsymbol{\Theta}_1^*) \in \mathcal{X}_1$ and $(\boldsymbol{W}_2^*, \boldsymbol{\Theta}_2^*) \in \mathcal{X}_2$. This directly implies $\mathcal{X} = \{(\boldsymbol{0}, \boldsymbol{0})\}$.

Case 2. If $\frac{\lambda_b^2}{nK^2(1+\lambda_b)^2} \le \lambda_W\lambda_H < \frac{1}{nK^2}$, then $c > 0$ and $c' = 0$. Therefore, we obtain $\mathcal{X}_2 = \{(\boldsymbol{0}, \boldsymbol{0})\}$ and

$$\mathcal{X}_1 = \left\{ (\boldsymbol{W}, \boldsymbol{\Theta}) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}} \sqrt{c}\boldsymbol{U}\bar{\boldsymbol{V}}^T, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{\lambda_H}} \boldsymbol{W}, \boldsymbol{U} \in \mathcal{O}^{d\times(K-1)} \right\}.$$

Then, we can verify that the inequalities become equalities in (55) and (56) for any $(\boldsymbol{W}_1^*, \boldsymbol{\Theta}_1^*) \in \mathcal{X}_1$ and $(\boldsymbol{W}_2^*, \boldsymbol{\Theta}_2^*) \in \mathcal{X}_2$. This directly implies $\mathcal{X} = \mathcal{X}_1$.

Case 3. If $\lambda_W\lambda_H < \frac{\lambda_b^2}{K^2(1+\lambda_b)^2}$, then $c > 0$ and $c' > 0$. Therefore, we obtain that $\mathcal{X}_1$ takes the same form as above and

$$\mathcal{X}_2 = \left\{ (\boldsymbol{W}, \boldsymbol{\Theta}) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}\sqrt{c'}}{\sqrt[4]{\lambda_W}\sqrt{K}} \boldsymbol{u}\boldsymbol{1}^T, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}} \boldsymbol{W}, \|\boldsymbol{u}\| = 1 \right\}.$$

For any $(\boldsymbol{W}_1^*, \boldsymbol{\Theta}_1^*) \in \mathcal{X}_1$ and $(\boldsymbol{W}_2^*, \boldsymbol{\Theta}_2^*) \in \mathcal{X}_2$, we can verify that ${\boldsymbol{W}_1^*}^T \boldsymbol{W}_2^* = \boldsymbol{0}$ holds if and only if $\boldsymbol{U}^T\boldsymbol{u} = 0$ due to $\bar{\boldsymbol{V}}^T\bar{\boldsymbol{V}} = \boldsymbol{I}$ and $\|\boldsymbol{1}_K\| = \sqrt{K}$. As a result, the inequality in (55) becomes equality. It follows from $\boldsymbol{P} = \bar{\boldsymbol{V}}\bar{\boldsymbol{V}}^T$ and $\boldsymbol{P}^\perp = \boldsymbol{1}\boldsymbol{1}^T/K$ that the inequality in (56) becomes equality. Using (54), $\boldsymbol{U}^T\boldsymbol{u} = 0$, $\boldsymbol{U} \in \mathcal{O}^{d\times(K-1)}$, and $\|\boldsymbol{u}\| = 1$, we have

$$\mathcal{X} = \left\{ (\boldsymbol{W}, \boldsymbol{\Theta}) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}} \boldsymbol{U} \begin{bmatrix} \sqrt{c}\bar{\boldsymbol{V}}^T \\ \sqrt{c'/K}\boldsymbol{1}_K^T \end{bmatrix}, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}} \boldsymbol{W}, \boldsymbol{U} \in \mathcal{O}^{d\times K} \right\}.$$

Combining the above cases with (52) yields that desired result. $\square$

Theorem 2. *Suppose that $\lambda_b^2/(nK^2(1+\lambda_b)^2) < \lambda_W\lambda_H < 1/nK^2$. For all $(\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b})$ satisfying*

$$\operatorname{dist}((\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b}), \mathcal{X}_f) \le \delta := \frac{1}{2}\min\left\{ \frac{\sqrt[4]{\lambda_W}}{\sqrt[4]{n\lambda_H}}, \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}} \right\} \left(1 - K\sqrt{n\lambda_W\lambda_H}\right)^{1/2}, \tag{57}$$

it holds that

$$\operatorname{dist}((\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b}), \mathcal{X}_f) \le \kappa \|\nabla f(\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b})\|_F, \tag{58}$$

where κ is a constant that depends on λ_W, λ_H, and λ_b.

Proof. To simplify our notation, let

$$\alpha := \frac{1 + \sqrt{1 - 1/(1+\lambda_b)}}{K},\ \beta := \frac{\lambda_b}{1+\lambda_b},\ c := 1 - K\sqrt{n\lambda_W\lambda_H},\ \rho := \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}. \tag{59}$$

According to the proof of Proposition 2 and $\lambda_b^2/(nK^2(1+\lambda_b)^2) \le \lambda_W\lambda_H < 1/nK^2$, the optimal solution set of Problem (53) takes the form of

$$\mathcal{X}_h := \left\{(\boldsymbol{W},\boldsymbol{\Theta}) : \boldsymbol{W} = \rho\sqrt{c}\boldsymbol{U}\bar{\boldsymbol{V}}^T, \boldsymbol{\Theta} = \frac{1}{\rho^2}\boldsymbol{W}, \boldsymbol{U} \in \mathcal{O}^{d\times(K-1)}\right\}. \tag{60}$$

We claim that for all $(\boldsymbol{W},\boldsymbol{\Theta})$ satisfying

$$\operatorname{dist}\left((\boldsymbol{W},\boldsymbol{\Theta}),\mathcal{X}_h\right) \le \delta_1 := \frac{1}{2}\min\left\{\rho,\frac{1}{\rho}\right\}\sqrt{c}, \tag{61}$$

it holds that

$$\operatorname{dist}\left((\boldsymbol{W},\boldsymbol{\Theta}),\mathcal{X}_h\right) \le \kappa_1\|\nabla h(\boldsymbol{W},\boldsymbol{\Theta})\|_F, \tag{62}$$

where δ_1,κ_1 are positive constants that depend on λ_W and λ_H. Let $(\boldsymbol{W}^*,\boldsymbol{\Theta}^*) \in \mathcal{X}_h$ be such that $\operatorname{dist}((\boldsymbol{W},\boldsymbol{H}),\mathcal{X}_h) = \|(\boldsymbol{W},\boldsymbol{H}) - (\boldsymbol{W}^*,\boldsymbol{H}^*)\|_F \le \delta_1$. According to (60) and $\delta_1 \le \min\{\rho,1/\rho\}\sqrt{c}$, we have

$$\|\boldsymbol{W}\| \le \|\boldsymbol{W}-\boldsymbol{W}^*\| + \|\boldsymbol{W}^*\| \le \delta_1 + \rho\sqrt{c} \le 2\rho\sqrt{c},\ \|\boldsymbol{\Theta}\| \le \|\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\| + \|\boldsymbol{\Theta}^*\| \le \frac{2\sqrt{c}}{\rho}. \tag{63}$$

Since $f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b})$ is strongly convex w.r.t. $\boldsymbol{b}$ with constant λ_b, we obtain

$$\lambda_b\|\boldsymbol{b}-\boldsymbol{b}^*\|^2 \le \langle \nabla_{\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}) - \nabla_{\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}^*), \boldsymbol{b}-\boldsymbol{b}^*\rangle,$$

where $\boldsymbol{b}^*$ is defined in (52). This, together with $\nabla_{\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}^*) = \boldsymbol{0}$ and the Cauchy–Schwarz inequality, implies

$$\|\boldsymbol{b}-\boldsymbol{b}^*\| \le \frac{1}{\lambda_b}\|\nabla_{\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b})\| \tag{64}$$

We compute

$$\begin{cases} \nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}) := \frac{1}{K}\boldsymbol{\Theta}(\boldsymbol{W}^T\boldsymbol{\Theta} + \boldsymbol{b}\boldsymbol{1}_K^T - \boldsymbol{I}_K)^T + \lambda_W\boldsymbol{W}, \\ \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}) := \frac{1}{K}\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{\Theta} + \boldsymbol{b}\boldsymbol{1}_K^T - \boldsymbol{I}_K) + n\lambda_H\boldsymbol{H}. \end{cases} \tag{65}$$

It follows from (52) and (53) that $h(\boldsymbol{W},\boldsymbol{\Theta}) = f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}^*)$, which implies $\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta}) = \nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}^*)$ and $\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta}) = \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}^*)$. Therefore, we have

$$\begin{aligned} \|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}) - \nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F &= \|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}) - \nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b}^*)\|_F \\ &= \frac{1}{K}\left\|\boldsymbol{\Theta}\boldsymbol{1}_K(\boldsymbol{b}-\boldsymbol{b}^*)\right\|_F \le \frac{2\sqrt{c}}{\rho\sqrt{K}}\|\boldsymbol{b}-\boldsymbol{b}^*\|, \end{aligned}$$

where the inequality follows from (63). This implies

$$\begin{aligned} \|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F &\le \|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b})\|_F + \frac{2\sqrt{c}}{\rho\sqrt{K}}\|\boldsymbol{b}-\boldsymbol{b}^*\| \\ &\le \|\nabla_{\boldsymbol{W}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b})\|_F + \frac{2\sqrt{c}}{\lambda_b\rho\sqrt{K}}\|\nabla_{\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b})\|, \end{aligned} \tag{66}$$

where the last inequality uses (64). By the same argument, we obtain

$$\|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F \le \|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b})\|_F + \frac{2\rho\sqrt{c}}{\lambda_b\sqrt{K}}\|\nabla_{\boldsymbol{b}} f(\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{b})\| \tag{67}$$

According to (51) and (60), we have for all $(\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b})$,

$$
\begin{aligned}
\operatorname{dist}^2\left((\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b}), \mathcal{X}_f\right) &= \operatorname{dist}^2\left((\boldsymbol{W}, \boldsymbol{\Theta}), \mathcal{X}_h\right) + \|\boldsymbol{b} - \boldsymbol{b}^*\|^2 \leq \kappa_1^2 \|\nabla h(\boldsymbol{W}, \boldsymbol{\Theta})\|_F^2 + \|\boldsymbol{b} - \boldsymbol{b}^*\|^2 \\
&\leq 2\kappa_1^2 \left(\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b})\|_F^2 + \frac{4c}{\lambda_b^2 \rho^2 K} \|\nabla_{\boldsymbol{b}} f(\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b})\|^2 \right) + \\
&\quad 2\kappa_1^2 \left(\|\nabla_{\boldsymbol{\Theta}} f(\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b})\|_F^2 + \frac{4\rho^2 c}{\lambda_b^2 K} \|\nabla_{\boldsymbol{b}} f(\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b})\|^2 \right) + \frac{1}{\lambda_b^2} \|\nabla_{\boldsymbol{b}} f(\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b})\|^2 \\
&\leq 2\kappa_1^2 \max\left\{1, \frac{4c}{\lambda_b^2 K}\left(\frac{1}{\rho^2} + \rho^2\right) + \frac{1}{2\kappa_1^2 \lambda_b^2}\right\} \|\nabla f(\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{b})\|_F^2,
\end{aligned}
$$

where the first inequality follows from (57) and (62), where the second inequality uses (66), (67), and (64).

The rest of the proof is devoted to proving our claim. For ease of exposition, let

$$
\boldsymbol{W}_1 := \boldsymbol{W}\boldsymbol{P},\ \boldsymbol{W}_2 := \boldsymbol{W}\boldsymbol{P}^{\perp},\ \boldsymbol{\Theta}_1 := \boldsymbol{\Theta}\boldsymbol{P},\ \boldsymbol{\Theta}_2 := \boldsymbol{\Theta}\boldsymbol{P}^{\perp}, \tag{68}
$$

and

$$
\gamma := K\sqrt{n\lambda_W \lambda_H} - \beta,\ \lambda_{\max} := \max\left\{\lambda_W, n\lambda_H\right\},\ \lambda_{\min} := \min\left\{\lambda_W, n\lambda_H\right\}. \tag{69}
$$

Since $(\boldsymbol{I} - \alpha \boldsymbol{1}\boldsymbol{1}^T)^2 = \boldsymbol{P} + \beta \boldsymbol{P}^{\perp}$, we compute

$$
\begin{cases}
K\nabla_{\boldsymbol{W}} h(\boldsymbol{W}, \boldsymbol{\Theta}) = \boldsymbol{\Theta}\left(\boldsymbol{P} + \beta \boldsymbol{P}^{\perp}\right)\left(\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{I}\right)^T + K\lambda_W \boldsymbol{W}, \\
K\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W}, \boldsymbol{\Theta}) = \boldsymbol{W}\left(\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{I}\right)\left(\boldsymbol{P} + \beta \boldsymbol{P}^{\perp}\right) + nK\lambda_H \boldsymbol{\Theta}.
\end{cases} \tag{70}
$$

According to (60), we compute

$$
\begin{aligned}
\operatorname{dist}^2\left((\boldsymbol{W}, \boldsymbol{\Theta}), \mathcal{X}_h\right) &= \min_{\boldsymbol{U} \in \mathcal{O}^{d \times (K-1)}} \left\{ \|\boldsymbol{W} - \rho\sqrt{c}\boldsymbol{U}\bar{\boldsymbol{V}}^T\|_F^2 + \left\|\boldsymbol{\Theta} - \frac{\sqrt{c}}{\rho}\boldsymbol{U}\bar{\boldsymbol{V}}^T\right\|_F^2 \right\} \\
= \min_{\boldsymbol{U} \in \mathcal{O}^{d \times (K-1)}} & \left\{ \|\boldsymbol{W}_1 - \rho\sqrt{c}\boldsymbol{U}\bar{\boldsymbol{V}}^T\|_F^2 + \left\|\boldsymbol{\Theta}_1 - \frac{\sqrt{c}}{\rho}\boldsymbol{U}\bar{\boldsymbol{V}}^T\right\|_F^2 \right\} + \|\boldsymbol{W}_2\|_F^2 + \|\boldsymbol{\Theta}_2\|_F^2 \qquad (71) \\
\leq \|\boldsymbol{W}_2\|_F^2 + \|\boldsymbol{\Theta}_2\|_F^2 &+ 2\left\|\boldsymbol{\Theta}_1 - \frac{1}{\rho^2}\boldsymbol{W}_1\right\|_F^2 + \left(1 + \frac{2\lambda_W}{n\lambda_H}\right) \min_{\boldsymbol{U} \in \mathcal{O}^{d \times (K-1)}} \|\boldsymbol{W}_1 - \rho\sqrt{c}\boldsymbol{U}\bar{\boldsymbol{V}}^T\|_F^2, \qquad (72)
\end{aligned}
$$

where the second equality is due to $\bar{\boldsymbol{V}}^T\boldsymbol{P}^{\perp} = \boldsymbol{0}$. Then, we bound each term above in turn. According to (70), we compute

$$
\begin{aligned}
K\nabla_{\boldsymbol{W}} h(\boldsymbol{W}, \boldsymbol{\Theta})\boldsymbol{P} &= \boldsymbol{\Theta}\left(\boldsymbol{P} + \beta \boldsymbol{P}^{\perp}\right)\left(\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{I}\right)^T \boldsymbol{P} + K\lambda_W \boldsymbol{W}\boldsymbol{P} \\
&= \left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T + \beta\boldsymbol{\Theta}_2\boldsymbol{\Theta}_2^T + K\lambda_W \boldsymbol{I}\right)\boldsymbol{W}_1 - \boldsymbol{\Theta}_1,
\end{aligned} \tag{73}
$$

and

$$
\begin{aligned}
K\nabla_{\boldsymbol{W}} h(\boldsymbol{W}, \boldsymbol{\Theta})\boldsymbol{P}^{\perp} &= \boldsymbol{\Theta}\left(\boldsymbol{P} + \beta \boldsymbol{P}^{\perp}\right)\left(\boldsymbol{W}^T\boldsymbol{\Theta} - \boldsymbol{I}\right)^T \boldsymbol{P}^{\perp} + K\lambda_W \boldsymbol{W}\boldsymbol{P}^{\perp} \\
&= \left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T + \beta\boldsymbol{\Theta}_2\boldsymbol{\Theta}_2^T + K\lambda_W \boldsymbol{I}\right)\boldsymbol{W}_2 - \beta\boldsymbol{\Theta}_2,
\end{aligned}
$$

which implies

$$
K\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W}, \boldsymbol{\Theta})\boldsymbol{P}^{\perp}\|_F \geq K\lambda_W \|\boldsymbol{W}_2\|_F - \beta\|\boldsymbol{\Theta}_2\|_F. \tag{74}
$$

Using the same computation, we obtain

$$\begin{cases} K\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P} = (\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H \boldsymbol{I})\boldsymbol{\Theta}_1 - \boldsymbol{W}_1, \\ K\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}^\perp = (\beta\boldsymbol{W}\boldsymbol{W}^T + nK\lambda_H \boldsymbol{I})\boldsymbol{\Theta}_2 - \beta\boldsymbol{W}_2. \end{cases} \tag{75}$$

Using the same argument in (73), we show

$$K\|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}^\perp\|_F \geq nK\lambda_H\|\boldsymbol{\Theta}_2\|_F - \beta\|\boldsymbol{W}_2\|_F. \tag{76}$$

It follows from (74), (76), and $\gamma := K\sqrt{n\lambda_W\lambda_H} - \beta$ that

$$\begin{aligned} & K\sqrt{n\lambda_H}\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}^\perp\|_F + K\sqrt{\lambda_W}\|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}^\perp\|_F \\ & \geq \gamma\left(\sqrt{\lambda_W}\|\boldsymbol{W}_2\|_F + \sqrt{n\lambda_H}\|\boldsymbol{H}_2\|_F\right). \end{aligned} \tag{77}$$

This implies

$$\begin{aligned} \|\boldsymbol{W}_2\|_F^2 + \|\boldsymbol{H}_2\|_F^2 & \leq \frac{2K^2\lambda_{\max}}{\gamma^2\lambda_{\min}}\left(\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}^\perp\|_F^2 + \|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}^\perp\|_F^2\right) \\ & \leq \frac{2K^2\lambda_{\max}}{\gamma^2\lambda_{\min}}\|\nabla h(\boldsymbol{W},\boldsymbol{\Theta})\|_F^2. \end{aligned} \tag{78}$$

Let $(\boldsymbol{W}^*,\boldsymbol{\Theta}^*) \in \mathcal{X}_h$ be such that $\mathrm{dist}((\boldsymbol{W},\boldsymbol{H}),\mathcal{X}_h) = \|(\boldsymbol{W},\boldsymbol{H}) - (\boldsymbol{W}^*,\boldsymbol{H}^*)\|_F \leq \delta_1$. According to (71) and $\delta_1 \leq \min\{\rho, 1/\rho\}\sqrt{c}$, we have

$$\|\boldsymbol{W}_1\| \leq \|\boldsymbol{W}_1 - \boldsymbol{W}_1^*\| + \|\boldsymbol{W}_1^*\| \leq 2\rho\sqrt{c},\ \|\boldsymbol{\Theta}_1\| \leq \|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_1^*\| + \|\boldsymbol{\Theta}_1^*\| \leq \frac{2\sqrt{c}}{\rho}, \tag{79}$$

$$\|\boldsymbol{W}_2\| \leq \delta_1,\ \|\boldsymbol{\Theta}_2\| \leq \delta_1. \tag{80}$$

According to (73), we compute

$$\begin{aligned} K\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}\|_F & \geq \|\left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T + K\lambda_W\boldsymbol{I}\right)\boldsymbol{W}_1 - \boldsymbol{\Theta}_1\|_F - \|\boldsymbol{\Theta}_2\boldsymbol{\Theta}_2^T\boldsymbol{W}_1\|_F \\ & \geq \|\left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T + K\lambda_W\boldsymbol{I}\right)\boldsymbol{W}_1 - \boldsymbol{\Theta}_1\|_F - c\|\boldsymbol{\Theta}_2\|_F, \end{aligned}$$

where the first inequality follows from the triangular inequality and $\beta < 1$, where the second inequality uses $\|\boldsymbol{\Theta}_2\boldsymbol{\Theta}_2^T\boldsymbol{W}_1\|_F \leq \|\boldsymbol{W}_1\|\|\boldsymbol{\Theta}_2\|\boldsymbol{\Theta}_2\|_F \leq c$ due to (80) and (61). By the same argument, we compute

$$K\|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}\|_F \geq \|\left(\boldsymbol{W}_1\boldsymbol{W}_1^T + nK\lambda_H\boldsymbol{I}\right)\boldsymbol{\Theta}_1 - \boldsymbol{W}_1\|_F - c\|\boldsymbol{W}_2\|_F.$$

Summing up $K\sqrt{n\lambda_H}\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}\|_F + K\sqrt{\lambda_W}\|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}\|_F$ yields

$$\begin{aligned} & K\left(\sqrt{n\lambda_H}\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}\|_F + \sqrt{\lambda_W}\|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\boldsymbol{P}\|_F\right) + c\left(\sqrt{n\lambda_H}\|\boldsymbol{\Theta}_2\|_F + \sqrt{\lambda_W}\|\boldsymbol{W}_2\|_F\right) \\ & \geq \sqrt{n\lambda_H}\|\left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T + K\lambda_W\boldsymbol{I}\right)\boldsymbol{W}_1 - \boldsymbol{\Theta}_1\|_F + \sqrt{\lambda_W}\|\left(\boldsymbol{W}_1\boldsymbol{W}_1^T + nK\lambda_H\boldsymbol{I}\right)\boldsymbol{\Theta}_1 - \boldsymbol{W}_1\|_F. \end{aligned}$$

Substituting (77) into the above inequality yields

$$\begin{aligned} & 2K\max\left\{1, \frac{c}{\gamma}\right\}\sqrt{\lambda_{\max}}\left(\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right) \\ & \geq \sqrt{n\lambda_H}\|\left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T + K\lambda_W\boldsymbol{I}\right)\boldsymbol{W}_1 - \boldsymbol{\Theta}_1\|_F + \sqrt{\lambda_W}\|\left(\boldsymbol{W}_1\boldsymbol{W}_1^T + nK\lambda_H\boldsymbol{I}\right)\boldsymbol{\Theta}_1 - \boldsymbol{W}_1\|_F. \end{aligned} \tag{81}$$

Using this and (79), we obtain

$$\begin{aligned}
&4K\max\left\{1,\frac{c}{\gamma}\right\}\max\left\{\rho,\frac{1}{\rho}\right\}\frac{\sqrt{\lambda_{\max}c}}{\sqrt{\lambda_{\min}}}\left(\|\nabla_{\boldsymbol{W}}h(\boldsymbol{W},\boldsymbol{\Theta})\|_F+\|\nabla_{\boldsymbol{\Theta}}h(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right)\\
&\geq \|\left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T+K\lambda_W\boldsymbol{I}\right)\boldsymbol{W}_1-\boldsymbol{\Theta}_1\|_F\|\boldsymbol{W}_1\|+\|\left(\boldsymbol{W}_1\boldsymbol{W}_1^T+nK\lambda_H\boldsymbol{I}\right)\boldsymbol{\Theta}_1-\boldsymbol{W}_1\|_F\|\boldsymbol{\Theta}_1\|\\
&\geq \|\left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T+K\lambda_W\boldsymbol{I}\right)\boldsymbol{W}_1\boldsymbol{W}_1^T-\boldsymbol{\Theta}_1\boldsymbol{W}_1^T\|_F+\|\left(\boldsymbol{W}_1\boldsymbol{W}_1^T+nK\lambda_H\boldsymbol{I}\right)\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T-\boldsymbol{W}_1\boldsymbol{\Theta}_1^T\|_F\\
&\geq K\|\lambda_W\boldsymbol{W}_1\boldsymbol{W}_1^T-n\lambda_H\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T\|_F.
\end{aligned}$$

Therefore, we have

$$\|\lambda_W\boldsymbol{W}_1\boldsymbol{W}_1^T-n\lambda_H\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T\|_F\leq\kappa_1\left(\|\nabla_{\boldsymbol{W}}h(\boldsymbol{W},\boldsymbol{\Theta})\|_F+\|\nabla_{\boldsymbol{\Theta}}h(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right), \tag{82}$$

where

$$\kappa_1:=4\max\left\{1,\frac{c}{\gamma}\right\}\max\left\{\rho,\frac{1}{\rho}\right\}\frac{\sqrt{\lambda_{\max}c}}{\sqrt{\lambda_{\min}}}. \tag{83}$$

According to (81), we compute

$$\begin{aligned}
&2K\max\left\{1,\frac{c}{\gamma}\right\}\sqrt{\lambda_{\max}}\left(\|\nabla_{\boldsymbol{W}}h(\boldsymbol{W},\boldsymbol{\Theta})\|_F+\|\nabla_{\boldsymbol{\Theta}}h(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right)\\
&\geq\|\sqrt{n\lambda_H}\left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T+K\lambda_W\boldsymbol{I}\right)\boldsymbol{W}_1-\sqrt{\lambda_W}\left(\boldsymbol{W}_1\boldsymbol{W}_1^T+nK\lambda_H\boldsymbol{I}\right)\boldsymbol{\Theta}_1+\sqrt{\lambda_W}\boldsymbol{W}_1-\sqrt{n\lambda_H}\boldsymbol{\Theta}_1\|_F\\
&=\left\|\frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}(\boldsymbol{W}_1\boldsymbol{W}_1^T+nK\lambda_H\boldsymbol{I})\left(\sqrt{\lambda_W}\boldsymbol{W}_1-\sqrt{n\lambda_H}\boldsymbol{\Theta}_1\right)+\sqrt{n\lambda_H}\left(\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T-\frac{\lambda_W}{n\lambda_H}\boldsymbol{W}_1\boldsymbol{W}_1^T\right)\boldsymbol{W}_1\right.\\
&\qquad\left.+\left(\sqrt{\lambda_W}\boldsymbol{W}_1-\sqrt{n\lambda_H}\boldsymbol{\Theta}_1\right)\right\|_F\\
&\geq\left\|\frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\left(\boldsymbol{W}_1\boldsymbol{W}_1^T+nK\lambda_H\boldsymbol{I}+\frac{\sqrt{n\lambda_H}}{\sqrt{\lambda_W}}\boldsymbol{I}\right)\left(\sqrt{\lambda_W}\boldsymbol{W}_1-\sqrt{n\lambda_H}\boldsymbol{\Theta}_1\right)\right\|_F-\\
&\qquad\sqrt{n\lambda_H}\|\boldsymbol{W}_1\|\left\|\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T-\frac{\lambda_W}{n\lambda_H}\boldsymbol{W}_1\boldsymbol{W}_1^T\right\|_F\\
&\geq\left(1+K\sqrt{n\lambda_W\lambda_H}\right)\left\|\sqrt{\lambda_W}\boldsymbol{W}_1-\sqrt{n\lambda_H}\boldsymbol{\Theta}_1\right\|_F-\frac{2\rho\sqrt{c}}{\sqrt{n\lambda_H}}\left\|n\lambda_H\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^T-\lambda_W\boldsymbol{W}_1\boldsymbol{W}_1^T\right\|_F,
\end{aligned}$$

where the last inequality follows from (79). This, together with (82), yields

$$\left\|\boldsymbol{\Theta}_1-\frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}_1\right\|_F=\frac{\left\|\sqrt{\lambda_W}\boldsymbol{W}_1-\sqrt{n\lambda_H}\boldsymbol{\Theta}_1\right\|_F}{\sqrt{n\lambda_H}}\leq\kappa_2\left(\|\nabla_{\boldsymbol{W}}h(\boldsymbol{W},\boldsymbol{\Theta})\|_F+\|\nabla_{\boldsymbol{\Theta}}h(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right), \tag{84}$$

where

$$\kappa_2:=\frac{1}{\sqrt{n\lambda_H}(1+K\sqrt{n\lambda_W\lambda_H})}\left(\frac{2\kappa_1\rho\sqrt{c}}{\sqrt{n\lambda_H}}+2K\max\left\{1,\frac{c}{\gamma}\right\}\sqrt{\lambda_{\max}}\right).$$

It follows from (61) and (71) that $\|\boldsymbol{W}_1-\boldsymbol{W}_1^*\|_F\leq\delta_1$. This, together with Weyl's inequality, implies

$$\sigma_{K-1}(\boldsymbol{W}_1)\geq\sigma_{K-1}(\boldsymbol{W}_1^*)-\|\boldsymbol{W}_1-\boldsymbol{W}_1^*\|_F\geq\rho\sqrt{c}-\frac{1}{2}\rho\sqrt{c}\geq\frac{1}{2}\rho\sqrt{c}, \tag{85}$$

where the second inequality follows from (61). Using (81) again, we obtain

$$
\begin{aligned}
& 2K \max\left\{1, \frac{c}{\gamma}\right\} \frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_W}} \left(\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right) \\
\geq & \left\| \left(\boldsymbol{W}_1\boldsymbol{W}_1^T + nK\lambda_H \boldsymbol{I}\right) \left(\boldsymbol{\Theta}_1 - \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}_1 + \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}_1\right) - \boldsymbol{W}_1 \right\|_F \\
\geq & \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}} \left\| \boldsymbol{W}_1 \left(\boldsymbol{W}_1^T\boldsymbol{W}_1 + nK\lambda_H\boldsymbol{I} - \frac{\sqrt{n\lambda_H}}{\sqrt{\lambda_W}}\boldsymbol{I}\right)\right\|_F - \left\| \left(\boldsymbol{W}_1\boldsymbol{W}_1^T + nK\lambda_H \boldsymbol{I}\right)\left(\boldsymbol{\Theta}_1 - \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}_1\right)\right\|_F \\
\geq & \frac{\sqrt{c}}{2\rho} \left\| \boldsymbol{W}_1^T\boldsymbol{W}_1 - \frac{c\sqrt{n\lambda_H}}{\sqrt{\lambda_W}}\boldsymbol{P}\right\|_F - \left(\|\boldsymbol{W}_1\|^2 + nK\lambda_H\right) \left\|\boldsymbol{\Theta}_1 - \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}_1\right\|_F
\end{aligned}
$$

where the last inequality follows from $c := 1 - K\sqrt{n\lambda_W\lambda_H}$, and $\left\|\boldsymbol{W}_1\left(\boldsymbol{W}_1^T\boldsymbol{W}_1 - c\sqrt{n\lambda_H}\boldsymbol{I}/\sqrt{\lambda_W}\right)\right\|_F = \left\|\boldsymbol{W}_1\left(\boldsymbol{W}_1^T\boldsymbol{W}_1 - c\sqrt{n\lambda_H}\boldsymbol{P}/\sqrt{\lambda_W}\right)\right\|_F \geq \sigma_{K-1}(\boldsymbol{W}_1)\left\|\boldsymbol{W}_1^T\boldsymbol{W}_1 - c\sqrt{n\lambda_H}\boldsymbol{P}/\sqrt{\lambda_W}\right\|_F$ due to $\boldsymbol{W} = \boldsymbol{W}_1\boldsymbol{P}$ and (85). This, together with (79) and (84), yields

$$
\left\| \boldsymbol{W}_1^T\boldsymbol{W}_1 - \frac{c\sqrt{n\lambda_H}}{\sqrt{\lambda_W}}\boldsymbol{P}\right\|_F \leq \kappa_3 \left(\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right), \tag{86}
$$

where

$$
\kappa_3 := \frac{2\rho}{\sqrt{c}} \left(2K\max\left\{1, \frac{c}{\gamma}\right\}\frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_W}} + \left(4\rho^2 c + nK\lambda_H\right)\kappa_2\right).
$$

Let $\boldsymbol{W}_1 = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T$ be the thin singular value decomposition of $\boldsymbol{W}_1$, where $\boldsymbol{U}_1 = [\bar{\boldsymbol{U}}_1\ \boldsymbol{u}_1] \in \mathcal{O}^{d\times K}$ with $\bar{\boldsymbol{U}}_1 \in \mathbb{R}^{d\times(K-1)}$ and $\boldsymbol{u}_1 \in \mathbb{R}^d$, $\boldsymbol{V}_1 = [\bar{\boldsymbol{V}}_1\ \boldsymbol{v}_1] \in \mathcal{O}^K$ with $\bar{\boldsymbol{V}}_1 \in \mathbb{R}^{K\times(K-1)}$ and $\boldsymbol{v}_1 \in \mathbb{R}^K$, and $\boldsymbol{\Sigma}_1 = \mathrm{diag}(\sigma_1,\dots,\sigma_K)$ is a diagonal matrix with $\sigma_1 \geq \cdots \geq \sigma_K \geq 0$. It follows from (85) that $\boldsymbol{W}_1 = \boldsymbol{W}\boldsymbol{P}$ is of rank $K-1$, which implies $\sigma_K = 0$. Therefore, we have $\boldsymbol{W}_1 = \bar{\boldsymbol{U}}_1\mathrm{diag}(\sigma_1,\dots,\sigma_{K-1})\bar{\boldsymbol{V}}_1^T$. This, together with $\boldsymbol{W}_1\boldsymbol{1}_K = \boldsymbol{W}\boldsymbol{P}\boldsymbol{1}_K = \boldsymbol{0}$, yields $\bar{\boldsymbol{V}}_1^T\boldsymbol{1}_K = \boldsymbol{0}$. This directly implies $\boldsymbol{v}_1 = \boldsymbol{1}_K/\sqrt{K}$. Using this and $\boldsymbol{V}_1\boldsymbol{V}_1^T = \boldsymbol{I}$, we obtain

$$
\bar{\boldsymbol{V}}_1\bar{\boldsymbol{V}}_1^T = \boldsymbol{I} - \boldsymbol{v}_1\boldsymbol{v}_1^T = \boldsymbol{P}. \tag{87}
$$

Let $\boldsymbol{\Sigma}_2 = \begin{bmatrix} \boldsymbol{I}_{K-1} & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{bmatrix}$. Noting that $\rho^2 = \sqrt{n\lambda_H}/\sqrt{\lambda_W}$, we compute

$$
\begin{aligned}
\left\|\boldsymbol{W}_1^T\boldsymbol{W}_1 - \rho^2 c\boldsymbol{P}\right\|_F &= \left\|\boldsymbol{V}_1\left(\boldsymbol{\Sigma}_1^2 - \rho^2 c\boldsymbol{\Sigma}_2\right)\boldsymbol{V}_1^T\right\|_F = \left\|\boldsymbol{\Sigma}_1^2 - \rho^2 c\boldsymbol{\Sigma}_2\right\|_F \\
&= \left\|(\boldsymbol{\Sigma}_1 - \rho\sqrt{c}\boldsymbol{\Sigma}_2)(\boldsymbol{\Sigma}_1 + \rho\sqrt{c}\boldsymbol{\Sigma}_2)\right\|_F \geq \rho\sqrt{c}\left\|\boldsymbol{\Sigma}_1 - \rho\sqrt{c}\boldsymbol{\Sigma}_2\right\| \\
&= \rho\sqrt{c}\left\|\boldsymbol{U}_1\left(\boldsymbol{\Sigma}_1 - \rho\sqrt{c}\boldsymbol{\Sigma}_2\right)\boldsymbol{V}_1^T\right\|_F \\
&= \rho\sqrt{c}\left\|\boldsymbol{W}_1 - \rho\sqrt{c}\bar{\boldsymbol{U}}_1\bar{\boldsymbol{V}}_1^T\right\|_F \geq \rho\sqrt{c}\min_{\boldsymbol{U}\in\mathcal{O}^{d\times(K-1)}}\left\|\boldsymbol{W}_1 - \rho\sqrt{c}\boldsymbol{U}\bar{\boldsymbol{V}}^T\right\|_F,
\end{aligned}
$$

where the last inequality uses the fact that there exists a $\boldsymbol{Q} \in \mathcal{O}^{K-1}$ such that $\bar{\boldsymbol{V}}_1 = \bar{\boldsymbol{V}}\boldsymbol{Q}$. Substituting this back into (86) yields

$$
\min_{\boldsymbol{U}\in\mathcal{O}^{d\times(K-1)}}\left\|\boldsymbol{W}_1 - \rho\sqrt{c}\boldsymbol{U}\bar{\boldsymbol{V}}^T\right\|_F \leq \frac{\kappa_3}{\rho\sqrt{c}}\left(\|\nabla_{\boldsymbol{W}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F + \|\nabla_{\boldsymbol{\Theta}} h(\boldsymbol{W},\boldsymbol{\Theta})\|_F\right). \tag{88}
$$

This, together with (72), (78), and (84), yields

$$
\mathrm{dist}^2\left((\boldsymbol{W},\boldsymbol{\Theta}),\mathcal{X}_h\right) \leq \left(\frac{2K^2\lambda_{\max}}{\gamma^2\lambda_{\min}} + 2\kappa_2^2 + \left(1 + \frac{2\lambda_W}{n\lambda_H}\right)\frac{\kappa_3^2}{\rho^2 c}\right)\|\nabla h(\boldsymbol{W},\boldsymbol{\Theta})\|_F^2.
$$

Then, we prove the claim. □

4 Cross-Entropy Loss

Suppose that $\mathcal{L}(\cdot,\cdot)$ is the cross-entropy (CE) loss:

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{y}_k) = -\log\left(\frac{\exp(z_k)}{\sum_{\ell=1}^{K}\exp(z_\ell)}\right). \tag{89}$$

We consider the following problems:

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times K}, \boldsymbol{H}\in\mathbb{R}^{d\times N}} F(\boldsymbol{W},\boldsymbol{H}) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{W}^T\boldsymbol{h}_{k,i}, \boldsymbol{y}_k) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2}\|\boldsymbol{H}\|_F^2, \tag{90}$$

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times K}, \boldsymbol{\Theta}\in\mathbb{R}^{d\times K}} f(\boldsymbol{W},\boldsymbol{\Theta}) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}(\boldsymbol{W}^T\boldsymbol{\theta}_k, \boldsymbol{y}_k) + \frac{\lambda_W}{2}\|\boldsymbol{W}\|_F^2 + \frac{n\lambda_H}{2}\|\boldsymbol{\Theta}\|_F^2, \tag{91}$$

where λ_W, λ_H are the penalties for $\boldsymbol{W}$ and $\boldsymbol{H}$, respectively. One can easily verify

$$F(\boldsymbol{W},\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{W},\boldsymbol{H}_i). \tag{92}$$

Lemma 3. *The optimal solution set of Problem* (91) *takes the form of*

$$\mathcal{X}_f = \left\{(\boldsymbol{W},\boldsymbol{\Theta}) : \boldsymbol{W} = \frac{\sqrt[4]{n\lambda_H}}{\sqrt[4]{\lambda_W}}\left(\max\{c,0\}\right)^{\frac{1}{2}}\boldsymbol{U}, \boldsymbol{\Theta} = \frac{\sqrt{\lambda_W}}{\sqrt{n\lambda_H}}\boldsymbol{W}, \boldsymbol{U}\in\mathcal{O}^{d\times K}\right\}, \tag{93}$$

where $c := 1 - K\sqrt{n\lambda_W\lambda_H}$.

Proof. $\square$

References

[1] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[2] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009.

[3] Z. Zhou and A. M.-C. So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.