Appendix of "Provable Knowledge Transfer using Successor Feature for Deep Reinforcement Learning."

Before moving into the technical details, we provide an overview of the structure of the appendix.

In Appendix A, we define some notations and useful lemmas to simplify the presentation and analysis. Some important notations for understanding the proof is summarized in Table 3.

In Appendix B, we provide some preliminary lemmas and proof for Theorem 1. A proof sketch is included as (**i**) characterization of the local convex region of the objective function in (12) and (9) (Lemma 6), (**ii**) Characterization of the difference between the empirical gradient in (15) and the gradient of the objective function (Lemma 7), (**iii**) Characterization of the relation of two consecutive iterations $\Theta^{(t+1)}$ and $\Theta^{(t)}$ in (62), and (**iv**) Mathematical induction over $(t+1)\cdot\|\Theta^{(t)}-\Theta^\star\|_2$ from $t=1$ to T to obtain the error bound between the learned model weights $\Theta^{(T)}$ and the optimal $\Theta^\star$.

In Appendix C, we provide the proof for Theorems 3 and 4. A proof sketch is included as follows: (1) Characterization of (26) by assuming knowledge of the optimal Q-function for previous tasks. (2) Characterization of the accumulated error resulting from the estimation error of the learned Q-function in previous tasks. (3) Combining the bounds from (1) and (2) leads to the error bound between (26) derived from the estimated Q-function of previous tasks and the optimal Q-function for the new tasks.

In Appendix D, we provide the proof for Theorem 2. The proof sketch is a direct application of the existing results of the convergence analysis as shown in Appendix B and the error bound between (26) derived from the estimated Q-function of previous tasks and the optimal Q-function for the new tasks as shown in Appendix C.

In Appendix E, we provide additional experiments to further support the proposed SF-DQN in Algorithm 1 and our theoretical findings.

In Appendix F, we provide the proofs for the preliminary lemmas in proving Theorems 1 and 2.

In Appendix G, we provide the proofs for the preliminary lemmas in proving Theorems 3 and 4.

In Appendix H, we provide the proof for some additional lemmas.

A Notations and preliminary results

Population risk function. We define a population risk function as

$$f_{\pi^\star}(\theta) := \mathbb{E}_{(\boldsymbol{s},a)\sim\pi^\star}\big\|\psi(\theta;\boldsymbol{s},a) - \mathbb{E}_{\boldsymbol{s}'|(\boldsymbol{s},a),a'\sim\pi^\star(\boldsymbol{s}')}\big(\phi(\boldsymbol{s},a,\boldsymbol{s}')+\gamma\cdot\psi(\theta^\star;\boldsymbol{s}',a')\big)\big\|_2^2. \tag{30}$$

We can see that $\theta^\star$ is the global minimal to (30) with Assumption 1. For the convenience of presentation, we simplify $f_{\pi^\star}$ as f in the supplementary materials.

Then, the gradient of (30) is

$$\begin{aligned}&\nabla f_{\pi^\star}(\theta)\\ &= \mathbb{E}_{(\boldsymbol{s},a)\sim\pi^\star,\boldsymbol{s}'|(\boldsymbol{s},a)\sim\mathcal{P},a'\sim\pi^\star}\big(\psi(\theta;\boldsymbol{s},\boldsymbol{a})-\phi(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')-\gamma\cdot\psi(\theta^\star;\boldsymbol{s}',a')\big)\cdot\nabla\psi(\theta;\boldsymbol{s},a).\end{aligned} \tag{31}$$

Given f is a smooth function, we have the gradient of f with respect to any θ_ℓ at the ground truth $\theta^\star$ equals to zero, namely,

$$\nabla_\ell f(\theta^\star) := \nabla_{\theta_\ell} f(\theta^\star) = \mathbf{0}, \qquad \forall \ell\in[L]. \tag{32}$$

Vectorized Gradient of θ and $\boldsymbol{w}$ at iteration t. To avoid unnecessary high-dimensional tensor analysis, the gradient with respect to θ, denoted as $\nabla_\theta H$ for some function H, is represented as its corresponding vectorized version, $\nabla_{\text{Vec}(\theta)}H$.

Let n denote the dimension of $\boldsymbol{W}$ defined in (1). We denote n_l as the dimension of the vectorized neuron weights in the ℓ-th layer, namely, $n_\ell = \dim(\text{vec}(\theta_\ell))$.

Then, the gradient in updating θ as

$$\begin{aligned} & g^{(t)}(\theta^{(t)}; \mathcal{D}_t) \\ = & \sum_{m \in \mathcal{D}_t} \Big(\psi(\theta^{(t)}; \boldsymbol{s}_m, a_m) - \phi(\boldsymbol{s}_m, a_m, \boldsymbol{s}'_m) - \gamma \cdot \psi(\theta^{(t)}; \boldsymbol{s}'_m, a'_m) \Big) \cdot \nabla_\theta \psi(\theta^{(t)}; \boldsymbol{s}_m, a_m) \end{aligned} \tag{33}$$

with $g^{(t)}(\theta^{(t)}; \mathcal{D}_t) \in \mathbb{R}^n$. Then, we have

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t \cdot g^{(t)}(\theta^{(t)}; \mathcal{D}_t). \tag{34}$$

Similary to (33), we define the gradient

$$l^{(t)}(\boldsymbol{w}^{(t)}; \mathcal{D}_t) = \sum_{m \in \mathcal{D}_t} \Big(\phi(\boldsymbol{s}_m, a_m, \boldsymbol{s}'_m)^\top \boldsymbol{w}^{(t)} - r(\boldsymbol{s}_m, a_m, \boldsymbol{s}'_m) \Big) \cdot \phi(\boldsymbol{s}_m, a_m, \boldsymbol{s}'_m). \tag{35}$$

In addition, without special descriptions, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^\top, \boldsymbol{\alpha}_2^\top, \cdots, \boldsymbol{\alpha}_K^\top]^\top$ stands for any unit vector that in $\mathbb{R}^{K_\ell K_{\ell-1}}$ with $\boldsymbol{\alpha}_j \in \mathbb{R}^{K_{\ell-1}}$ ($K_0 = d$). Therefore, we have

$$\begin{aligned} \|\nabla_\ell H\|_2 &= \max_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}^\top \nabla_\ell H\|_2 = \max_{\boldsymbol{\alpha}} \Big| \sum_{j=1}^K \boldsymbol{\alpha}_j^\top \frac{\partial H}{\partial \boldsymbol{w}_{\ell,j}} \Big|, \\ \|\nabla_\ell^2 H\|_2 &= \max_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}^\top \nabla_\ell^2 \, H \, \boldsymbol{\alpha}\|_2 = \max_{\boldsymbol{\alpha}} \Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^\top \frac{\partial H}{\partial \boldsymbol{w}_{\ell,j}} \Big)^2. \end{aligned} \tag{36}$$

Derivation of the gradient of deep neural networks. We use $h^{(\ell)}(\theta)$ to denote the input in the ℓ-th layer (or the output in the $(\ell-1)$-th layer) of deep neural network $\psi(\theta)$, and $h^{(1)} = \boldsymbol{x}(\boldsymbol{s}, a)$, where

$$\boldsymbol{h}^{(\ell)}(\theta; \boldsymbol{s}, a) = \sigma(\theta_{\ell-1}^\top \boldsymbol{h}^{(\ell-1)}) = \cdots = \sigma\Big(\theta_\ell^\top \sigma\big(\theta_{\ell-1} \cdots \sigma(\theta_1^\top \boldsymbol{x}(\boldsymbol{s}, a))\big)\Big). \tag{37}$$

Then, we denote the dimension of $\boldsymbol{h}^{(\ell)}$ as K_ℓ. Then, $\psi(\theta; \boldsymbol{s}, a)$ can be written as

$$\psi(\theta; \boldsymbol{s}, a) = \frac{\mathbf{1}^\top}{K_L} \sum_{k=1}^{K_L} \sigma(\theta_{L,k}^\top \boldsymbol{h}^{(L)}) = \frac{\mathbf{1}^\top}{K_L} \sigma\big(\theta_L^\top \sigma(\theta_{L-1}^\top \boldsymbol{h}^{(L-1)})\big), \tag{38}$$

where $\theta_{\ell,k}$ denotes the k-th neuron weights in the ℓ-th layer. Then, we define a group of functions $\mathcal{J}_\ell(\theta) \in \mathbb{R}^n \longrightarrow \mathbb{R}^K$ such that

$$\begin{aligned} & \mathcal{J}_\ell(\theta) \\ = & \begin{cases} \big[\mathbf{1}^\top \sigma'(\theta_L^\top \boldsymbol{h}^{(L)}) \theta_L^\top \cdot \sigma'(\theta_{L-1}^\top \boldsymbol{h}^{(L-1)}) \theta_{L-1}^\top \cdots \sigma'(\theta_{\ell+1}^\top \boldsymbol{h}^{(\ell+1)}) \theta_{\ell+1}^\top\big]^\top & \text{if} \quad \ell > 1 \\ \mathbf{1} \quad \text{if} \quad \ell = 1 \end{cases} . \end{aligned} \tag{39}$$

Then, the gradient of ψ can be represented as

$$\frac{\partial \psi}{\partial \theta_{\ell,k}}(\theta) = \frac{1}{K_\ell} \mathcal{J}_{\ell,k}(\theta) \sigma'\big(\theta_{\ell,k}^\top \boldsymbol{h}^{(\ell)}(\theta)\big) \boldsymbol{h}^{(\ell)}(\theta), \tag{40}$$

where $\mathcal{J}_{\ell,k}$ stands for the k-th component of $\mathcal{J}_\ell$.

Order-wise Analysis. Most constant numbers will be ignored in most steps. In particular, we use $h_1(z) \gtrsim$ (or $\lesssim, \eqsim$)$h_2(z)$ to denote there exists some positive constant C such that $h_1(z) \geq$ (or $\leq$, $=$)$C \cdot h_2(z)$ when $z \in \mathbb{R}$ is sufficiently large. In this paper, we consider the case where $\theta_\ell^\star$ is well-conditioned, such that its largest singular value $\Sigma_1(\ell)$ and the condition number $\Sigma_1(\ell)/\sigma_K(\ell)$ can be viewed as constants and will be hidden in the order-wise analysis.

A.1 Useful Lemmas for matrix concentration

Lemma 1 (Weyl's inequality, (Bhatia, 2013)). *Let $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{E}$ be a matrix with dimension $m \times m$. Let $\lambda_i(\boldsymbol{B})$ and $\lambda_i(\boldsymbol{A})$ be the i-th largest eigenvalues of $\boldsymbol{B}$ and $\boldsymbol{A}$, respectively. Then, we have*

$$|\lambda_i(\boldsymbol{B}) - \lambda_i(\boldsymbol{A})| \leq \|\boldsymbol{E}\|_2, \quad \forall \quad i \in [m]. \tag{41}$$

Table 3: Notations for the proofs

d	Dimension of the feature mappings of the state-action pair $(\boldsymbol{s}, a) \in \mathcal{S} \times \mathcal{A}$.
K	Number of neurons in the hidden layer.
L	Number of hidden layers.
T	Number of iterations.
$\boldsymbol{w}_i^{(t)}$	The estimated value for reward mapping of task i at t-th iteration.
$\Theta_i^{(t)}$	The estimated neuron weights for the successor feature of task i at t-th iteration.
$\theta^{(t)}$	The value of $\Theta_1^{(t)}$ to simplify the notation in the analyses without GPI.
$g^{(t)}(\theta^{(t)}; \mathcal{D}_t)$	The pseudo-gradient function defined in (33) at point $\theta^{(t)}$ with respect to the dataset $\mathcal{D}_t$.
$f_{\pi^\star}$ or f	The population risk function defined in (30).
$\nabla_\ell H(\hat{\theta})$	The gradient of a function H with respect to the components of θ_ℓ at point $\hat{\theta}$.
$\nabla_\ell^2 H(\hat{\theta})$	The Hessian matrix of a function H with respect to the components of θ_ℓ at point $\hat{\theta}$.
Q_i^π	The Q-function of task i for policy π.
$Q_i^\star$	The Q-function of task i for the optimal policy $\pi^\star$.
$q^\star$	A constant defined in (80), depending on task relevance $\|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2$.
η_t	The step size for updating neuron weights Θ_i for the successor feature.
κ_t	The step size for updating the parameter for the weight mapping.
c_N	A constant in the order of $1/\sqrt{N}$.
n	The dimension of θ.
n_ℓ	The dimension of vectorized θ_ℓ.
K_ℓ	The dimension of the input for the ℓ-th layer for the deep neural network. $K_0 = d$.
$\mathcal{J}_\ell(\boldsymbol{W})$	A function in $\mathbb{R}^n \longrightarrow \mathbb{R}^K$, defined in (39).
C_t	The distribution shift between the optimal policy and behavior policy at iteration t, defined in Assumption (3).
N	The size of the experience replay buffer.
$\phi_{\max}$	The upper bound of the transition feature.
ρ_1	A constant defined in (84).
ρ_2	The smallest eigenvalue of $\mathbb{E}\phi(\boldsymbol{s}, a)\phi(\boldsymbol{s}, a)^\top \in \mathbb{R}^{d\times d}$.
$\phi_{\max}$	The upper bound of the transition feature.

Lemma 2 ((Tropp, 2012), Theorem 1.6)**.** *Consider a finite sequence $\{\boldsymbol{Z}_k\}$ of independent, random matrices with dimensions $d_1 \times d_2$. Assume that such random matrix satisfies*

$$\mathbb{E}(\boldsymbol{Z}_k) = 0 \quad \textit{and} \quad \|\boldsymbol{Z}_k\| \le R \quad \textit{almost surely}.$$

Define

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

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

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

Lemma 3 (Lemma 5.2, (Vershynin, 2010)). *Let $\mathcal{B}(0,1) \in \{\boldsymbol{\alpha} \big| \|\boldsymbol{\alpha}\|_2 = 1, \boldsymbol{\alpha} \in \mathbb{R}^d\}$ denote a unit ball in $\mathbb{R}^d$. Then, a subset $\mathcal{S}_\xi$ is called a ξ-net of $\mathcal{B}(0,1)$ if every point $\boldsymbol{z} \in \mathcal{B}(0,1)$ can be approximated to within ξ by some point $\boldsymbol{\alpha} \in \mathcal{B}(0,1)$, i.e., $\|\boldsymbol{z} - \boldsymbol{\alpha}\|_2 \leq \xi$. Then the minimal cardinality of a ξ-net $\mathcal{S}_\xi$ satisfies*

$$|\mathcal{S}_\xi| \leq (1 + 2/\xi)^d. \tag{42}$$

Lemma 4 (Lemma 5.3, (Vershynin, 2010)). *Let $\boldsymbol{A}$ be an $d_1 \times d_2$ matrix, and let $\mathcal{S}_\xi(d)$ be a ξ-net of $\mathcal{B}(0,1)$ in $\mathbb{R}^d$ for some $\xi \in (0,1)$. Then*

$$\|\boldsymbol{A}\|_2 \leq (1-\xi)^{-1} \max_{\boldsymbol{\alpha}_1 \in \mathcal{S}_\xi(d_1), \boldsymbol{\alpha}_2 \in \mathcal{S}_\xi(d_2)} |\boldsymbol{\alpha}_1^T \boldsymbol{A} \boldsymbol{\alpha}_2|. \tag{43}$$

Lemma 5 (Mean Value Theorem). *Let $\boldsymbol{U} \subset \mathbb{R}^{n_1}$ be open and $\boldsymbol{f} : \boldsymbol{U} \longrightarrow \mathbb{R}^{n_2}$ be continuously differentiable, and $\boldsymbol{x} \in \boldsymbol{U}$, $\boldsymbol{h} \in \mathbb{R}^{n_1}$ vectors such that the line segment $\boldsymbol{x} + t\boldsymbol{h}$, $0 \leq t \leq 1$ remains in $\boldsymbol{U}$. Then we have:*

$$\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x}) = \left(\int_0^1 \nabla \boldsymbol{f}(\boldsymbol{x} + t\boldsymbol{h}) dt \right) \cdot \boldsymbol{h},$$

where $\nabla \boldsymbol{f}$ denotes the Jacobian matrix of $\boldsymbol{f}$.

A.2 Defintions of Sub-Gaussian and Sub-exponential.

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

$$(\mathbb{E}|X|^p)^{1/p} \leq c_1 \sqrt{p} \tag{44}$$

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

$$\mathbb{E}e^{s(X - \mathbb{E}X)} \leq e^{c_2 \|X\|_{\psi_2}^2 s^2} \tag{45}$$

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

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

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

$$(\mathbb{E}|X|^p)^{1/p} \leq c_3 p \tag{46}$$

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

$$\mathbb{E}e^{s(X - \mathbb{E}X)} \leq e^{c_4 \|X\|_{\psi_1}^2 s^2} \tag{47}$$

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

B Proof of Theorem 1

Lemma 6 (Local convexity of $f_{\pi^\star}$). *Given any $\theta \in \mathbb{R}^n$, let θ satisfy*

$$\|\theta - \theta^\star\|_2 \lesssim \frac{c_N \cdot \sigma_K}{\rho_1 \cdot K} \tag{48}$$

for some constant $c_N \in (0,1)$. Then, for the $f_{\pi^\star}$ defined in (30), we have

$$\frac{(1 - c_N)\rho_1}{K^2} \preceq \nabla_\ell^2 f_{\pi^\star}(\theta) \preceq \frac{7}{K}. \tag{49}$$

Lemma 7 (Upper bound of the error gradient)**.** *Let $f_{\pi^\star}$ be the function defined in (30). Let g_t be the function defined in (33). Then, with probability at least $1 - q^{-K_{\ell-1}}$, we have*

$$\begin{aligned}\Big\|\nabla_\ell f_{\pi^\star}(\theta) - g_\ell(\theta^{(t)}; \mathcal{D}_t)\Big\|_2 \lesssim & \frac{1}{K_\ell} \cdot \|\theta - \theta^\star\|_2 \cdot \sqrt{\frac{K_{\ell-1}\log q}{|\mathcal{D}_t|}} + \frac{\gamma}{K_\ell} \cdot \|\theta^{(t)} - \theta^\star\|_2 \\ & + \frac{R_{\max}}{1-\gamma} \cdot (1+\gamma)\tau^\star \cdot \eta_{t-\tau^\star} \\ & + |\mathcal{A}| \cdot \frac{R_{\max}}{1-\gamma} \cdot (1 + \log_\nu \lambda^{-1} + \frac{1}{1-\nu}) \cdot C_t,\end{aligned} \tag{50}$$

where $\tau^\star = \min\{t \mid \lambda\nu^t \le \eta_T\}$, and ν & λ are defined in Assumption 2.

Lemma 8 (Convergence of $\boldsymbol{w}^{(t)}$)**.** *With probability at least $1 - q^{-d}$, $\boldsymbol{w}$ enjoys a linear convergence rate to $\boldsymbol{w}^\star$ as*

$$\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^\star\|_2 \le \Big(1 - \frac{\rho - c_N}{\phi_{\max}}\Big) \cdot \|\boldsymbol{w}^{(t)} - \boldsymbol{w}^\star\|_2. \tag{51}$$

Proof of Theorem 1. From Algorithm 1, the update of θ can be written as

$$\begin{aligned}\theta^{(t+1)} =& \theta^{(t)} - \eta_t \cdot g^{(t)}(\theta^{(t)}; \mathcal{D}_t) \\ =& \theta^{(t)} - \eta_t \cdot \nabla f(\theta^{(t)}) + \eta_t \cdot \big(\nabla f(\theta^{(t)}) - g^{(t)}(\theta^{(t)}; \mathcal{D}_t)\big).\end{aligned} \tag{52}$$

Since ∇f is a smooth function and θ^* is a local (global) optimal to f, then we have

$$\begin{aligned}\nabla f(\theta^{(t)}) =& \nabla f(\theta^{(t)}) - \nabla f(\theta^\star) \\ =& \int_0^1 \nabla^2 f\Big(\theta^{(t)} + u \cdot (\theta^{(t)} - \theta^\star)\Big) du \cdot (\theta^{(t)} - \theta^\star),\end{aligned} \tag{53}$$

where the last equality comes from Mean Value Theory in Lemma 5. For notational convenience, we use $\boldsymbol{A}^{(t)}$ to denote the integration as

$$\boldsymbol{A}^{(t)} := \int_0^1 \nabla^2 f\Big(\theta^{(t)} + u \cdot (\theta^{(t)} - \theta^\star)\Big) du. \tag{54}$$

Then, we have

$$\begin{aligned}\|\theta^{(t+1)} - \theta^\star\|_2 \le & \|\boldsymbol{I} - \eta_t \boldsymbol{A}^{(t)}\|_2 \cdot \|\theta^{(t)} - \theta^\star\|_2 + \eta_t \cdot \|\nabla f(\theta^{(t)}) - g^{(t)}(\theta^{(t)}; \mathcal{D}_t)\|_2 \\ \le & \|\boldsymbol{I} - \eta_t \boldsymbol{A}^{(t)}\|_2 \cdot \|\theta^{(t)} - \theta^\star\|_2 + \eta_t \cdot \sum_{\ell=1}^{L} \Big\|\nabla_\ell f(\theta^{(t)}) - g^{(t)}(\theta_\ell^{(t)}; \mathcal{D}_t)\Big\|_2.\end{aligned} \tag{55}$$

From Lemma 6, we have

$$\|\boldsymbol{I} - \eta_t \boldsymbol{A}^{(t)}\|_2 \le 1 - \eta_t \cdot \frac{(1 - c_N) \cdot \rho_1}{K^2}. \tag{56}$$

From Lemma 7, we have

$$\begin{aligned}\Big\|\nabla_\ell f_{\pi^\star}(\theta^{(t)}) - g_\ell(\theta^{(t)}; \mathcal{D}_t)\Big\|_2 \lesssim & \frac{1}{K_\ell} \cdot \|\theta^{(t)} - \theta^\star\|_2 \cdot \sqrt{\frac{K_{\ell-1}\log q}{|\mathcal{D}_t|}} + \frac{\gamma}{K_\ell} \cdot \|\theta^{(t)} - \theta^\star\|_2 \\ & + \frac{R_{\max}}{1-\gamma} \cdot (1+\gamma)\tau^\star \cdot \eta_{t-\tau^\star} \\ & + |\mathcal{A}| \cdot \frac{R_{\max}}{1-\gamma} \cdot (1 + \log_\nu \lambda^{-1} + \frac{1}{1-\nu}) \cdot C_t.\end{aligned} \tag{57}$$

With Assumption 3, we have

$$C_t \le C \cdot \big(\|\theta^{(t)} - \theta^\star\|_2 + \|\boldsymbol{w}^{(t)} - \boldsymbol{w}^\star\|_2\big).$$

When we have a sufficiently large number of samples at iteration t as

$$|\mathcal{D}_t| \gtrsim c_N^{-2} \cdot \rho_1^{-1} \cdot \big(\sum_{\ell=1}^{L} K_\ell \sqrt{K_{\ell-1}}\big)^2 \cdot \log q, \tag{58}$$

(55) can be simplified as

$$\|\theta^{(t+1)} - \theta^\star\|_2 \leq (1 - \eta_t \cdot \xi) \cdot \|\theta^{(t)} - \theta^\star\|_2 + \eta_t \cdot \Delta_t + \eta_t \cdot C^\star \|\boldsymbol{w}^{(t)} - \boldsymbol{w}^\star\|_2. \tag{59}$$

where

$$\begin{aligned} C^\star &= |\mathcal{A}| \cdot \frac{R_{\max}}{1-\gamma} \cdot (1 + \log_\nu \lambda^{-1} + \frac{1}{1-\nu}) \cdot C \\ \xi &= \frac{(1-\gamma-c_N)\rho_1}{K^2} - C^\star \\ \Delta_t &= \frac{R_{\max}}{1-\gamma} \cdot (1+\gamma)\tau^\star \cdot \eta_{t-\tau^\star}. \end{aligned} \tag{60}$$

Let $\eta_t = \frac{1}{\xi \cdot (t+1)}$, we have

$$(t+1) \cdot \|\theta^{(t+1)} - \theta^\star\|_2 \leq t \cdot \|\theta^{(t)} - \theta^\star\|_2 + \xi^{-1} \cdot \Delta_t + \xi^{-1} \cdot C^\star \|\boldsymbol{w}^{(t)} - \boldsymbol{w}^\star\|_2. \tag{61}$$

Next, we have

$$\begin{aligned} &\sum_{t=0}^{T-1} (t+1) \cdot \|\theta^{(t+1)} - \theta^\star\|_2 - t \cdot \|\theta^{(t)} - \theta^\star\|_2 \\ \leq &\sum_{t=0}^{T-1} \xi^{-1} \cdot (\Delta_t + C^\star \|\boldsymbol{w}^{(t)} - \boldsymbol{w}^\star\|_2). \end{aligned} \tag{62}$$

With the definition of Δ_t in (60), we have

$$\begin{aligned} \sum_{t=0}^{T-1} \Delta_t \leq &\sum_{t=0}^{\tau^\star} \Delta_t + \sum_{t=\tau^\star}^{T-1} \lambda^{-1} \cdot \Delta_t \\ \leq &\sum_{t=0}^{\tau^\star} \tau^\star \cdot \frac{R_{\max}}{1-\gamma} + \sum_{t=\tau^\star}^{T-1} \cdot \frac{R_{\max} \cdot (1+\gamma)}{1-\gamma} \cdot \tau^\star \cdot \frac{1}{T - \tau^\star + 1} \\ \lesssim &\frac{R_{\max} \cdot \log^2 T}{1-\gamma} + \frac{R_{\max} \cdot (1+\gamma) \cdot \log^2 T}{1-\gamma}. \end{aligned} \tag{63}$$

With Lemma 8 that $\boldsymbol{w}$ enjoys a geometric decay, we have

$$\sum_{t=0}^{T-1} \|\boldsymbol{w}^{(t)} - \boldsymbol{w}^\star\|_2 \lesssim \|\boldsymbol{w}^{(0)} - \boldsymbol{w}^\star\|_2. \tag{64}$$

By multiplying $1/T$ on both sides of (62), we have

$$\|\theta^{(T)} - \theta^\star\|_2 \leq \frac{(2+\gamma) \cdot R_{\max} \cdot \log^2 T + C^\star \|\boldsymbol{w}^{(0)} - \boldsymbol{w}^\star\|_2}{(1-\gamma-c_N)\rho_1 K^{-2} - C^\star} \cdot \frac{1}{T}. \tag{65}$$

□

C PROOFS OF THEOREMS 3 AND 4

Lemma 9. *Let $Q_i^\star$ be the Q-function for the optimal policy of task i, we have*

$$|Q_i^\star - Q_j^\star| \leq \frac{1+\gamma}{1-\gamma} \phi_{\max} \cdot \|\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star\|_2. \tag{66}$$

Proof of Theorem 3. For any task $j \in [n]$, we have

$$\begin{aligned} Q_{n+1}^{\pi_{n+1}}(\boldsymbol{s},a) - Q_{n+1}^{\pi_j}(\boldsymbol{s},a) &= \max_{i\in[n]} Q_{n+1}^{\pi_i^\star}(\boldsymbol{s},a) - Q_{n+1}^{\pi_j}(\boldsymbol{s},a) \\ &\geq Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a) - Q_{n+1}^{\pi_j}(\boldsymbol{s},a) \\ &= \big(\psi_j(\Theta_j^\star) - \psi_j(\Theta_j^{(T)})\big)\cdot \boldsymbol{w}_{n+1}^\star. \end{aligned} \tag{67}$$

According to Theorem 1, we have

$$\|\psi_j(\Theta_j^\star) - \psi_j(\Theta_j^{(T)})\|_2 \leq \frac{(2+\gamma)\cdot R_{\max}\cdot \log^2 T + C^\star \|\boldsymbol{w}_j^{(0)} - \boldsymbol{w}_j^\star\|_2}{(1-\gamma-c_N)\rho_1 K^{-2} - C^\star}\cdot\frac{1}{T} := \frac{C_3}{T} \tag{68}$$

Then, we have

$$\mathcal{T}^\pi Q_{n+1}^{\pi_j}(\boldsymbol{s},a) \geq Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma\cdot\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}. \tag{69}$$

Therefore, with the contraction property of the Bellman operator $\mathcal{T}^\pi$, we have

$$\begin{aligned} Q_{n+1}^{\pi_{n+1}}(\boldsymbol{s},a) =& \lim_{k\to\infty}(\mathcal{T}^\pi)^k Q_{n+1}^{\pi_j}(\boldsymbol{s},a) \\ \geq& \lim_{k\to\infty}(\mathcal{T}^\pi)^{k-1}\Big(Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}\Big) \\ =& \lim_{k\to\infty}(\mathcal{T}^\pi)^{k-2}\cdot\mathcal{T}^\pi\Big(Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}\Big) \\ =& \lim_{k\to\infty}(\mathcal{T}^\pi)^{k-2}\cdot\Big(\mathcal{T}^\pi Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma^2\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}\Big) \\ =& \lim_{k\to\infty}(\mathcal{T}^\pi)^{k-2}\Big(Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} - \gamma^2\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}\Big) \\ =& Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \sum_{k=1}^{\infty}\gamma^k\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} \\ =& Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \frac{\gamma}{1-\gamma}\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} \\ \geq& Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a) - \frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} - \frac{\gamma}{1-\gamma}\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} \\ =& Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a) - \frac{1}{1-\gamma}\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} \end{aligned} \tag{70}$$

For any policy $\pi_j^\star$ with $j \in [n]$, we have

$$\begin{aligned} & Q_{n+1}^\star(\boldsymbol{s},a) - Q_{n+1}^{\pi_{n+1}}(\boldsymbol{s},a) \\ =& \big(Q_{n+1}^\star(\boldsymbol{s},a) - Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a)\big) + \big(Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a) - Q_{n+1}^{\pi_j}(\boldsymbol{s},a)\big) \\ \leq& \frac{2\gamma}{1-\gamma}\cdot\max_{\boldsymbol{s},a}|r_{n+1}(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)| + \frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{(1-\gamma)T} \\ \leq& \frac{2\gamma\cdot\phi_{\max}}{1-\gamma}\|\boldsymbol{w}_{n+1} - \boldsymbol{w}_j\|_2 + \frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{(1-\gamma)T}. \end{aligned} \tag{71}$$

Since (71) holds for any j, we have

$$|Q_{n+1}^\star(\boldsymbol{s},a) - Q_{n+1}^{\pi_{n+1}}(\boldsymbol{s},a)| \leq \frac{2\gamma\cdot\phi_{\max}}{1-\gamma}\min_{j\in[n]}\|\boldsymbol{w}_{n+1} - \boldsymbol{w}_j\|_2 + \frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{(1-\gamma)T}. \tag{72}$$

From Lemma 9, we know that

$$Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a) - Q_{n+1}^{\pi_j}(\boldsymbol{s},a) \leq \frac{1+\gamma}{1-\gamma}\cdot\max_{\boldsymbol{s},a}|r_{n+1}(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)|. \tag{73}$$

Similar to (70) to (72), we have

$$|Q^{\star}_{n+1}(\boldsymbol{s},a)-Q^{\pi_{n+1}}_{n+1}(\boldsymbol{s},a)|\leq\frac{(1+\gamma)\cdot\phi_{\max}}{1-\gamma}\min_{j\in[n]}\|\boldsymbol{w}_{n+1}-\boldsymbol{w}_j\|_2+\frac{C_3\|\boldsymbol{w}^{\star}_{n+1}\|_2}{(1-\gamma)T}. \tag{74}$$

□

Proof of Theorem 4. Let π'_{n+1} be generalized policy with DQN via GPI. Similar to (67), we have

$$\begin{aligned}
&Q^{\pi'_{n+1}}_{n+1}(\boldsymbol{s},a)-Q^{\pi'_j}_{n+1}(\boldsymbol{s},a)\\
=&\max_{i\in[n]}Q^{\pi^{\star}_i}_{n+1}(\boldsymbol{s},a)-Q^{\pi'_j}_{n+1}(\boldsymbol{s},a)\\
\geq& Q^{\pi^{\star}_j}_{n+1}(\boldsymbol{s},a)-Q^{\pi'_j}_{n+1}(\boldsymbol{s},a)\\
=&\psi_j(\Theta^{\star}_j)\boldsymbol{w}^{\star}_{n+1}-\psi_j(\Theta^{(T)}_j)\boldsymbol{w}^{(t)}_j\\
\approx&\psi_j(\Theta^{\star}_j)\boldsymbol{w}^{\star}_{n+1}-\psi_j(\Theta^{(T)}_j)\boldsymbol{w}^{\star}_j\\
=&\psi_j(\Theta^{\star}_j)\boldsymbol{w}^{\star}_{n+1}-\psi_j(\Theta^{(T)}_j)\boldsymbol{w}^{\star}_{n+1}+\psi_j(\Theta^{(T)}_j)\boldsymbol{w}^{\star}_{n+1}-\psi_j(\Theta^{(T)}_j)\boldsymbol{w}^{\star}_j\\
\geq&-\|\Theta^{\star}_j-\Theta^{(T)}_j\|\cdot\|\boldsymbol{w}^{\star}_{n+1}\|_2-\frac{1}{1-\gamma}\phi_{\max}\cdot\|\boldsymbol{w}^{\star}_{n+1}-\boldsymbol{w}^{\star}_j\|_2.
\end{aligned} \tag{75}$$

Following similar steps in the proof of Theorem 3, we have

$$\begin{aligned}
|Q^{\star}_{n+1}(\boldsymbol{s},a)-Q^{\pi'_{n+1}}_{n+1}(\boldsymbol{s},a)|\leq&\frac{(1+\gamma)\cdot\phi_{\max}}{1-\gamma}\min_{j\in[n]}\|\boldsymbol{w}_{n+1}-\boldsymbol{w}_j\|_2+\frac{C_3\|\boldsymbol{w}^{\star}_{n+1}\|_2}{(1-\gamma)T}\\
&+\frac{1}{1-\gamma}\phi_{\max}\cdot\min_{j\in[n]}\|\boldsymbol{w}^{\star}_{n+1}-\boldsymbol{w}^{\star}_j\|_2\\
\leq&\frac{2\cdot\phi_{\max}}{1-\gamma}\min_{j\in[n]}\|\boldsymbol{w}_{n+1}-\boldsymbol{w}_j\|_2+\frac{C_3\|\boldsymbol{w}^{\star}_{n+1}\|_2}{(1-\gamma)T}.
\end{aligned} \tag{76}$$

□

D PROOF OF THEOREM 2

Proof of Theorem 2. For task i, let π_j be the policy derived from $\psi_j(\Theta^{(T)}_j)\boldsymbol{w}^{\star}_i$ with $1\leq j\leq i$, where $\Theta^{(T)}_j$ is the returned neuron weights for the successor feature of task j.

Similar to (74), we have

$$Q^{\star}_i(\boldsymbol{s},a)-Q^{\pi_j}_i(\boldsymbol{s},a)\leq\frac{(1+\gamma)\cdot\phi_{\max}}{1-\gamma}\|\boldsymbol{w}_j-\boldsymbol{w}_i\|_2+\frac{C_3\|\boldsymbol{w}^{\star}_i\|_2}{(1-\gamma)T}. \tag{77}$$

Let π' be the policy derived from $\psi_i(\Theta^{(t)}_i)\boldsymbol{w}^{\star}_i$ at iteration t for task i, we have

$$Q^{\star}_i(\boldsymbol{s},a)-Q^{\pi'}_i\leq\|\Theta^{(t)}_i-\Theta^{\star}_i\|_2\cdot\|\boldsymbol{w}^{\star}_i\|_2. \tag{78}$$

Therefore, at iteration t for task i, we have

$$\begin{aligned}
C_t=&|Q^{\star}_i(\boldsymbol{s},a)-Q^{\pi^{(t)}_i}_i|\\
\leq&\min\left\{\frac{(1+\gamma)\cdot\phi_{\max}}{1-\gamma}\min_{1\leq j\leq i}\|\boldsymbol{w}_j-\boldsymbol{w}_i\|_2+\frac{C_3\|\boldsymbol{w}^{\star}_i\|_2}{(1-\gamma)T},\|\Theta^{(t)}_i-\Theta^{\star}_i\|_2\cdot\|\boldsymbol{w}^{\star}_i\|_2\right\}\\
\lesssim&\min\left\{\frac{(1+\gamma)\cdot\phi_{\max}}{1-\gamma}\min_{1\leq j\leq i}\|\boldsymbol{w}_j-\boldsymbol{w}_i\|_2,\|\Theta^{(t)}_i-\Theta^{\star}_i\|_2\cdot\|\boldsymbol{w}^{\star}_i\|_2\right\}\text{ (As } T\text{ is sufficiently large)}\\
=&\min\{q_t,1\}\cdot\|\Theta^{(t)}_i-\Theta^{\star}_i\|_2,
\end{aligned} \tag{79}$$

where

$$q_t = \frac{(1+\gamma)R_{\max}}{1-\gamma} \cdot \frac{\min_{1\leq i\leq j-1} \|\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star\|_2}{\|\Theta_j^{(t)} - \Theta_j^\star\|_2} \tag{80}$$

Following similar steps in (59) in the proof of Theorem 1, with C_t satisfying (79), we have

$$\|\theta^{(T)} - \theta^\star\|_2 \leq \frac{1}{T}\sum_{t=1}^{T-1} \frac{(2+\gamma)\cdot R_{\max}\cdot \log^2 T + C^\star \|\boldsymbol{w}^{(0)} - \boldsymbol{w}^\star\|_2}{(1-\gamma-c_N)\rho_1 K^{-2} - \min\{1, q_t\}\cdot C^\star} \cdot \frac{1}{T}. \tag{81}$$

□

E ADDITIONAL NUMERICAL EXPERIMENTS

In this section we empirically validate the theoretical results obtained in the previous section, using synthetic and real-world RL benchmarks.

E.1 SYNTHETIC DATA SETTINGS

Here, we define an MDP that contains two tasks with shared state transition dynamics. The MDP consists of a state space with $|\mathcal{S}| = 10,000$, an action space with $|\mathcal{A}| = 4$. For the first task, its successor feature is parameterized by a deep neural network with the randomly generated neuron weights $\Theta_1^\star$, and $\boldsymbol{w}_1^\star$ are randomly generated as the corresponding reward mapping. We then generate ϕ based on (10) with $\psi(\Theta_1^\star)$. Since ϕ is shared across all tasks, for Task 2, we randomly generate the reward mapping $\boldsymbol{w}_2^\star$ and then calculate $\psi_2^\star$ accordingly.

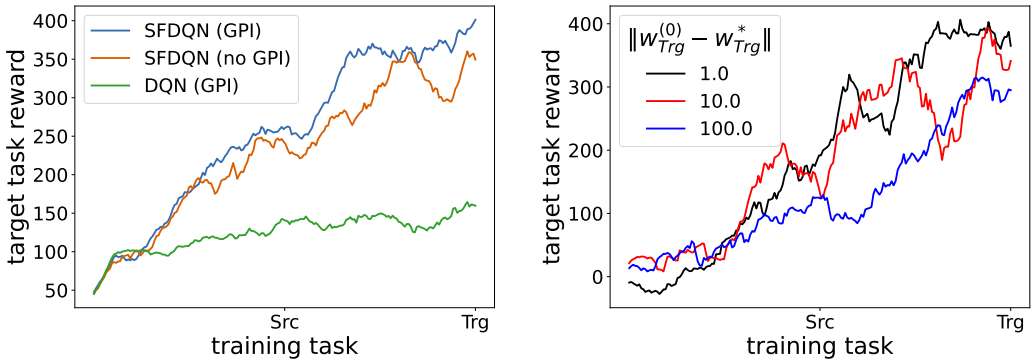
E.2 REAL DATA: REACHER ENVIRONMENT

The reacher environment is a robotic arm manipulation task consisting of a robotic arm with two joint torque controls. The state space is continuous, and the state features consist of angular displacement and angular velocity of the two joints. The actual action space for the robot arm is continuous that consists of the torques applied to the two joints, and is discretized for 3 values (for each joint torque). Thus, the total discretized action space consists of 9 actions ($|\mathcal{A}| = 9$). The discount factor used is $\gamma = 0.9$. Multiple tasks in this environment is defined by goal locations, and the objective of each task is to move the tip of the robotic arm towards the goal location.

The reward of each task is defined by the distance δ, measured from the tip of the robotic arm to the corresponding goal location. Specifically, a reward of $1 - \delta$ is given to the agent at each time step. There are 12 predefined tasks, and ϕ for a given state (common to all 12 tasks) is defined by stacking the reward for each of the 12 tasks for a given state as a vector. The corresponding reward weights $\boldsymbol{w}_i^\star$ for $i = 1, \ldots, 12$ are defined by one hot vectors, where the i^{th} element of $\boldsymbol{w}_i^\star$ is 1 and other elements are 0. Thus, the inner product $\phi^\top \boldsymbol{w}_i^\star$ naturally recovers the reward for the i^{th} task. For running experiments with this task, we use the open source code base `https://github.com/mike-gimelfarb/deep-successor-features-for-transfer.git`.

We first provide a comparison of the performance of SF-DQN with GPI, SF-DQN without GPI, and DQN with GPI, in Figure 3a. Here we consider the average transfer performance for four tasks, after training on a source task. It can be seen that SFDQN with GPI performs better compared to its no GPI counterpart. Both of these agents perform significantly better compared to DQN with GPI. hence, this result validates our theoretical results for the performance of these three methods.

Next, we investigate the performance of the SFDQN agent when the target task reward mappings are not known and learned simultaneously with successor features. We consider varying distances from the initial target task reward mapping to the true target task reward mapping. The results are shown in Figure 3b. It can be seen that when the reward mappings are initialized far away from the true reward mappings, the convergence of the SF-DQN agent is slower compared to that is initialized closer to the true reward mappings. This aligns with our convergence analysis for the SF-DQN agent with GPI.



(a) Comparison of DQN (GPI) and SF-DQN (with and without GPI)

(b) Comparison of different initializations for target task reward mapping

Figure 3: Experiments on Reacher environment

E.3 Additional experiments on synthetic RL benchmarks

Effect of $\|w_1^\star - w_2^\star\|$ on knowledge transfer. We investigate the effect of the distance between $\boldsymbol{w}_1^\star$ to $\boldsymbol{w}_2^\star$, on the transfer performance of the SFDQN. For this purpose, we assume SF-DQN agents have access to optimal reward mappings when training on Tasks 1 and 2. After obtaining ϕ as described earlier, we initialize and train Θ_2 using ϕ and $\boldsymbol{w}_2^\star$, with GPI. Reward defined by $\phi \cdot w_2^\star$ is used to obtain the average reward for Task 2. We repeat the process for different choices of $\boldsymbol{w}_2^\star$, and the results are shown in Figure 4. It can be seen that, when the task similarity is low (i.e. $\|w_1^\star - w_2^\star\|$ is large), the performance of the SF-DQN agent with GPI is poor. On the other hand, when the task similarity is high, the performance becomes significantly better.

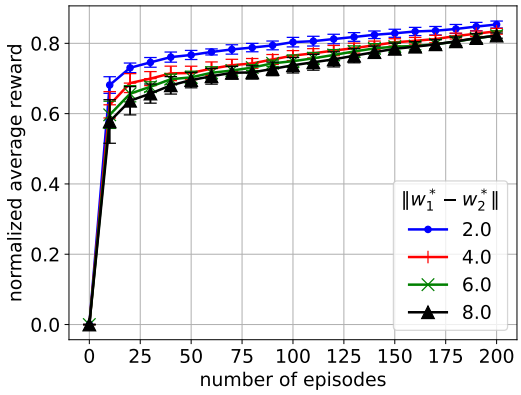


Figure 4: Effect $\|w_1^\star - w_2^\star\|$ on the convergence of SF-DQN agent when training on task 2 with GPI

F Proof of lemmas in Appendix B

F.1 Proof of Lemma 6

Lemma 6 provides the lower and upper bounds for the eigenvalues of the Hessian matrix of population risk function in (30). According to Weyl's inequality in Lemma 1, the eigenvalues of $\nabla_\ell^2 f(\cdot)$ at any fixed point θ can be bounded in the form of (86). Therefore, we first provide the lower and upper bounds for $\nabla_\ell^2 f$ at the desired ground truth $\theta^\star$. Then, the bounds for $\nabla_\ell^2 f$ at any other point θ is bounded through (30) by utilizing the conclusion in Lemma 10. Lemma 10 illustrates the distance between the Hessian matrix of f at θ and θ^*. Lemma 11 provides the lower bound of

$\mathbb{E}_{\boldsymbol{x}}\big(\sum_{j=1}^{K}\boldsymbol{\alpha}_j^\top\frac{\partial\psi}{\partial\theta_{\ell,k}}(\theta^\star)\big)^2$ when $\boldsymbol{x}$ belongs to sub-Gaussian distribution, which is used in proving the lower bound of the Hessian matrix in (87).

Lemma 10. *Let $f(\theta)$ be the population risk function defined in (30). If θ is close to $\theta^\star$ such that*

$$\|\theta-\theta^\star\|_2\lesssim\frac{\rho_1}{K} \tag{82}$$

we have

$$\|\nabla_\ell^2 f(\theta)-\nabla_\ell^2 f(\theta^\star)\|_2\lesssim\frac{1}{K}\cdot\|\theta-\theta^\star\|_2. \tag{83}$$

Lemma 11. *Suppose the following assumptions hold:*

1. *$\{\theta_j\}_{j=1}^K\in\mathbb{R}^{K_\ell}$ are linear independent,*
2. *Let $p(\boldsymbol{h}):\mathbb{R}^{K_\ell}\longrightarrow[\,0\ \ 1\,]$ be the probability density for $\boldsymbol{h}$ such that $\mathbb{E}_{\boldsymbol{h}}\|\boldsymbol{h}\|_2^2\le+\infty$.*

Let $\boldsymbol{\alpha}\in\mathbb{R}^{K_\ell K_{\ell-1}}$ be the unit vector defined in (36), we have

$$\rho_1:=\min_{\|\boldsymbol{\alpha}\|_2=1}\int_{\mathcal{R}}\Big(\sum_{j=1}^{K}\boldsymbol{\alpha}^\top\boldsymbol{h}\phi'(\theta_{\ell,j}^\top\boldsymbol{h})\Big)^2 p_H(\boldsymbol{h})\cdot d\boldsymbol{h}>0, \tag{84}$$

where $\mathcal{R}\subset\mathbb{R}^{K_\ell}$ with $\int_{\mathcal{R}}f_H(\boldsymbol{h})>0$. Moreover, if further assuming $\boldsymbol{h}$ belongs to Gaussian distribution, we have $\rho_1>0.091$.

Lemma 12. *Let $\boldsymbol{h}^{(\ell)}(\theta)$ be the function defined in (37). When θ is sufficiently close to $\theta^\star$, i.e., $\|\theta-\theta^\star\|_2$ is smaller than some positive constant $c<1$, we have*

$$\begin{aligned}\|\boldsymbol{h}^{(\ell)}(\theta)\|_2&\lesssim\|\boldsymbol{x}\|_2,\\ \|\boldsymbol{h}^{(\ell)}(\theta)-\boldsymbol{h}^{(\ell)}(\theta^\star)\|_2&\lesssim\|\theta-\theta^\star\|_2\cdot\|\boldsymbol{x}\|_2.\end{aligned} \tag{85}$$

Proof of Lemma 6. Let $\lambda_{\max}(\theta)$ and $\lambda_{\min}(\theta)$ denote the largest and smallest eigenvalues of $\nabla_\ell^2 f(\theta)$ at θ, respectively. Then, from Lemma 1, we have

$$\begin{aligned}\lambda_{\max}(\theta)&\le\lambda_{\max}(\theta^\star)+\|\nabla_\ell^2 f(\theta)-\nabla_\ell^2 f(\theta^\star)\|_2,\\ \lambda_{\min}(\theta)&\ge\lambda_{\min}(\theta^\star)-\|\nabla_\ell^2 f(\theta)-\nabla_\ell^2 f(\theta^\star)\|_2.\end{aligned} \tag{86}$$

Then, we provide the lower bound of the Hessian matrix of the population function at $\theta^\star$. Let $\mathcal{P}$ be the distribution for $\boldsymbol{h}^{(\ell)}(\theta)$ when $\boldsymbol{x}\sim\mu^\star$ with probability density function denoted as p_H. For any $\boldsymbol{\alpha}\in\mathbb{R}^{K_\ell K}$ with $\|\boldsymbol{\alpha}\|_2=1$, we have

$$\begin{aligned}&\min_{\|\boldsymbol{\alpha}\|_2=1}\boldsymbol{\alpha}^\top\nabla_\ell^2 f(\theta^\star)\boldsymbol{\alpha}\\ =\ &\frac{1}{K^2}\min_{\|\boldsymbol{\alpha}\|_2=1}\mathbb{E}_{\boldsymbol{h}\sim\mathcal{P}}\Big(\sum_{j=1}^{K}\boldsymbol{\alpha}_j^\top\boldsymbol{h}^{(\ell)}\mathcal{J}_{\ell,k}\phi'(\theta_{\ell,j}^{\star\top}\boldsymbol{h}^{(\ell)})\Big)^2\\ =\ &\frac{1}{K^2}\min_{\|\boldsymbol{\alpha}\|_2=1}\int_{\mathbb{R}^{K_{\ell-1}}}\Big(\sum_{j=1}^{K}\boldsymbol{\alpha}_j^\top\boldsymbol{h}^{(\ell)}\mathcal{J}_{\ell,k}\phi'(\theta_{\ell,j}^{\star\top}\boldsymbol{h}^{(\ell)})\Big)^2 p_H(\boldsymbol{h}^{(\ell)})\cdot d\mathbf{h}^{(\ell)}\\ =\ &\frac{1}{K^2}\min_{\|\boldsymbol{\alpha}\|_2=1}\int_{\{\boldsymbol{h}^{(\ell)}|\mathcal{J}_{\ell,k}\neq0\}}\Big(\sum_{j=1}^{K}\boldsymbol{\alpha}_j^\top\boldsymbol{h}^{(\ell)}\phi'(\theta_{\ell,j}^{\star\top}\boldsymbol{h}^{(\ell)})\Big)^2 p_H(\boldsymbol{h}^{(\ell)})\cdot d\mathbf{h}^{(\ell)}\\ \gtrsim\ &\frac{\rho_1}{K^2},\end{aligned} \tag{87}$$

where the last inequality comes from Lemma 11, and Lemma 11 holds since $\boldsymbol{h}^{(\ell)}$ belongs to sub-Gaussian distribution and θ_ℓ is full rank.

Next, the upper bound of $\nabla_\ell^2 f$ can be bounded as

$$
\begin{aligned}
&\max_{\|\boldsymbol{\alpha}\|_2=1} \boldsymbol{\alpha}^\top \nabla_\ell^2 f(\theta^\star)\boldsymbol{\alpha} \\
=&\frac{1}{K^2}\max_{\|\boldsymbol{\alpha}\|_2=1} \mathbb{E}_{\boldsymbol{x}}\Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^\top \boldsymbol{h}^{(\ell)} \cdot \mathcal{J}_{\ell,k}\phi'(\theta_{\ell,j}^{\star\top}\boldsymbol{h}^{(\ell)})\Big)^2 \\
=&\frac{1}{K^2}\max_{\|\boldsymbol{\alpha}\|_2=1} \mathbb{E}_{\boldsymbol{x}}\sum_{j_1=1}^K\sum_{j_2=1}^K \boldsymbol{\alpha}_{j_1}^\top \boldsymbol{h}^{(\ell)} \cdot \mathcal{J}_{\ell,k}\phi'(\theta_{\ell,j_1}^{\star\top}\boldsymbol{h}^{(\ell)}) \cdot \boldsymbol{\alpha}_{j_2}^\top \boldsymbol{h}^{(\ell)} \cdot \mathcal{J}_{\ell,k}\phi'(\theta_{\ell,j_2}^{\star\top}\boldsymbol{h}^{(\ell)}) \\
=&\frac{1}{K^2}\sum_{j_1=1}^K\sum_{j_2=1}^K \mathbb{E}_{\boldsymbol{x}}\boldsymbol{\alpha}_{j_1}^\top \boldsymbol{h}^{(\ell)} \cdot \mathcal{J}_{\ell,k}\phi'(\theta_{\ell,j_1}^{\star T}\boldsymbol{h}^{(\ell)}) \cdot \boldsymbol{\alpha}_{j_2}^\top \boldsymbol{h}^{(\ell)} \cdot \mathcal{J}_{\ell,k}\phi'(\theta_{\ell,j_2}^{\star\top}\boldsymbol{h}^{(\ell)}) \\
\leq&\frac{1}{K^2}\max_{\|\boldsymbol{\alpha}\|_2=1}\sum_{j_1=1}^K\sum_{j_2=1}^K \Big[\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{\alpha}_{j_1}^\top \boldsymbol{h}^{(\ell)})^4 \cdot \mathbb{E}(\phi'(\theta_{\ell,j_1}^{\star\top}\boldsymbol{h}^{(\ell)}))^4 \cdot \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{\alpha}_{j_2}^\top \boldsymbol{h}^{(\ell)})^4 \cdot \mathbb{E}_{\boldsymbol{x}}(\phi'(\theta_{\ell,j_2}^{\star\top}\boldsymbol{h}^{(\ell)}))^4\Big]^{1/4} \\
\leq&\frac{1}{K^2}\max_{\|\boldsymbol{\alpha}\|_2=1}\sum_{j_1=1}^K\sum_{j_2=1}^K \Big[\mathbb{E}_{\boldsymbol{x}}(\boldsymbol{\alpha}_{j_1}^\top \boldsymbol{x})^4 \cdot \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{\alpha}_{j_2}^\top \boldsymbol{x})^4\Big]^{1/4} \\
\leq&\frac{3}{K^2}\sum_{j_1=1}^K\sum_{j_2=1}^K \|\boldsymbol{\alpha}_{j_1}\|_2 \cdot \|\boldsymbol{\alpha}_{j_2}\|_2 \leq \frac{6}{K^2}\sum_{j_1=1}^K\sum_{j_2=1}^K \frac{1}{2}\Big(\|\boldsymbol{\alpha}_{j_1}\|_2^2 + \|\boldsymbol{\alpha}_{j_2}\|_2^2\Big) \\
=&\frac{6}{K}.
\end{aligned}
\tag{88}
$$

Therefore, we have

$$
\lambda_{\max}(\theta^\star) = \max_{\|\boldsymbol{\alpha}\|_2=1} \boldsymbol{\alpha}^\top \nabla_\ell^2 f(\theta^\star; p)\boldsymbol{\alpha} \leq \frac{6}{K}. \tag{89}
$$

Then, given (82), we have

$$
\|\theta - \theta^\star\|_2 \lesssim \frac{2\rho_1}{K}. \tag{90}
$$

Combining (90) and Lemma 10, we have

$$
\|\nabla_\ell^2 f(\theta) - \nabla_\ell^2 f(\theta^\star)\|_2 \lesssim \frac{\rho_1}{K^2}. \tag{91}
$$

Therefore, from (91) and (86), we have

$$
\begin{aligned}
\lambda_{\max}(\theta) \leq \lambda_{\max}(\theta^\star) + \|\nabla_\ell^2 f(\theta) - \nabla_\ell^2 f(\theta^\star)\|_2 \leq \frac{6}{K} + \frac{\rho_1}{2K^2} \leq \frac{7}{K}, \\
\lambda_{\min}(\theta) \geq \lambda_{\min}(\theta^\star) - \|\nabla_\ell^2 f(\theta) - \nabla_\ell^2 f(\theta^\star)\|_2 \geq \frac{\rho_1}{K^2} - \frac{\rho_1}{2K^2} = \frac{\rho_1}{2K^2},
\end{aligned}
\tag{92}
$$

which completes the proof. □

F.2 Proof of Lemma 7

The error bound between $\|\nabla_\ell f - g_t\|_2$ is divided into bounding $\boldsymbol{I}_1$, $\boldsymbol{I}_2$, $\boldsymbol{I}_3$, and I_4 as shown in (98). $\boldsymbol{I}_1$ represents the deviation of the gradient of $\mathcal{D}_t$ to their expectation, which can be bounded through concentration inequality. $\boldsymbol{I}_2$ is derived from the distribution shift between the trajectory and its stationary distribution, which can be bounded with assumption 2. $\boldsymbol{I}_3$ come from the data distribution shift between the behavior policy and optimal policy. $\boldsymbol{I}_4$ comes from the inconsistency of the "noisy" label and the "ground truth" label in the population risk function (30). To ensure a smooth presentation, we will defer the proof of $I_1 - I_4$ until we have completed the main proof of Lemma 7.

Proof of Lemma 7. From (33), we know that

$$\begin{aligned}
&g^{(t)}(\theta_{\ell,k}^{(t)};\mathcal{X}_m)\\
&=\sum_{m\in\mathcal{D}_t}\big(\psi(\theta^{(t)};\boldsymbol{s}_m,a_m)-y_m^{(t)}\big)\cdot\frac{\partial\psi(\theta^{(t)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}\\
&=\sum_{m\in\mathcal{D}_t}\Big(\psi(\theta^{(t)};\boldsymbol{s}_m,a_m)-\phi(\theta^\star;\boldsymbol{s}_m,a_m)-\gamma\cdot\psi(\boldsymbol{s}'_m,a'_m;\theta^{(t)})\Big)\cdot\frac{\partial\psi(\theta^{(t)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}\\
&=\sum_{m\in\mathcal{D}_t}\Big(\psi(\theta^{(t,n)};\boldsymbol{s}_m,a_m)-\psi(\theta^\star;\boldsymbol{s}_m,a_m)+\gamma\cdot\max_{a'}\psi(\boldsymbol{s}'_m,a';\theta^\star)\\
&\qquad\qquad-\gamma\cdot\psi(\boldsymbol{s}'_m,a'_m;\theta^{(t)})\Big)\cdot\frac{\partial\psi(\theta^{(t,n)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}\\
&=\sum_{m\in\mathcal{D}_t}\Big(\psi(\theta^{(t)};\boldsymbol{s}_m,a_m)-\psi(\theta^\star;\boldsymbol{s}_m,a_m)\Big)\cdot\frac{\partial\psi(\theta^{(t)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}\\
&\quad+\gamma\cdot\Big(\max_{a'}\psi(\boldsymbol{s}'_m,a';\theta^\star)-\psi(\boldsymbol{s}'_m,a'_m;\theta^{(t)})\Big)\cdot\frac{\partial\psi(\theta^{(t)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}\\
&:=\sum_{m\in\mathcal{D}_t}b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)+\Delta b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m),
\end{aligned}\tag{93}$$

where we have

$$b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)=\Big(\psi(\theta^{(t)};\boldsymbol{s}_m,a_m)-\psi(\theta^\star;\boldsymbol{s}_m,a_m)\Big)\cdot\frac{\partial\psi(\theta^{(t)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}\tag{94}$$

and

$$\Delta b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)=\Big(\max_{a'}\psi(\theta^\star;\boldsymbol{s}'_m,a')-\psi(\theta^{(t-1)};\boldsymbol{s}'_m,a'_m)\Big)\cdot\frac{\partial\psi(\theta^{(t)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}.\tag{95}$$

Then, let us define $\bar{b}_{\ell,k}^{(t)}$ as

$$\bar{b}_{\ell,k}^{(t)}(\theta;\mathcal{X})=\mathbb{E}_{(\boldsymbol{s},a)\sim\mu_t}\Big(\psi(\theta;\boldsymbol{s},a)-\psi(\theta^\star;\boldsymbol{s},a)\Big)\cdot\nabla_\theta\psi(\theta;\boldsymbol{s},a).\tag{96}$$

From (30), we know that

$$\frac{\partial f_{\pi^\star}}{\partial\theta_{\ell,k}}(\theta^{(t)})=\mathbb{E}_{(\boldsymbol{s},a)\sim\mu^\star}\Big(\phi(\theta^{(t)};\boldsymbol{s},a)-\phi(\theta^\star;\boldsymbol{s},a)\Big)\cdot\frac{\partial\phi(\theta^{(t)};\boldsymbol{s},a)}{\partial\theta_{\ell,k}}.\tag{97}$$

Then, from (93) and (97), we have

$$\begin{aligned}
&g^{(t)}(\theta_{\ell,k}^{(t)};\mathcal{X}_m)-\frac{\partial f_{\pi^\star}}{\partial\theta_{\ell,k}}(\theta^{(t)};\mathcal{X}_m)\\
&=\sum_{m\in\mathcal{D}_t}b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)+\Delta b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)-\frac{\partial f_{\pi^\star}}{\partial\theta_{\ell,k}}(\theta^{(t)};\mathcal{X}_m)\\
&=\Big[b_{\ell,k}^{(t)}(\theta_{\ell,k}^{(t)};\mathcal{X}_m)-\mathbb{E}_{\mathcal{X}_m\sim\mathcal{D}_t}b_{\ell,k}^{(t)}(\theta_{\ell,k}^{(t)};\mathcal{X}_m)\Big]+\Big[\mathbb{E}_{\mathcal{X}_m\sim\mathcal{D}_t}b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)-\bar{b}_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)\Big]\\
&\quad+\Big[\bar{b}_{\ell,k}^{(t)}(\theta^{(t)})-\frac{\partial f_{\pi^\star}}{\partial\theta_{\ell,k}}(\theta^{(t)})\Big]+\mathbb{E}_{\mathcal{X}_m\sim\mathcal{D}_t}\Delta b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)\\
&:=\boldsymbol{I}_1+\boldsymbol{I}_2+\boldsymbol{I}_3+\boldsymbol{I}_4.
\end{aligned}\tag{98}$$

Therefore, we have

$$\Big\|g^{(t)}(\theta_{\ell,k}^{(t)};\mathcal{X}_m)-\frac{\partial f_{\pi^\star}}{\partial\theta_{\ell,k}}(\theta^{(t)})\Big\|_2\le\|\boldsymbol{I}_1\|_2+\|\boldsymbol{I}_2\|_2+\|\boldsymbol{I}_3\|_2+\|\boldsymbol{I}_4\|_2.\tag{99}$$

Next, we first provide the bound for $\|\boldsymbol{I}_1\|_2$, $\|\boldsymbol{I}_2\|_2$, $\|\boldsymbol{I}_3\|_2$, and $\|\boldsymbol{I}_4\|_2$ as

$$\begin{aligned}\|\boldsymbol{I}_1\|_2 &\leq \frac{1}{K_\ell}\cdot\|\theta-\theta^\star\|_2\cdot\sqrt{\frac{d\log q}{|\mathcal{D}_t|}},\\ \|\boldsymbol{I}_2\|_2 &\leq \frac{R_{\max}}{1-\gamma}\cdot(1+\gamma)\tau^\star\cdot\eta_{t-\tau^\star},\\ \|\boldsymbol{I}_3\|_2 &\leq |\mathcal{A}|\cdot\frac{R_{\max}}{1-\gamma}\cdot(1+\log_\nu\lambda^{-1}+\frac{1}{1-\nu})\cdot C_t,\\ \|\boldsymbol{I}_4\|_2 &\leq \frac{\gamma}{K_\ell}\cdot\|\theta^{(t}-\theta^\star\|_2,\end{aligned} \tag{100}$$

where $|\mathcal{A}|$ is the size of action space. The details for the derivation of I_1- I_4 can be found after the proof.

Let $\boldsymbol{\alpha}\in\mathbb{R}^{Kd}$ and $\boldsymbol{\alpha}_j\in\mathbb{R}^d$ with $\boldsymbol{\alpha}=[\boldsymbol{\alpha}_1^T,\boldsymbol{\alpha}_2^T,\cdots,\boldsymbol{\alpha}_K^T]^T$, with probability at least $1-q^{-d}$, we have

$$\begin{aligned}\|g^{(t)}(\theta_\ell;\theta)-\nabla_\ell f_{\pi^\star}(\theta)\|_2^2 =&\Big|\boldsymbol{\alpha}^T\big(g^{(t)}(\theta)-\nabla f_{\pi^\star}(\theta)\big)\Big|^2\\ \leq&\sum_{k=1}^K\Big|\boldsymbol{\alpha}_k^T\big(g^{(t)}(\theta_{\ell,k};\theta)-\frac{\partial f_{\pi^\star}}{\partial\theta_{\ell,k}}(\theta)\big)\Big|^2\\ \leq&\sum_{k=1}^K\Big\|g^{(t)}(\theta_{\ell,k};\theta)-\frac{\partial f_{\pi^\star}}{\partial\theta_{\ell,k}}(\theta)\Big\|_2^2\cdot\|\boldsymbol{\alpha}_k\|_2^2\\ \leq&\max_k\Big\|g^{(t)}(\theta_{\ell,k};\theta)-\frac{\partial f_{\pi^\star}}{\partial\theta_{\ell,k}}(\theta)\Big\|_2^2.\end{aligned} \tag{101}$$

In conclusion, we have

$$\begin{aligned}&\|g^{(t)}(\theta_\ell;\theta)-\nabla_\ell f_{\pi^\star}(\theta)\|_2\\ \leq&\max_k\Big\|g^{(t)}(\theta_{\ell,k};\theta)-\frac{\partial f_{\pi^\star}}{\partial\theta_{\ell,k}}(\theta)\Big\|_2^2\\ \leq&\max_k\|\boldsymbol{I}_1(k)\|_2+\|\boldsymbol{I}_2(k)\|_2+\|\boldsymbol{I}_3(k)\|_2+\|\boldsymbol{I}_4(k)\|_2\\ \leq&\frac{1}{K_\ell}\cdot\|\theta-\theta^\star\|_2\cdot\sqrt{\frac{d\log q}{|\mathcal{D}_t|}}+\frac{R_{\max}}{1-\gamma}\cdot(1+\gamma)\tau^\star\cdot\eta_{t-\tau^\star}\\ &+|\mathcal{A}|\cdot\frac{R_{\max}}{1-\gamma}\cdot(1+\log_\nu\lambda^{-1}+\frac{1}{1-\nu})\cdot C_t+\frac{\gamma}{K_\ell}\cdot\|\theta^{(t)}-\theta^\star\|_2,\end{aligned} \tag{102}$$

where $\tau^\star=\min\{t\mid\lambda\nu^t\leq\eta_T\}$ □

F.2.1 Proof of upper bound of I_1

Proof. We define a random variable

$$Z^{(\ell)}(k)=\big(\psi(\theta;\boldsymbol{s},a)-\psi(\theta^\star;\boldsymbol{s},a)\big)\cdot\mathcal{J}_{\ell,k}\cdot\boldsymbol{\alpha}^T\boldsymbol{h}^{(\ell)}(\theta)$$

with $(\boldsymbol{s},a)\sim\mathcal{D}_t$ and

$$Z_m^{(\ell)}(k)=\big(Q(\boldsymbol{x}_m;\theta)-Q(\boldsymbol{x}_m;\theta^\star)\big)\cdot\mathcal{J}_{\ell,k}\cdot\boldsymbol{\alpha}^T\boldsymbol{h}_n^{(\ell)}(\theta)$$

as the realization of $Z^{(\ell)}$ for $m\in\mathcal{D}_t$, where $\boldsymbol{\alpha}$ is any fixed unit vector.

According to the definition of I_1 in (98), we can rewrite $\boldsymbol{I}_1$ as

$$\boldsymbol{I}_1=\frac{1}{K_\ell}\Big[\sum_{m\in\mathcal{D}_t}Z_m^{(\ell)}(k)-\mathbb{E}_{(\boldsymbol{s},a)\sim\mathcal{D}_t}Z^{(\ell)}(k)\Big]. \tag{103}$$

Then, for any $p \in \mathbb{N}^+$, we have

$$\begin{aligned}\big(\mathbb{E}|Z^{(\ell)}|^p\big)^{1/p} =& \Big(\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_t}|\psi(\theta;\boldsymbol{s},a)-\psi(\theta^\star;\boldsymbol{s},a)|^p\cdot|\mathcal{J}_{\ell,k}\sigma'(\boldsymbol{w}_{\ell,k}^\top\boldsymbol{x})|\cdot|\boldsymbol{\alpha}^T\boldsymbol{h}^{(\ell)}|^p\Big)^{1/p}\\ \le& \Big(\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{t,1}}|\psi(\theta;\boldsymbol{s},a)-\psi(\theta^\star;\boldsymbol{s},a)|^p\cdot|\boldsymbol{\alpha}^T\boldsymbol{h}^{(\ell)}|^p\Big)^{1/p}\\ \le& \Big(\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_t}\Big|\|\theta-\theta^\star\|_2\cdot\|\boldsymbol{x}(\boldsymbol{s},a)\|_2\Big|^p\cdot\big|\boldsymbol{\alpha}^T\boldsymbol{x}(\boldsymbol{s},a)\big|^p\Big)^{1/p}\\ \lesssim& \cdot\|\theta-\theta^\star\|_2\cdot p.\end{aligned} \tag{104}$$

From Definition 2, we know that $Z^{(\ell)}$ belongs to sub-exponential distribution with $\|Z^{(\ell)}\|_{\psi_1} \lesssim \|\theta-\theta^\star\|_2$. Therefore, by Chernoff inequality, for any $s\in\mathbb{R}$, we have

$$\mathbb{P}\bigg\{\Big|\frac{1}{|\mathcal{D}_t|}\sum_{m\in\mathcal{D}_t}Z_m^{(\ell)}(k)-\mathbb{E}Z^{(\ell)}(k)\Big|<t\bigg\}\le 1-\frac{e^{-(\|\theta-\theta^\star\|_2)^2\cdot|\mathcal{D}_t|\cdot s^2}}{e^{|\mathcal{D}_t|\cdot st}}. \tag{105}$$

Let $t=\|\theta-\theta^\star\|_2\sqrt{\frac{d\log q}{N}}$ and $s=\frac{2}{\|\theta-\theta^\star\|_2}\cdot t$ for some large constant $q>0$. Then, with probability at least $1-q^{-d}$, we have

$$\Big|\frac{1}{|\mathcal{D}_t|}\sum_{m\in\mathcal{D}_t}Z_m^{(\ell)}(k)-\mathbb{E}Z^{(\ell)}(k)\Big|\lesssim\|\theta-\theta^\star\|_2\cdot\sqrt{\frac{d\log q}{|\mathcal{D}_t|}}. \tag{106}$$

From Lemma 4 and (103), with probability at least $1-|\mathcal{S}_{\frac{1}{2}}(d)|\cdot q^{-d}$, we have

$$\|\boldsymbol{I}_1\|_2\le 2\cdot\frac{1}{K_\ell}\bigg|\frac{1}{|\mathcal{D}_t|}\sum_{m\in\mathcal{D}_t}Z_m^{(\ell)}-\mathbb{E}Z^{(\ell)}\bigg|\lesssim\frac{1}{K_\ell}\|\theta-\theta^\star\|_2\cdot\sqrt{\frac{d\log q}{|\mathcal{D}_t|}}. \tag{107}$$

From Lemma 3, we know that $|\mathcal{S}_{\frac{1}{2}}(d)|\le 5^d$. Therefore, the probability for (107) holds is at least $1-\big(\frac{q}{5}\big)^{-d}$. Because $q\gg 5$, we denote the probability as $1-q^{-d}$ for convenience. □

F.2.2 Proof of upper bound of I_2

Proof. $\boldsymbol{I}_2$ is the bias of the data because the data $(\boldsymbol{s},a)$ at iteration t depends on the neural network parameters $\theta^{(t)}$. Recall the definition of $b_{\ell,k}^{(t)}$ and $\bar{b}_{\ell,k}^{(t)}$, we define

$$\Delta_t=b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)-\bar{b}_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m). \tag{108}$$

It is easy to verify that

$$\begin{aligned}&\|b_{\ell,k}^{(t)}(\theta;\mathcal{X}_m)-b_{\ell,k}^{(t)}(\tilde{\theta};\mathcal{X}_m)\|_2\le(1+\gamma)\cdot\|\theta-\tilde{\theta}\|_2,\\ &\|\bar{b}_{\ell,k}^{(t)}(\theta;\mathcal{X}_m)-\bar{b}_{\ell,k}^{(t)}(\tilde{\theta};\mathcal{X}_m)\|_2\le(1+\gamma)\cdot\|\theta-\tilde{\theta}\|_2,\\ &\qquad\textit{and}\qquad\|b_{\ell,k}^{(t)}\|\lesssim\frac{R_{\max}}{1-\gamma}.\end{aligned} \tag{109}$$

Then, we have

$$\Delta_t(\theta)-\Delta_t(\tilde{\theta})\lesssim(1+\gamma)\cdot\|\theta-\tilde{\theta}\|_2. \tag{110}$$

Therefore, we have

$$\Delta_t(\theta^{(t)})\le\Delta_t(\theta^{(t-\tau)})+\frac{1+\gamma}{1-\gamma}\cdot R_{\max}\cdot\sum_{i=t-\tau}^{t-1}\eta_i. \tag{111}$$

Then, we need to bound $\delta_t(\theta^{(t-\tau)})$.

Let us define the observed tuple $O_t(\boldsymbol{s},a,s')$ as the collection of the state, action, and the next state at the t-th iteration. Note that

$$\theta^{(t-\tau)}\longrightarrow\boldsymbol{s}_{t-\tau}\longrightarrow\boldsymbol{s}_t\longrightarrow O_t \tag{112}$$

forms a Markov chain introduced by the policy π_t.

Let $\tilde{\theta}^{(t-\tau,0)}$ and $\widetilde{O}_t$ be independently drawn from the marginal distributions of $\theta^{(t-\tau,0)}$ and O_t, respectively.

With Lemma 9 in Bhandari et al. (2018), we have

$$\mathbb{E}\ \Delta_t(\theta^{(t-\tau)}, O_t) - \mathbb{E}\ \Delta_t(\tilde{\theta}^{(t-\tau)}, \tilde{O}_t) \lesssim\ 2 \sup_{\theta,O} |\Delta_t(\theta, O)| \cdot \lambda \cdot \nu^\tau. \tag{113}$$

By definition, we have $\mathbb{E}\ \Delta_m(\tilde{\theta}^{(t-\tau)}, \widetilde{O}_t) = 0$ and

$$|\Delta_t(\theta, O)| \leq \frac{2\ R_{\max}}{1-\gamma}. \tag{114}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}\Delta_t(\theta^{(t)}) \leq& \mathbb{E}\Delta_t(\theta^{(t-\tau)}) + \frac{1+\gamma}{1-\gamma} \cdot R_{\max} \cdot \sum_{i=t-\tau}^{t-1} \eta_i \\ \leq& \frac{R_{\max}}{1-\gamma}\Big(\lambda \cdot \nu^\tau + (1+\gamma) \cdot \tau \cdot \eta_{t-\tau}\Big),\end{aligned} \tag{115}$$

where the last inequality comes from the fact that the step size η_m is non-increasing.

Choose $\tau^\star = \min\big\{t = 0, 1, 2, \cdots \mid \lambda\nu^\tau \leq \eta_T\big\}$. When $t \leq \tau^\star$, we choose $\tau = t$ and have

$$\mathbb{E}\Delta_t(\theta^{(t)}) \leq \frac{R_{\max}}{1-\gamma} \cdot \tau^\star \cdot \eta_0. \tag{116}$$

When $n > \tau^\star$, we can choose $\tau = \tau^\star$ and obtain

$$\mathbb{E}\Delta_t(\theta^{(t)}) \leq \frac{R_{\max}}{1-\gamma} \cdot (1+\gamma)\tau^\star \cdot \eta_{t-\tau^\star}. \tag{117}$$

Combining (116) and (117), we have

$$|\boldsymbol{I}_2| \leq \frac{R_{\max}}{1-\gamma} \cdot (1+\gamma)\tau^\star \cdot \eta_{\max\{0, t-\tau^\star\}}, \tag{118}$$

where $\tau^\star = \min\{t \mid \lambda\nu^t \leq \eta_T\}$. □

F.2.3 Proof of bound of I_3

Proof. We have

$$\begin{aligned}\boldsymbol{I}_3 =& \bar{b}_{\ell,k}^{(t)}(\theta^{(t)}) - \frac{\partial f_{\pi^\star}}{\partial \theta_{\ell,k}}(\theta^{(t)}) \\ =& \mathbb{E}_{(\boldsymbol{s},a)\sim\mu_t}\Big(\psi(\theta; \boldsymbol{s}, a) - \psi(\theta^\star; \boldsymbol{s}, a)\Big) \cdot \frac{\partial \psi(\theta; \boldsymbol{s}, a)}{\partial \theta_{\ell,k}} \\ & - \mathbb{E}_{(\boldsymbol{s},a)\sim\mu^\star}\Big(\psi(\theta; \boldsymbol{s}, a) - \psi(\theta^\star; \boldsymbol{s}, a)\Big) \cdot \frac{\partial \psi(\theta; \boldsymbol{s}, a)}{\partial \theta_{\ell,k}} \\ =& \mathbb{E}_{(\boldsymbol{s},a)\sim\mu_t}\Big(\psi(\theta; \boldsymbol{s}, a) - r(\boldsymbol{s}, a) - \gamma \cdot \mathbb{E}_{\boldsymbol{s}'\sim p^a_{\boldsymbol{s},\boldsymbol{s}'}} \max_{a'} \psi(\theta^\star; \boldsymbol{s}', a')\Big) \cdot \frac{\partial \psi(\theta; \boldsymbol{s}, a)}{\partial \theta_{\ell,k}} \\ & - \mathbb{E}_{(\boldsymbol{s},a)\sim\mu^\star}\Big(\psi(\theta; \boldsymbol{s}, a) - r(\boldsymbol{s}, a) - \gamma \cdot \mathbb{E}_{\boldsymbol{s}'\sim p^a_{\boldsymbol{s},\boldsymbol{s}'}} \max_{a'} \psi(\theta^\star; \boldsymbol{s}', a')\Big) \cdot \frac{\partial \psi(\theta; \boldsymbol{s}, a)}{\partial \theta_{\ell,k}} \\ =& \mathbb{E}_{(\boldsymbol{s},a)\sim\mu_t, \boldsymbol{s}'\sim p^a_{\boldsymbol{s},\boldsymbol{s}'}}\Big(\psi(\theta; \boldsymbol{s}, a) - r(\boldsymbol{s}, a) - \gamma \cdot \max_{a'} \psi(\theta^\star; \boldsymbol{s}', a')\Big) \cdot \frac{\partial \psi(\theta; \boldsymbol{s}, a)}{\partial \theta_{\ell,k}} \\ & - \mathbb{E}_{(\boldsymbol{s},a)\sim\mu^\star, \boldsymbol{s}'\sim p^a_{\boldsymbol{s},\boldsymbol{s}'}}\Big(\psi(\theta; \boldsymbol{s}, a) - r(\boldsymbol{s}, a) - \gamma \cdot \max_{a'} \psi(\theta^\star; \boldsymbol{s}', a')\Big) \cdot \frac{\partial \psi(\theta; \boldsymbol{s}, a)}{\partial \theta_{\ell,k}}\end{aligned} \tag{119}$$

Then, we have

$$\begin{aligned}
&\Big|\int_{(\boldsymbol{s},a)}\int_{\boldsymbol{s}'}\big(\mu^\star(d\boldsymbol{s},da)\mathcal{P}(d\boldsymbol{s}'|\boldsymbol{s},a)-\mu_t(d\boldsymbol{s},da)\mathcal{P}(d\boldsymbol{s}'|\boldsymbol{s},a)\big)\Big|\\
=&\Big|\int_{(\boldsymbol{s},a)}\int_{\boldsymbol{s}'}\big(\mathcal{P}^\star(d\boldsymbol{s})\pi^\star(da|\boldsymbol{s})\mathcal{P}(d\boldsymbol{s}'|\boldsymbol{s},a)-\mathcal{P}_t(d\boldsymbol{s})\pi_t(da|d\boldsymbol{s})\mathcal{P}(d\boldsymbol{s}'|\boldsymbol{s},a)\big)\Big|\\
\leq&\Big|\int_{(\boldsymbol{s},a)}\int_{\boldsymbol{s}'}\big(\mathcal{P}^\star(d\boldsymbol{s})-\mathcal{P}_t(d\boldsymbol{s})\big)\pi^\star(da|\boldsymbol{s})\mathcal{P}(d\boldsymbol{s}'|\boldsymbol{s},a)\Big|\\
&+\Big|\int_{(\boldsymbol{s},a)}\int_{\boldsymbol{s}'}\mathcal{P}_t(d\boldsymbol{s})\big(\pi_t(da|d\boldsymbol{s})-\pi^\star(da|d\boldsymbol{s})\big)\mathcal{P}(d\boldsymbol{s}'|\boldsymbol{s},a)\Big|.
\end{aligned}\tag{120}$$

From Theorem 3.1 in Mitrophanov (2005), we know that

$$\begin{aligned}
&\Big|\int_{(\boldsymbol{s},a)}\big(\mathcal{P}^\star(d\boldsymbol{s})-\mathcal{P}_t(d\boldsymbol{s})\big)\Big|\leq|\mathcal{A}|(\log_\nu\lambda^{-1}+\frac{1}{1-\nu})C_t\\
&\qquad\text{and}\qquad\big\|\pi_t(da|d\boldsymbol{s})-\pi^\star(da|d\boldsymbol{s})\big\|\leq C_t.
\end{aligned}\tag{121}$$

Therefore, the bound of $\boldsymbol{I}_3$ can be found as

$$\begin{aligned}
\|\boldsymbol{I}_3\|_2\leq&\frac{R_{\max}}{1-\gamma}\cdot|\mathcal{A}|\cdot C_t\cdot(1+\log_\nu\lambda^{-1}+\frac{1}{1-\nu})\\
=&|\mathcal{A}|\cdot\frac{R_{\max}}{1-\gamma}\cdot(1+\log_\nu\lambda^{-1}+\frac{1}{1-\nu})\cdot C_t.
\end{aligned}\tag{122}$$

□

F.2.4 PROOF OF BOUND OF I_4

Proof. We have

$$\begin{aligned}
\|\boldsymbol{I}_4\|=&\|\Delta b_{\ell,k}^{(t)}(\theta^{(t)};\mathcal{X}_m)\|_2\\
=&\max_{\boldsymbol{s},a}\gamma\cdot\Big(\max_{a'}\psi(\boldsymbol{s}'_m,a';\theta^\star)-\psi(\boldsymbol{s}'_m,a'_m;\theta^{(t)})\Big)\cdot\Big\|\frac{\partial\psi(\theta^{(t)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}\Big\|_2\\
\leq&\max_{\boldsymbol{s},a}\gamma\cdot\Big(\max_{a'}\psi(\boldsymbol{s}'_m,a';\theta^\star)-\max_{a'}\psi(\boldsymbol{s}'_m,a';\theta^{(t)})\Big)\cdot\Big\|\frac{\partial\psi(\theta^{(t)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}\Big\|_2\\
\leq&\gamma\cdot\max_{s,a,a'}\Big|\psi(\boldsymbol{s}'_m,a';\theta^\star)-\psi(\boldsymbol{s}'_m,a';\theta^{(t)})\Big|\cdot\Big\|\frac{\partial\psi(\theta^{(t)};\mathcal{X}_m)}{\partial\theta_{\ell,k}}\Big\|_2\\
\lesssim&\gamma\cdot\|\theta^{(t)}-\theta^\star\|_2\cdot\frac{1}{K_\ell}\\
\leq&\frac{\gamma}{K_\ell}\|\theta^{(t)}-\theta^\star\|_2.
\end{aligned}\tag{123}$$

□

F.3 PROOF OF LEAMMA 8

Proof of Lemma 8. From the update rule of $\boldsymbol{w}$ in Algorithm 1, we have

$$\begin{aligned}
\boldsymbol{w}^{(t+1)}-\boldsymbol{w}^\star=&\boldsymbol{w}^{(t)}-\boldsymbol{w}^\star-\kappa_t\cdot\sum_{m\in\mathcal{D}_t}(\phi_m^\top\boldsymbol{w}^{(t)}-r_m)\cdot\phi_m\\
=&\boldsymbol{w}^{(t)}-\boldsymbol{w}^\star-\kappa_t\cdot\sum_{m\in\mathcal{D}_t}(\phi_m^\top\boldsymbol{w}^{(t)}-\phi_m\boldsymbol{w}^\star)\cdot\phi_m\\
=&\Big(\boldsymbol{I}-\kappa_t\sum_{m\in\mathcal{D}_m}\phi_m^\top\phi_m\Big)\cdot(\boldsymbol{w}^{(t)}-\boldsymbol{w}^\star).
\end{aligned}\tag{124}$$

For any unit vector $\alpha \in dim(\boldsymbol{w})$, we have

$$\begin{aligned} |\boldsymbol{\alpha}^\top \mathbb{E}_{\mathcal{D}_t} \phi^\top \phi \boldsymbol{\alpha}| &\leq \max_{\|\phi\|_2} |\boldsymbol{\alpha}^\top \phi|^2 \leq \phi_{\max}^2, \\ |\boldsymbol{\alpha}^\top \mathbb{E}_{\mathcal{D}_t} \phi^\top \phi \boldsymbol{\alpha}| &\geq |\boldsymbol{\alpha}^\top \phi_{\min}|^2 \geq 0. \end{aligned} \tag{125}$$

Also, it is easy to verify that $|\boldsymbol{\alpha}^\top \mathbb{E}_{\mathcal{D}_t} \phi^\top \phi \boldsymbol{\alpha}| = 0$ if only and if ϕ_m are all parallel to each other. As ϕ_m does not parallel to each other, let $\rho_2 > 0$ denote the minimal eigenvalue of $\mathbb{E}_{\mathcal{D}_t} \phi^\top \phi$.

Given ϕ is bounded, ϕ belongs to the sub-Gaussian distribution. Similar to (106), with Chebyshev's inequality, we have

$$\left\| \sum_{m \in \mathcal{D}_m} \phi_m^\top \phi_m - \mathbb{E}_{\mathcal{D}_t} \phi^\top \phi \right\|_2 \leq \sqrt{\frac{d \log q}{|\mathcal{D}_t|}} \tag{126}$$

with probability at least $1 - d^{-q}$. Let $N \geq c_N^{-2} d \log q$, according to Lemma 1, we have

$$\lambda_{\min}(\sum_{m \in \mathcal{D}_m} \phi_m^\top \phi_m) \leq \lambda_{\min}(\mathbb{E}_{\mathcal{D}_t} \phi^\top \phi) - c_N \leq \rho_2 - c_N. \tag{127}$$

When we choose $\kappa_t = \frac{1}{\phi_{\max}}$, we have

$$\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^\star\|_2 \leq \Big(1 - \frac{\rho_2 - c_N}{\phi_{\max}}\Big) \cdot \|\boldsymbol{w}^{(t)} - \boldsymbol{w}^\star\|_2. \tag{128}$$

□

G PROOF OF LEMMAS IN APPENDIX C

Lemma 13. *We have*

$$\left|Q_i^{\pi_i^\star}(\boldsymbol{s}, a) - Q_i^{\pi_j^\star}(\boldsymbol{s}, a)\right| \leq \frac{2\gamma}{1-\gamma} \cdot \max_{\boldsymbol{s}, a} |r_i(\boldsymbol{s}, a) - r_j(\boldsymbol{s}, a)|. \tag{129}$$

Proof. $\left|Q_i^{\pi_i^\star}(\boldsymbol{s}, a) - Q_i^{\pi_j^\star}(\boldsymbol{s}, a)\right|$ can be upper bounded as

$$\begin{aligned} &\left|Q_i^{\pi_i^\star}(\boldsymbol{s}, a) - Q_i^{\pi_j^\star}(\boldsymbol{s}, a)\right| \\ =&\Big|r_i + \gamma \cdot \sum_{\boldsymbol{s}'} p_{\boldsymbol{s},\boldsymbol{s}'}^a Q_i^{\pi_i^\star}\big(\boldsymbol{s}', \pi_i^\star(\boldsymbol{s}')\big) - \Big(r_i + \gamma \cdot \sum_{\boldsymbol{s}'} p_{\boldsymbol{s},\boldsymbol{s}'}^a Q_i^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big)\Big)\Big| \\ =&\gamma \cdot \Big| \sum_{\boldsymbol{s}'} p_{\boldsymbol{s},\boldsymbol{s}'}^a Q_i^{\pi_i^\star}\big(\boldsymbol{s}', \pi_i^\star(\boldsymbol{s}')\big) - \sum_{\boldsymbol{s}'} p_{\boldsymbol{s},\boldsymbol{s}'}^a Q_i^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big)\Big| \\ \leq&\gamma \cdot \sum_{\boldsymbol{s}'} p_{\boldsymbol{s},\boldsymbol{s}'}^a \cdot \Big|Q_i^{\pi_i^\star}\big(\boldsymbol{s}', \pi_i^\star(\boldsymbol{s}')\big) - Q_i^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big)\Big| \\ \leq&\gamma \cdot \sum_{\boldsymbol{s}'} p_{\boldsymbol{s},\boldsymbol{s}'}^a \cdot \Big[\Big|Q_i^{\pi_i^\star}\big(\boldsymbol{s}', \pi_i^\star(\boldsymbol{s}')\big) - Q_j^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big)\Big| + \Big|Q_j^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big) - Q_i^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big)\Big|\Big] \\ =&\gamma \cdot \sum_{\boldsymbol{s}'} p_{\boldsymbol{s},\boldsymbol{s}'}^a \cdot \Big[\Big|\max_{a'} Q_i^{\pi_i^\star}\big(\boldsymbol{s}', a'\big) - \max_{a'} Q_j^{\pi_j^\star}\big(\boldsymbol{s}', a'\big)\Big| + \Big|Q_j^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big) - Q_i^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big)\Big|\Big] \\ \leq&\gamma \cdot \sum_{\boldsymbol{s}'} p_{\boldsymbol{s},\boldsymbol{s}'}^a \cdot \Big[\max_{a'} \Big|Q_i^{\pi_i^\star}\big(\boldsymbol{s}', a'\big) - Q_j^{\pi_j^\star}\big(\boldsymbol{s}', a'\big)\Big| + \Big|Q_j^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big) - Q_i^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big)\Big|\Big] \\ \leq&\gamma \cdot \sum_{\boldsymbol{s}'} p_{\boldsymbol{s},\boldsymbol{s}'}^a \cdot \Big[\max_{\boldsymbol{s}', a'} \Big|Q_i^{\pi_i^\star}\big(\boldsymbol{s}', a'\big) - Q_j^{\pi_j^\star}\big(\boldsymbol{s}', a'\big)\Big| + \max_{\boldsymbol{s}'} \Big|Q_j^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big) - Q_i^{\pi_j^\star}\big(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')\big)\Big|\Big] \end{aligned} \tag{130}$$

Let

$$I_5 = \max_{\boldsymbol{s}, a} \Big|Q_i^{\pi_i^\star}\big(\boldsymbol{s}, a\big) - Q_j^{\pi_j^\star}\big(\boldsymbol{s}, a\big)\Big|$$

and

$$I_6 = \max_{\boldsymbol{s},a} \left| Q_j^{\pi_j^\star}(\boldsymbol{s},a) - Q_i^{\pi_j^\star}(\boldsymbol{s},a) \right| \geq \max_{\boldsymbol{s}} \left| Q_j^{\pi_j^\star}(\boldsymbol{s},\pi_j^\star(\boldsymbol{s})) - Q_i^{\pi_j^\star}(\boldsymbol{s},\pi_j^\star(\boldsymbol{s})) \right|.$$

Then, we have

$$\begin{aligned} I_5 =& \max_{\boldsymbol{s},a} \left| r_i + \gamma \cdot \sum_{\boldsymbol{s}'} p^a_{\boldsymbol{s},\boldsymbol{s}'} \max_{a'} Q_i^{\pi_i^\star}(\boldsymbol{s}',a') - r_j - \gamma \cdot \sum_{\boldsymbol{s}'} p^a_{\boldsymbol{s},\boldsymbol{s}'} \max_{a'} Q_j^{\pi_j^\star}(\boldsymbol{s}',a') \right| \\ \leq& \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)| + \gamma \max_{\boldsymbol{s},a} \sum_{\boldsymbol{s}'} p^a_{\boldsymbol{s},\boldsymbol{s}'} \cdot \max_{a'} |Q_i^{\pi_i^\star}(\boldsymbol{s}',a') - Q_j^{\pi_j^\star}(\boldsymbol{s}',a')| \\ \leq& \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)| + \gamma \cdot I_5. \end{aligned} \tag{131}$$

Therefore, we have

$$I_5 \leq \frac{1}{1-\gamma} \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)|. \tag{132}$$

Similar to (131), we have

$$\begin{aligned} I_6 \leq& \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)| + \gamma \max_{\boldsymbol{s},a} \sum_{\boldsymbol{s}'} p^a_{\boldsymbol{s},\boldsymbol{s}'} \cdot |Q_j^{\pi_j^\star}(\boldsymbol{s}',\pi_j^\star(\boldsymbol{s}')) - Q_i^{\pi_j^\star}(\boldsymbol{s}',\pi_j^\star(\boldsymbol{s}'))| \\ \leq& \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)| + \gamma \cdot I_6. \end{aligned} \tag{133}$$

Therefore, we have

$$I_6 \leq \frac{1}{1-\gamma} \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)|. \tag{134}$$

Therefore, we have

$$\left| Q_i^{\pi_i^\star}(\boldsymbol{s},a) - Q_j^{\pi_i^\star}(\boldsymbol{s},a) \right| \leq \gamma(I_5 + I_6) \leq \frac{2\gamma}{1-\gamma} \cdot \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)|. \tag{135}$$

□

H ADDITIONAL PROOF OF THE LEMMAS

H.1 PROOF OF LEMMA 10

The distance of the second order derivatives of the population risk function $f(\cdot)$ at point θ and $\theta^\star$ can be converted into bounding $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, which are defined in (137). The major idea in proving $\boldsymbol{P}_1$ is to connect the error bound to the angle between θ and $\theta^\star$ given $\boldsymbol{h}^{(\ell)}$ belongs to the sub-Gaussian distribution.

Proof of Lemma 10. From the definition of f in (30), we have

$$\begin{aligned} &\frac{\partial^2 f}{\partial\theta_{\ell,j_1}\partial\theta_{\ell,j_2}}(\theta^\star) = \frac{1}{K^2}\mathbb{E}_{\boldsymbol{x}} \mathcal{J}_{\ell,k}\sigma'(\theta_{j_1}^{\star\top}\boldsymbol{h}) \cdot \mathcal{J}_{\ell,k}\sigma'(\theta_{j_2}^{\star\top}\boldsymbol{h}) \cdot \boldsymbol{h}^\star\boldsymbol{h}^{\star\top}, \\ \text{and} \quad &\frac{\partial^2 f}{\partial\theta_{\ell,j_1}\partial\theta_{\ell,j_2}}(\theta) = \frac{1}{K^2}\mathbb{E}_{\boldsymbol{x}}\sigma' \mathcal{J}^\star_{\ell,k}(\theta_{\ell,j_1}^\top\boldsymbol{h}) \cdot \mathcal{J}^\star_{\ell,k}\sigma'(\theta_{\ell,j_2}^\top\boldsymbol{h}) \cdot \boldsymbol{h}\boldsymbol{h}^\top, \end{aligned} \tag{136}$$

where $\boldsymbol{h} = \boldsymbol{h}^{(\ell)}(\theta)$ and $\boldsymbol{h}^\star = \boldsymbol{h}^{(\ell)}(\theta^\star)$.

Then, we have

$$\begin{aligned} &\frac{\partial^2 f}{\partial\theta_{\ell,j_1}\partial\theta_{\ell,j_2}}(\theta^*) - \frac{\partial^2 f}{\partial\theta_{\ell,j_1}\partial\theta_{\ell,j_2}}(\theta) \\ =& \frac{1}{K^2}\mathbb{E}_{\boldsymbol{x}}\big[\mathcal{J}^\star_{\ell,k}\sigma'(\theta^{\star T}_{\ell,j_1}\boldsymbol{h}^\star)\mathcal{J}^\star_{\ell,k}\sigma'(\theta^{\star T}_{\ell,j_2}\boldsymbol{h}^\star)\boldsymbol{h}^\star\boldsymbol{h}^{\star\top} \\ &\qquad - \mathcal{J}_{\ell,k}\sigma'(\theta^\top_{\ell,j_1}\boldsymbol{h})\mathcal{J}_{\ell,k}\mathcal{J}_{\ell,k}\sigma'(\theta^\top_{\ell,j_2}\boldsymbol{h})\boldsymbol{h}\boldsymbol{h}^\top\big] \\ =& \frac{1}{K^2}\mathbb{E}_{\boldsymbol{x}}\big[\mathcal{J}^\star_{\ell,k}\sigma'(\theta^{\star T}_{\ell,j_1}\boldsymbol{h}^\star)\big(\mathcal{J}^\star_{\ell,k}\sigma'(\theta^{\star T}_{\ell,j_2}\boldsymbol{h}^\star)\boldsymbol{h}^\star\boldsymbol{h}^{\star\top} - \mathcal{J}_{\ell,k}\sigma'(\theta^\top_{\ell,j_2}\boldsymbol{h})\boldsymbol{h}\boldsymbol{h}^\top\big) \\ &\qquad + \mathcal{J}_{\ell,k}\sigma'(\theta^\top_{\ell,j_2}\boldsymbol{h})\big(\mathcal{J}^\star_{\ell,k}\sigma'(\theta^{\star T}_{\ell,j_1}\boldsymbol{h})\boldsymbol{h}^\star\boldsymbol{h}^{\star\top} - \mathcal{J}_{\ell,k}\sigma'(\theta^\top_{\ell,j_1}\boldsymbol{h})\boldsymbol{h}\boldsymbol{h}^\top\big)\big] \\ :=& \frac{1}{K^2}(\boldsymbol{P}_1 + \boldsymbol{P}_2). \end{aligned} \tag{137}$$

For any $\boldsymbol{a} \in \mathbb{R}^{K_\ell}$ with $\|\boldsymbol{a}\|_2 = 1$, we have

$$\boldsymbol{a}^\top \boldsymbol{P}_1 \boldsymbol{a} = \mathbb{E}_{\boldsymbol{x}} \mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_1} \boldsymbol{h}^\star) \Big(\mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}^\star)(\boldsymbol{a}^\top \boldsymbol{h}^\star)^2 - \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h})(\boldsymbol{a}^\top \boldsymbol{h})^2 \Big). \tag{138}$$

Then, we have

$$\begin{aligned}
|\boldsymbol{a}^\top \boldsymbol{P}_1 \boldsymbol{a}| =& \Big| \mathbb{E}_{\boldsymbol{x}} \mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_1} \boldsymbol{h}^\star) \Big(\mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}^\star)(\boldsymbol{a}^\top \boldsymbol{h}^\star)^2 - \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h})(\boldsymbol{a}^\top \boldsymbol{h})^2 \Big) \Big| \\
\leq & \mathbb{E}_{\boldsymbol{x}} \Big| \mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}^\star)(\boldsymbol{a}^\top \boldsymbol{h}^\star)^2 - \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h})(\boldsymbol{a}^\top \boldsymbol{h})^2 \Big| \\
\leq & \mathbb{E}_{\boldsymbol{x}} \Big| \mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}^\star)(\boldsymbol{a}^\top \boldsymbol{h}^\star)^2 - \mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h}^\star)(\boldsymbol{a}^\top \boldsymbol{h})^2 \Big| \\
& + \mathbb{E}_{\boldsymbol{x}} \Big| \mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h}^\star)(\boldsymbol{a}^\top \boldsymbol{h})^2 - \mathcal{J}_{\ell,k} \sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h})(\boldsymbol{a}^\top \boldsymbol{h})^2 \Big| \\
& + \mathbb{E}_{\boldsymbol{x}} \Big| \mathcal{J}_{\ell,k} \sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h})(\boldsymbol{a}^\top \boldsymbol{h})^2 - \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h})(\boldsymbol{a}^\top \boldsymbol{h})^2 \Big| \\
\lesssim & \|\theta - \theta^\star\|_2 + \|\theta - \theta^\star\|_2 + \mathbb{E}_{\boldsymbol{x}} \Big| \big(\sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h}) - \sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h}) \big) \cdot (\boldsymbol{a}^\top \boldsymbol{h})^2 \Big| \\
\lesssim & \|\theta - \theta^\star\|_2 + \mathbb{E}_{\boldsymbol{x}} \Big| \big(\sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h}) - \sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h}) \big) \cdot (\boldsymbol{a}^\top \boldsymbol{h})^2 \Big|.
\end{aligned} \tag{139}$$

Utilizing the Gram-Schmidt process, we can demonstrate the existence of a set of normalized orthonormal vectors denoted as $\mathcal{B} = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{a}_4^\perp, \cdots, \boldsymbol{a}_d^\perp\} \in \mathbb{R}^d$. This set forms an orthogonal and normalized basis for $\mathbb{R}^d$, wherein the subspace spanned by $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ includes $\boldsymbol{a}, \theta_{\ell,j_2}$, and θ^*_{ℓ,j_2}. Then, for any $\boldsymbol{x} \in \mathbb{R}^d$, we have a unique $\boldsymbol{z} = [z_1,\ z_2,\ \cdots,\ z_d]^\top$ such that

$$\boldsymbol{h} = z_1 \boldsymbol{a} + z_2 \boldsymbol{b} + z_3 \boldsymbol{c} + \cdots + z_d \boldsymbol{a}_d^\perp.$$

Because (i) $\boldsymbol{a}, \theta_{\ell,j_2}$, and θ^*_{ℓ,j_2} belongs to the subspace spanned by vectors $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ and (ii) $\boldsymbol{a}_4^\perp, \cdots, \boldsymbol{a}_d^\perp, \cdots$ are orthogonal to $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$. Then, we know that

$$\begin{aligned}
\theta^{\star \top}_{\ell,j_2} \boldsymbol{h} =& \theta^{\star \top}_{\ell,j_2} (z_1 \boldsymbol{a} + z_2 \boldsymbol{b} + z_3 \boldsymbol{c} + \cdots + z_d \boldsymbol{a}_d^\perp) \\
=& z_1 \theta^{\star \top}_{\ell,j_2} \boldsymbol{a} + z_2 \theta^{\star \top}_{\ell,j_2} \boldsymbol{b} + z_3 \theta^{\star \top}_{\ell,j_2} \boldsymbol{c} + \cdots + z_d \theta^{\star \top}_{\ell,j_2} \boldsymbol{a}_d^\perp \\
=& z_1 \theta^{\star \top}_{\ell,j_2} \boldsymbol{a} + z_2 \theta^{\star \top}_{\ell,j_2} \boldsymbol{b} + z_3 \theta^{\star \top}_{\ell,j_2} \boldsymbol{c} + 0 \\
=& \theta^{\star \top}_{\ell,j_2} (z_1 \boldsymbol{a} + z_2 \boldsymbol{b} + z_3 \boldsymbol{c}) \\
:=& \theta^{\star \top}_{\ell,j_2} \widetilde{\boldsymbol{h}}.
\end{aligned} \tag{140}$$

where $\widetilde{\boldsymbol{h}} = z_1 \boldsymbol{a} + z_2 \boldsymbol{b} + z_3 \boldsymbol{c}$. Similar to (140), we have $\theta^\top_{\ell,j_2} \boldsymbol{h} = \theta^\top_{\ell,j_2} \widetilde{\boldsymbol{h}}$ and $\boldsymbol{a}^\top \boldsymbol{h} = \boldsymbol{a}^\top \widetilde{\boldsymbol{h}}$.

Then, we define I_4 as

$$\begin{aligned}
I_4 :=& \mathbb{E}_{\boldsymbol{h}} \Big| \big(\sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h}) - \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h}) \big) \cdot \big(\boldsymbol{a}^\top \boldsymbol{h} \big) \Big| \\
=& \int_{\mathcal{R}_{\boldsymbol{h}}} |\sigma'\big(\theta^\top_{\ell,j_2} \boldsymbol{h}\big) - \sigma'\big(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}\big)| \cdot |\boldsymbol{a}^\top \boldsymbol{h}|^2 \cdot f_H(\boldsymbol{h}) d\boldsymbol{h} \\
=& \int_{\mathcal{R}_{\boldsymbol{z}}} |\sigma'\big(\theta^\top_{\ell,j_2} \boldsymbol{h}\big) - \sigma'\big(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}\big)| \cdot |\boldsymbol{a}^\top \boldsymbol{h}|^2 \cdot f_Z(\boldsymbol{z}) \cdot |\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{z})| d\boldsymbol{z}
\end{aligned} \tag{141}$$

where $|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{z})|$ is the determinant of the Jacobian matrix $\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}}$. Since $\boldsymbol{z}$ is a representation of $\boldsymbol{h}$ based on an orthogonal and normalized basis, we have $|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{z})| = 1$. According to (140), I_4 can be rewritten as

$$\begin{aligned}
I_4 =& \int_{\mathcal{R}_z} |\sigma'\big(\theta^\top_{\ell,j_2} \widetilde{\boldsymbol{h}}\big) - \sigma'\big(\theta^{\star T}_{\ell,j_2} \widetilde{\boldsymbol{h}}\big)| \cdot |\boldsymbol{a}^\top \widetilde{\boldsymbol{h}}|^2 \cdot f_Z(\boldsymbol{z}) d\boldsymbol{z} \\
=& \int_{\mathcal{R}_z} |\sigma'\big(\theta^\top_{\ell,j_2} \widetilde{\boldsymbol{h}}\big) - \sigma'\big(\theta^{\star T}_{\ell,j_2} \widetilde{\boldsymbol{h}}\big)| \cdot |\boldsymbol{a}^\top \widetilde{\boldsymbol{h}}|^2 \cdot f_Z(z_1, z_2, z_3) dz_1 dz_2 dz_3
\end{aligned} \tag{142}$$

where in the last equality we abuse $f_Z(z_1, z_2, z_3)$ to represent the probability density function of (z_1, z_2, z_3) defined in region $\mathcal{R}_z$.

Next, we show that $\boldsymbol{z}$ is rotational invariant over $\mathcal{R}_z$. Let $\boldsymbol{R} = [\boldsymbol{a}\ \boldsymbol{b}\ \boldsymbol{c}\ \cdots\ \boldsymbol{a}_d^{\perp}]$, we have $\boldsymbol{h} = \boldsymbol{R}\boldsymbol{z}$. For any $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ with $\|\boldsymbol{z}^{(1)}\|_2 = \|\boldsymbol{z}^{(2)}\|_2$. We define $\boldsymbol{h}^{(1)} = R\boldsymbol{z}^{(1)}$ and $\boldsymbol{h}^{(2)} = R\boldsymbol{z}^{(2)}$. Since $\boldsymbol{x}$ is rotational invariant and $\|\boldsymbol{h}^{(1)}\|_2 = \|\boldsymbol{h}^{(2)}\|_2 = \|\boldsymbol{z}^{(1)}\|_2 = \|\boldsymbol{z}^{(2)}\|_2$, then we know $\boldsymbol{h}^{(1)}$ and $\boldsymbol{h}^{(2)}$ has the same distribution density. Then, $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ has the same distribution density as well. Therefore, $\boldsymbol{z}$ is rotational invariant over $\mathcal{R}_z$.

Then, we consider spherical coordinates with $z_1 = Rcos\sigma_1, z_2 = Rsin\sigma_1 sin\sigma_2, z_3 = Rsin\sigma_1 cos\sigma_2$. Hence, we have

$$I_4 = \int |\sigma'\big(\theta_{\ell,j_2}^{\top}\widetilde{\boldsymbol{h}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star\top}\widetilde{\boldsymbol{h}}\big)| \cdot |R\cos\sigma_1|^2 \cdot f_Z(R,\sigma_1,\sigma_2) \cdot R^2\sin\sigma_1 \cdot dRd\sigma_1 d\sigma_2. \tag{143}$$

Since $\boldsymbol{z}$ is rotational invariant, we have that

$$f_Z(R,\sigma_1,\sigma_2) = f_Z(R). \tag{144}$$

Then, we have

$$\begin{aligned} I_4 &= \int |\sigma'\big(\theta_{\ell,j_2}^{\top}(\widetilde{\boldsymbol{h}}/R)\big) - \sigma'\big(\theta_{\ell,j_2}^{\star T}(\widetilde{\boldsymbol{h}}/R)\big)| \cdot |R\cos\sigma_1|^2 \cdot f_Z(R)R^2\sin\sigma_1 dRd\sigma_1 d\sigma_2 \\ &= \int_0^{\infty} R^4 f_z(R)dR \int_0^{\psi_1(R)}\int_0^{\psi_2(R)} |\cos\sigma_1|^2 \cdot \sin\sigma_1 \\ &\quad \cdot |\sigma'\big(\theta_{\ell,j_2}^{\top}(\widetilde{\boldsymbol{h}}/R)\big) - \sigma'\big(\theta_{\ell,j_2}^{\star T}(\widetilde{\boldsymbol{h}}/R)\big)|d\sigma_1 d\sigma_2 \\ &\leq \int_0^{\infty} R^4 f_z(R)dR \int_0^{\pi}\int_0^{2\pi} \sin\sigma_1 \cdot |\sigma'\big(\theta_{\ell,j_2}^{\top}\bar{\boldsymbol{x}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star T}\bar{\boldsymbol{x}}\big)|d\sigma_1 d\sigma_2, \end{aligned} \tag{145}$$

where the first equality holds because $\sigma'\big(\theta_{i,,j_2}^{\top}\boldsymbol{h}\big)$ only depends on the direction of $\boldsymbol{h}$, and $\bar{\boldsymbol{x}} := \boldsymbol{h}/R = (\cos\sigma_1, \sin\sigma_1\sin\sigma_2, \sin\sigma_1\cos\sigma_2)$ in the last inequality.

Because $\boldsymbol{z}$ belongs to the sub-Gaussian distribution, we have $F_z(R) \geq 1 - 2e^{-\frac{R^2}{\sigma^2}}$ for some constant $\sigma > 0$. Then, the integration of R can be represented as

$$\begin{aligned} \int_0^{\infty} R^4 f_Z(R)dR &= \int_0^{\infty} R^4 d\big(1 - F_z(R)\big) \\ &\leq \int_0^{\infty} 4R^3\big(1 - F_z(R)\big)dR \\ &\leq \int_0^{\infty} 8R^3 e^{-\frac{R^2}{\sigma^2}}dR \\ &\leq \frac{32}{\sqrt{2\pi}}\sigma\int_0^{\infty} R^2 e^{-\frac{R^2}{\sigma^2}}dR \\ &= 32\sigma^2\int_0^{\infty} R^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{R^2}{\sigma^2}}dR, \end{aligned} \tag{146}$$

where the last inequality comes from the calculation that

$$\begin{aligned} \int_0^{\infty} 2R^2 e^{-\frac{R^2}{\sigma^2}}dR &= \sqrt{2\pi}\sigma^3, \\ \int_0^{\infty} 2R^3 e^{-\frac{R^2}{\sigma^2}}dR &= 4\sigma^4. \end{aligned} \tag{147}$$

Then, we define $\widetilde{\boldsymbol{x}} \in \mathbb{R}^{K_\ell}$ belongs to Gaussian distribution as $\widetilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Therefore, we have

$$\begin{aligned} I_4 &\leq 32\sigma^2 \cdot \int_0^{\infty} R^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{R^2}{\sigma^2}}dR \int_0^{\pi}\int_0^{2\pi} \sin\sigma_1 \cdot |\sigma'\big(\theta_{\ell,j_2}^{\top}\bar{\boldsymbol{x}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star\top}\bar{\boldsymbol{x}}\big)|d\sigma_1 d\sigma_2 \\ &= 32\sigma^2 \cdot \mathbb{E}_{z_1,z_2,z_3}\big|\sigma'\big(\theta_{\ell,j_2}^{\top}\widetilde{\boldsymbol{x}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star\top}\widetilde{\boldsymbol{x}}\big)| \\ &\eqsim \mathbb{E}_{\widetilde{\boldsymbol{x}}}\big|\sigma'\big(\theta_{\ell,j_2}^{\top}\widetilde{\boldsymbol{x}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}}\big)|, \end{aligned} \tag{148}$$

where $\widetilde{\boldsymbol{x}}$ belongs to Gaussian distribution.

Therefore, the inequality bound over a sub-Gaussian distribution is bounded by the one over a Gaussian distribution. In the following contexts, we provide the upper bound of $\mathbb{E}_{\widetilde{\boldsymbol{x}}}\big|\sigma'\big(\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}}\big)\big|$.

Define a set $\mathcal{A}_1 = \{\boldsymbol{x}|(\theta_{\ell,j_2}^{\star\top}\widetilde{\boldsymbol{x}})(\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}}) < 0\}$. If $\widetilde{\boldsymbol{x}} \in \mathcal{A}_1$, then $\theta_{\ell,j_2}^{\star\top}\widetilde{\boldsymbol{x}}$ and $\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}}$ have different signs, which means the value of $\sigma'(\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}})$ and $\sigma'(\theta_{\ell,j_2}^{\star\top}\widetilde{\boldsymbol{x}})$ are different. This is equivalent to say that

$$|\sigma'(\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}}) - \sigma'(\theta_{\ell,j_2}^{\star\top}\widetilde{\boldsymbol{x}})| = \begin{cases} 1, \text{ if } \widetilde{\boldsymbol{x}} \in \mathcal{A}_1 \\ 0, \text{ if } \widetilde{\boldsymbol{x}} \in \mathcal{A}_1^c \end{cases}. \tag{149}$$

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

$$|\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}}| \leq |\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}} - \theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}}| \leq \|\theta_{\ell,j_2}^\star - \theta_{\ell,j_2}\|_2 \cdot \|\widetilde{\boldsymbol{x}}\|_2. \tag{150}$$

Let us define a set $\mathcal{A}_2$ such that

$$\begin{aligned}\mathcal{A}_2 =& \Big\{\widetilde{\boldsymbol{x}}\Big|\frac{|\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}}|}{\|\theta_{\ell,j_2}^*\|_2\|\widetilde{\boldsymbol{x}}\|_2} \leq \frac{\|\theta_{\ell,j_2}^* - \theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^*\|_2}\Big\} \\ =& \Big\{\theta_{\widetilde{\boldsymbol{x}},\theta_{\ell,j_2}^*}\Big||\cos\theta_{\widetilde{\boldsymbol{x}},\theta_{\ell,j_2}^\star}| \leq \frac{\|\theta_{\ell,j_2}^\star - \theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2}\Big\}.\end{aligned} \tag{151}$$

Hence, we have that

$$\begin{aligned}\mathbb{E}_{\widetilde{\boldsymbol{x}}}|\sigma'(\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}}) - \sigma'(\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}})|^2 =& \mathbb{E}_{\widetilde{\boldsymbol{x}}}|\sigma'(\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}}) - \sigma'(\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}})| \\ =& \text{Prob}(\widetilde{\boldsymbol{x}} \in \mathcal{A}_1) \\ \leq& \text{Prob}(\widetilde{\boldsymbol{x}} \in \mathcal{A}_2).\end{aligned} \tag{152}$$

Since $\widetilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \|\boldsymbol{a}\|_2^2\boldsymbol{I})$, $\theta_{\widetilde{\boldsymbol{x}},\theta_{\ell,j_2}^\star}$ belongs to the uniform distribution on $[-\pi,\pi]$, we have

$$\begin{aligned}\text{Prob}(\widetilde{\boldsymbol{x}} \in \mathcal{A}_2) = \frac{\pi - \arccos\frac{\|\theta_{\ell,j_2}^\star - \theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2}}{\pi} \leq& \frac{1}{\pi}\tan(\pi - \arccos\frac{\|\theta_{\ell,j_2}^\star - \theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2}) \\ =& \frac{1}{\pi}\cot(\arccos\frac{\|\theta_{\ell,j_2}^\star - \theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2}) \\ \leq& \frac{2}{\pi}\frac{\|\theta_{\ell,j_2}^\star - \theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2} \\ \leq& \|\theta_\ell^\star - \theta_\ell\|_2\end{aligned} \tag{153}$$

Hence, (145) and (153) suggest that

$$\begin{aligned} I_4 &\lesssim \|\theta_i - \theta_i^\star\|_2 \cdot \|\boldsymbol{a}\|_2^2, \\ \text{and} \qquad \|\boldsymbol{P}_1\|_2 &\leq \|\theta - \theta^\star\|_2 + I_4 \lesssim \|\theta - \theta^\star\|_2, \end{aligned} \tag{154}$$

The same bound that is shown in (154) holds for $\boldsymbol{P}_2$ as well.

Therefore, we have

$$\begin{aligned}\|\nabla_\ell^2 f(\theta^\star) - \nabla_\ell^2 f(\theta)\|_2 =& \max_{\|\boldsymbol{\alpha}\|_2\leq 1}\Big|\boldsymbol{\alpha}^\top\Big(\nabla_\ell^2 f(\theta^\star) - \nabla_\ell^2 f(\theta)\Big)\boldsymbol{\alpha}\Big| \\ \leq& \frac{1}{K^2}\sum_{j_1=1}^K\sum_{j_2=1}^K \|\boldsymbol{P}_1 + \boldsymbol{P}_2\|_2 \cdot \|\boldsymbol{\alpha}_{j_1}\|_2 \cdot \|\boldsymbol{\alpha}_{j_2}\|_2 \\ \lesssim& \frac{1}{K^2}\cdot\sum_{j_1=1}^K\sum_{j_2=1}^K \|\theta - \theta^\star\|_2 \cdot \|\boldsymbol{\alpha}_{j_1}\|_2\|\boldsymbol{\alpha}_{j_2}\|_2 \\ \lesssim& \frac{1}{K^2}\cdot\sum_{j_1=1}^K\sum_{j_2=1}^K \|\theta - \theta^\star\|_2 \cdot \Big(\frac{\|\boldsymbol{\alpha}_{j_1}\|_2^2 + \|\boldsymbol{\alpha}_{j_2}\|_2^2}{2}\Big) \\ \lesssim& \frac{1}{K}\cdot\|\theta^\star - \theta\|_2,\end{aligned} \tag{155}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{Kd}$ and $\boldsymbol{\alpha}_j \in \mathbb{R}^{K_\ell}$ with $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^\top, \boldsymbol{\alpha}_2^\top, \cdots, \boldsymbol{\alpha}_K^\top]^\top$. □

H.2 PROOF OF LEMMA 11

We aim to prove that $\int_{\mathcal{R}} \Big(\sum_{j=1}^{K} \boldsymbol{\alpha}^\top \boldsymbol{h} \sigma'(\theta_{\ell,j}^\top \boldsymbol{h}) \Big)^2 p_H(\boldsymbol{h}) \cdot d\boldsymbol{h}$ is strictly greater than zero for any $\boldsymbol{\alpha}$. Therefore, the ρ_1 in (6) is strictly greater than zero. The proof is inspired by Theorem 3.1 in (Du et al., 2019). It is obviously that $(\sum_{j=1}^{K} \boldsymbol{\alpha}^\top \boldsymbol{h} \sigma'(\theta_{\ell,j}^\top \boldsymbol{h}))^2$ is greater or equal to zero. Given $(\sum_{j=1}^{K} \boldsymbol{\alpha}^\top \boldsymbol{h} \sigma'(\theta_{\ell,j}^\top \boldsymbol{h}))^2$ is continuous, we only need to show that α such that $\sum_{j=1}^{K} \boldsymbol{\alpha}^\top \boldsymbol{h} \sigma'(\theta_{\ell,j}^\top \boldsymbol{h}) \neq 0$ for any α, namely, $\{\boldsymbol{h}\sigma'(\theta_{\ell,j}^\top \boldsymbol{h})\}_{j=1}^{K}$ are linear independent.

Proof of Lemma 11. Let $\mathcal{H}$ be a Hilbert space on $\mathbb{R}^{K_\ell}$, and the inner product of $\mathcal{H}$ is defined as

$$\langle f, g \rangle = \int_{\mathcal{R}} f(\boldsymbol{h})^\top g(\boldsymbol{h}) f_H(\boldsymbol{h}) \cdot d\boldsymbol{h}, \quad \forall f, g \in \mathcal{H}, \tag{156}$$

where the Lebesgue measure of $\mathcal{R}$ over $\mathbb{R}^{K_\ell}$ is non-zero. Instead of directly proving $\int_{\mathcal{R}} \Big(\sum_{k=1}^{K} \boldsymbol{\alpha}^\top \boldsymbol{h} \sigma'(\theta_k^\top \boldsymbol{h}) \Big)^2 f_H(\boldsymbol{h}) \cdot d\boldsymbol{h} > 0$ for any $\boldsymbol{\alpha}$, we note that it is sufficient to prove that $\{\boldsymbol{h}\sigma'(\theta_k^\top \boldsymbol{h})\}_{k\in[K]}$ are linear independent over the Hilbert space $\mathcal{H}$. Namely, if $\{\boldsymbol{h}\sigma'(\theta_k^\top \boldsymbol{h})\}_{k\in[K]}$ are linear independent, we have

$$\boldsymbol{\alpha}^\top \boldsymbol{h} \sigma'(\theta_k^\top \boldsymbol{h}) \neq 0 \quad \textit{almost everywhere}. \tag{157}$$

Therefore, we can know that $\int_{\mathcal{R}} \Big(\sum_{j=1}^{K} \boldsymbol{\alpha}^\top \boldsymbol{h} \sigma'(\theta_{\ell,j}^\top \boldsymbol{h}) \Big)^2 p_H(\boldsymbol{h}) \cdot d\boldsymbol{h}$ is strictly greater than zero.

Next, we provide the whole proof for that $\{x\sigma'(\theta_k^\top \boldsymbol{h})\}_{k\in[K]}$ are linear independent over the Hilbert space $\mathcal{H}$.

We define a group of functions $\{\psi_j(\boldsymbol{h})\}_{j=1}^{K}$, where $\psi_j(\boldsymbol{h}) = \boldsymbol{h}\sigma'(\theta_j^\top \boldsymbol{h})$. From the assumption in Lemma 11, we can justify that $\mathbb{E}_{\boldsymbol{h}\sim\mathcal{D}} |\psi_j(\boldsymbol{h})|^2 \leq \mathbb{E}_{\boldsymbol{h}\sim\mathcal{D}} |\boldsymbol{h}|^2 < \infty$.

Let $\mathcal{X}_i = \{\boldsymbol{h} \mid \theta_i^\top \boldsymbol{h} = 0\}$ for any $i \in [K]$. For any fixed k, we can justify that $\mathcal{X}_k$ cannot be covered by other sets $\{\mathcal{X}_k\}_{j\neq k}$ as long as θ_k does not parallel to any other weights θ_j with $j \neq k$. Namely, $\mathcal{X}_k \not\subset \cup_{j\neq k} \mathcal{X}_j$. The idea of proving the claim above is that the intersection of $\mathcal{X}_j$ and $\mathcal{X}_k$ is only a hyperplane in $\mathcal{X}_k$. The union of finite many hyperplanes is not even a measurable space and thus cannot cover the original space. Formally, we provide the formal proof for this claim as follows.

Let λ be the Lebesgue measure on $\mathcal{X}_k$, then $\lambda(\mathcal{X}_k) > 0$. When θ_j does not parallel to θ_k, $\mathcal{X}_k \cap \mathcal{X}_j$ is only a hyperplane in $\mathcal{X}_k$ for $j \neq k$. Hence, we have $\lambda(\mathcal{X}_j \cap \mathcal{X}_k) = 0$. Next, we have

$$\lambda\big(\mathcal{X}_k \cap (\cup_{j\neq k} \mathcal{X}_k)\big) \leq \sum_{j\neq k} \lambda(\mathcal{X}_k \cap \mathcal{X}_j) = 0. \tag{158}$$

Therefore, we have

$$\lambda\big(\mathcal{X}_k / (\cup_{j\neq k} \mathcal{X}_k)\big) = \lambda(\mathcal{X}_k) - \lambda\big(\mathcal{X}_k \cap (\cup_{j\neq k} \mathcal{X}_k)\big) = \lambda(\mathcal{X}_k) > 0. \tag{159}$$

Therefore, we have $\mathcal{X}_k / (\cup_{j\neq k} \mathcal{X}_j)$ is not empty, which means that $\mathcal{X}_k \not\subset \cup_{j\neq k} \mathcal{X}_j$.

Next, Since $\mathcal{X}_k / (\cup_{j\neq k} \mathcal{X}_j)$ is not an empty set, there exists a point $\boldsymbol{z}_k \in \mathcal{X}_k / (\cup_{j\neq k} \mathcal{X}_j)$ and $r_0 > 0$ such that

$$\mathcal{B}(\boldsymbol{z}_k, r) \cap \mathcal{D}_j = \emptyset \quad \textit{with} \quad \forall r \leq r_0 \textit{ and } j \neq k, \tag{160}$$

where $\mathcal{B}(\boldsymbol{z}_k, r)$ stands for a ball centered at $\boldsymbol{z}_k$ with a radius of r. Then, we divide $\mathcal{B}(\boldsymbol{z}_k, r)$ into two disjoint subsets such that

$$\begin{aligned} \mathcal{B}_r^+ &= \mathcal{B}(\boldsymbol{z}_k, r) \cap \{\boldsymbol{h} \mid \theta_k^\top \boldsymbol{h} > 0\}, \\ \mathcal{B}_r^- &= \mathcal{B}(\boldsymbol{z}_k, r) \cap \{\boldsymbol{h} \mid \theta_k^\top \boldsymbol{h} < 0\}. \end{aligned} \tag{161}$$

Because $\boldsymbol{z}_k$ is a boundary point of $\{\boldsymbol{h} | \theta_k^\top \boldsymbol{h} = 0\}$, both $\mathcal{B}_r^+$ and $\mathcal{B}_r^-$ are non-empty.

Note that $\psi_j(\boldsymbol{h})$ is continuous at any point except for the ones in $\mathcal{X}_j$. Then, for any $j \neq k$, we know that $\sigma_j(\theta_k^\top \boldsymbol{h})$ is continuous at point $\boldsymbol{z}_k$ since $\boldsymbol{z}_k \notin \mathcal{X}_j$. Hence, it is easy to verify that

$$\lim_{r\to 0_+} \frac{1}{\lambda(\mathcal{B}_r^+)} \int_{\mathcal{B}_r^+} \psi_k(\boldsymbol{h}) d\boldsymbol{h} = \lim_{r\to 0_-} \frac{1}{\lambda(\mathcal{B}_r^-)} \int_{\mathcal{B}_r^+} \psi_k(\boldsymbol{h}) d\boldsymbol{h} = \psi_k(\boldsymbol{z}_k). \tag{162}$$

While for ψ_k, we know that $\psi_k(\boldsymbol{h}) \equiv 0$ for $\boldsymbol{h} \in \mathcal{B}_r^-$, (ii) $\psi_k(\boldsymbol{h}) = \boldsymbol{h}$ for $\boldsymbol{h} \in \mathcal{B}_r^+$. Hence, it is easy to verify that

$$\begin{aligned}\lim_{r\to 0_+} \frac{1}{\lambda(\mathcal{B}_r^+)} \int_{\mathcal{B}_r^+} \psi_k(\boldsymbol{h}) d\boldsymbol{h} = \boldsymbol{z}_k \\ \lim_{r\to 0_-} \frac{1}{\lambda(\mathcal{B}_r^-)} \int_{\mathcal{B}_r^+} \psi_k(\boldsymbol{h}) d\boldsymbol{h} = 0.\end{aligned} \tag{163}$$

Now let us proof that $\{\psi_j\}_{j=1}^K$ are linear independent by contradiction. Suppose $\{\psi_j\}_{j=1}^K$ are linear dependent, we have

$$\sum_{j=1}^K \alpha_j \psi_j(\boldsymbol{h}) \equiv 0, \quad \forall \boldsymbol{h}. \tag{164}$$

Then, we have

$$\begin{aligned}\lim_{r\to 0_+} \frac{1}{\lambda(\mathcal{B}_r^+)} \int_{\mathcal{B}_r^+} \sum_{j=1}^K \alpha_j \psi_j(\boldsymbol{h}) d\boldsymbol{h} = 0 \\ \lim_{r\to 0_+} \frac{1}{\lambda(\mathcal{B}_r^-)} \int_{\mathcal{B}_r^+} \sum_{j=1}^K \alpha_j \psi_j(\boldsymbol{h}) d\boldsymbol{h} = 0\end{aligned} \tag{165}$$

Then, we have

$$\begin{aligned}0 =& \lim_{r\to 0_+} \frac{1}{\lambda(\mathcal{B}_r^+)} \int_{\mathcal{B}_r^+} \sum_{j=1}^K \alpha_j \psi_j(\boldsymbol{h}) d\boldsymbol{h} - \lim_{r\to 0_+} \frac{1}{\lambda(\mathcal{B}_r^-)} \int_{\mathcal{B}_r^+} \sum_{j=1}^K \alpha_j \psi_j(\boldsymbol{h}) d\boldsymbol{h} \\ =& \alpha_k \boldsymbol{z}_k\end{aligned} \tag{166}$$

where the last equality comes from (162) and (163).

Note that $\boldsymbol{z}_k$ cannot be $\mathbf{0}$ because $\boldsymbol{z}_k \notin \mathcal{X}_j$. Therefore, we have $\alpha_k = 0$. Similarly to (166), we can obtain that $\alpha_j = 0$ by define $\boldsymbol{z}_j$ following the definition of $\boldsymbol{z}_k$ for any $j \in [K]$. Then, we know that (164) holds if and only if $\boldsymbol{\alpha} = \mathbf{0}$, which contradicts the assumption that $\{\psi_j\}_{j=1}^K$ are linear dependent.

In conclusion, we know that $\{\psi_j\}_{j=1}^K$ are linear independent, and $\int_{\mathcal{R}} \left(\sum_{j=1}^K \boldsymbol{\alpha}^\top \boldsymbol{h} \sigma'(\theta_{\ell,j}^\top \boldsymbol{h}) \right)^2 p_H(\boldsymbol{h}) \cdot d\boldsymbol{h}$ is strictly greater than zero. □

H.3 Proof of Lemma 12

Proof of Lemma 12. From the definition of (37), we have

$$\begin{aligned}&\|\boldsymbol{h}^{(\ell)}(\theta) - \boldsymbol{h}^{(\ell)}(\theta^\star)\|_2 \\ =& \|\sigma\big(\theta_{\ell-1}^\top \boldsymbol{h}^{(\ell-1)}(\theta)\big) - \sigma\big(\theta_{\ell-1}^{\star\top} \boldsymbol{h}^{(\ell-1)}(\theta^\star)\big)\|_2 \\ =& \|\sigma\big(\theta_{\ell-1}^\top \boldsymbol{h}^{(\ell-1)}(\theta)\big) - \sigma\big(\theta_{\ell-1}^{\star\top} \boldsymbol{h}^{(\ell-1)}(\theta)\big) + \sigma\big(\theta_{\ell-1}^{\star\top} \boldsymbol{h}^{(\ell-1)}(\theta)\big) - \sigma\big(\theta_{\ell-1}^{\star\top} \boldsymbol{h}^{(\ell-1)}(\theta^\star)\big)\|_2 \\ \leq& \|\sigma\big(\theta_{\ell-1}^\top \boldsymbol{h}^{(\ell-1)}(\theta)\big) - \sigma\big(\theta_{\ell-1}^{\star\top} \boldsymbol{h}^{(\ell-1)}(\theta)\big)\|_2 + \|\sigma\big(\theta_{\ell-1}^{\star\top} \boldsymbol{h}^{(\ell-1)}(\theta)\big) - \sigma\big(\theta_{\ell-1}^{\star\top} \boldsymbol{h}^{(\ell-1)}(\theta^\star)\big)\|_2 \\ \leq& \|\theta_{\ell-1} - \theta_{\ell-1}^\star\|_2 \cdot \|\boldsymbol{h}^{(\ell-1)}(\theta)\|_2 + \|\boldsymbol{h}^{(\ell-1)}(\theta) - \boldsymbol{h}^{(\ell-1)}(\theta^\star)\|_2.\end{aligned} \tag{167}$$

With the assumption in the Lemma 12 such that θ is close enough to $\theta^\star$, we have

$$\|\theta_i\|_2 \leq \|\theta_i^\star\|_2 + \|\theta_i - \theta_i^\star\|_2 \lesssim 1. \tag{168}$$

Therefore, we have

$$\|\boldsymbol{h}^{(i)}(\theta)\|_2 \leq \|\theta_i\|_2 \cdots \|\theta_1\|_2 \cdot \|\boldsymbol{x}\|_2 \lesssim \|\boldsymbol{x}\|_2. \tag{169}$$

Then, we have

$$
\begin{aligned}
&\|\boldsymbol{h}^{(\ell)}(\theta) - \boldsymbol{h}^{(\ell)}(\theta^\star)\|_2 \\
\leq &\|\theta_{\ell-1} - \theta^\star_{\ell-1}\|_2 \cdot \|\boldsymbol{x}\|_2 + \|\boldsymbol{h}^{(\ell-1)}(\theta) - \boldsymbol{h}^{(\ell-1)}(\theta^\star)\|_2 \\
\leq &\sum_{i=1}^{\ell-1} \|\theta_i - \theta_i^\star\|_2 \cdot \|\boldsymbol{x}\|_2 + \|\boldsymbol{h}^{(1)}(\theta) - \boldsymbol{h}^{(1)}(\theta^\star)\|_2 \\
= &\sum_{i=1}^{\ell-1} \|\theta_i - \theta_i^\star\|_2 \cdot \|\boldsymbol{x}\|_2 + \|\boldsymbol{x} - \boldsymbol{x}\|_2 \\
= &\sum_{i=1}^{\ell-1} \|\theta_i - \theta_i^\star\|_2 \cdot \|\boldsymbol{h}^{(i-1)}(\theta)\|_2 \\
\leq &\|\theta - \theta^\star\|_2 \cdot \|\boldsymbol{x}\|_2,
\end{aligned}
\tag{170}
$$

which completes the proof. □