SF-DQN: Provable Knowledge Transfer using Successor Feature for Deep Reinforcement Learning

Shuai Zhang [1] **Heshan Devaka Fernando** [2] **Miao Liu** [3] **Keerthiram Murugesan** [3] **Songtao Lu** [3] **Pin-Yu Chen** [3] **Tianyi Chen** [2] **Meng Wang** [2]

Abstract

This paper studies the transfer reinforcement learning (RL) problem where multiple RL problems have different reward functions but share the same underlying transition dynamics. In this setting, the Q-function of each RL problem (task) can be decomposed into a successor feature (SF) and a reward mapping: the former characterizes the transition dynamics, and the latter characterizes the task-specific reward function. This Q-function decomposition, coupled with a policy improvement operator known as generalized policy improvement (GPI), reduces the sample complexity of finding the optimal Q-function, and thus the SF & GPI framework exhibits promising empirical performance compared to traditional RL methods like Q-learning. However, its theoretical foundations remain largely unestablished, especially when learning the successor features using deep neural networks (SF-DQN). This paper studies the provable knowledge transfer using SFs-DQN in transfer RL problems. We establish the first convergence analysis with provable generalization guarantees for SF-DQN with GPI. The theory reveals that SF-DQN with GPI outperforms conventional RL approaches, such as deep Q-network, in terms of both faster convergence rate and better generalization. Numerical experiments on real and synthetic RL tasks support the superior performance of SF-DQN & GPI, aligning with our theoretical findings.

[1]Department of Data Science, New Jersey Institute of Technology, Newark, NJ [2]Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY [3]IBM Thomas J. Watson Research Center, IBM Research, Yorktown Heights, NY. Correspondence to: Meng Wang <wangm7@rpi.edu>, Shuai Zhang <sz457@njit.edu>, Miao Liu <miao.liu1@ibm.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

In reinforcement learning (RL), the goal is to train an agent to perform a task within an environment in a desirable manner by allowing the agent to interact with the environment. Here, the agent is guided towards the desirable behavior by the rewards, and the optimal policy is derived from a learned value function (Q-function) in selecting the best actions to maximize the immediate and future rewards. This framework can effectively capture a wide array of real-world applications, such as gaming (Mnih et al., 2013; Silver et al., 2017), robotics (Kalashnikov et al., 2018), autonomous vehicles (Shalev-Shwartz et al., 2016; Schwarting et al., 2018), healthcare (Coronato et al., 2020), and natural language processing (Tenney et al., 2018). However, RL agents often need numerous interactions with the environment to manage complex tasks, especially when RL is equipped with deep neural networks (DNNs). For example, AlphaGo (Silver et al., 2017) required 29 million matches and 5000 TPUs at a cost exceeding $35 million, which is time-consuming and memory-intensive. Nevertheless, many complex real-world problems can naturally decompose into multiple interrelated sub-problems, all sharing the same environmental dynamics (Sutton et al., 1999; Bacon et al., 2017; Kulkarni et al., 2016a). In such scenarios, it becomes highly advantageous for an agent to harness knowledge acquired from previous tasks to enhance its performance in tackling new but related challenges. This practice of leveraging knowledge from one task to improve performance in others is known as transfer learning (Lazaric, 2012; Taylor & Stone, 2009; Barreto et al., 2017).

This paper focuses on an RL setting with learning multiple tasks, where each task is associated with a different reward function but shares the same environment. This setting naturally arises in many real-world applications such as robotics (Yu et al., 2020). We consider exploring the knowledge transfer among multiple tasks via the successor feature (SF) framework (Barreto et al., 2017) which disentangles the environment dynamic from the reward function at an incremental computational cost. The SF framework is derived from successor representation (SR) (Dayan, 1993) by introducing the value function approximation. Specifi-

cally, SR (Dayan, 1993) decouples the value function into a future state occupancy measure and a reward mapping. Here, the future state occupancy characterizes the transition dynamics of the environment, and the reward mapping characterizes the reward function of the task. SF is a natural application of SR in solving value function approximation. Furthermore, (Barreto et al., 2017) propose a generalization of the classic policy improvement, termed generalized policy improvement (GPI), enabling smooth knowledge transfer across learned policies. In contrast to traditional policy improvement, which typically considers only a single policy, Generalized Policy Improvement (GPI) operates by maintaining a set of policies, each associated with a distinct skill the agent has acquired. This approach enables the agent to switch among these policies based on the current state or task requirements, providing a flexible and adaptive framework for decision-making. Empirical findings presented in (Barreto et al., 2017) highlight the superior transfer performance of SF & GPI in deep RL when compared to conventional methods like Deep Q-Networks (DQNs). Subsequent works further justified the improved performance of SF in subgoal identification (Kulkarni et al., 2016b) and real-world robot navigation (Zhang et al., 2017).

Focus of this paper. While performance guarantees of SF-based learning are provided in the simple tabular setting (Barreto et al., 2017; 2018), less is known for such approaches in the widely used function approximation setting, especially for non-linear models including DNNs. In this context, this paper aims to close this gap by providing theoretical guarantees for SF learning in the context of DNNs. Our objective is to explore the convergence and generalization analysis of SF when paired with DNN approximation. We also seek to delineate the conditions under which SF learning can offer more effective knowledge transfer among tasks when contrasted with classical deep reinforcement learning (DRL) approaches.

Contributions. This paper presents the first convergence analysis with generalization guarantees for successor feature learning with deep neural network approximation (SF-DQN). This paper focuses on estimating the optimal Q-value function through the successor feature decomposition, where the successor feature decomposition component is approximated through a deep neural network. The paper offers a comprehensive analysis of the convergence of deep Q-networks with successor feature decomposition and provides insights into the improved performance of the learned Q-value function derived from successor feature decomposition. The key contributions are as follows:

(C1) The convergence analysis of the proposed SF-DQN to the optimal Q-function with generalization guarantees. By decomposing the reward into a linear combination of the transition feature and reward mapping, we demonstrate that the optimal Q-function can be learned by alternately updating the reward mapping and the successor feature using the collected data in online RL, where the corresponding successor feature is parameterized by a deep neural network. The learned Q-function converges to the optimal Q-function with generalization guarantees at a rate of $1/T$, where T is the number of iterations in updating transition features and reward mappings.

(C2) The theoretical characterization of enhanced performance by leveraging knowledge from previous tasks through GPI. This paper characterizes the convergence rate with generalization guarantees in transfer RL utilizing GPI. The convergence rate accelerates following the degree of correlation between the source and target tasks.

(C3) The theoretical characterization of the superior transfer learning performance with SF-DQN over non-representation learning approach DQNs. This paper quantifies the transfer learning ability of SF-DQN and DQN algorithms by evaluating their generalization error when transferring knowledge from one task to another. Our results indicate that SF-DQN achieves improved generalization compared to DQN, demonstrating the superiority of SF-DQN in transfer RL.

1.1. Related Works

Successor features in RL. In pioneering works, (Dayan, 1993) introduced the concept of SR, demonstrating that the value function can be decomposed into a reward mapping and a state representation that measures the future state occupancy from a given state, with learning feasibility proof in tabular settings. Subsequently, (Barreto et al., 2017) extended SR from three perspectives: (1) the feature domain of SR is extended from states to state-action pairs, known as SF; (2) DNNs are deployed as function approximators to represent the SF and reward mappings; (3) GPI algorithm is introduced to accelerate policy transfer for multi-tasks. Furthermore, (Kulkarni et al., 2016b; Zhang et al., 2017) apply SF learning with DNN-based schemes to subgoal identification (Kulkarni et al., 2016b) and robot navigation (Zhang et al., 2017). However, only (Barreto et al., 2017; 2018) provided transfer guarantees for Q-learning with SF and GPI for the tabular case under the assumption that the Q-function from the source task is well-estimated. However, to the best of our knowledge, none of works have provided any theoretical guarantees of SF in the function approximation with neural networks. In addition, instead of assuming that the Q-function of the source task is well estimated, our paper offers both convergence analysis and sample complexity for successor feature learning in both the source task training stage and transfer learning stages. We refer readers to a comprehensive comparison of rein-

forcement learning transfer using Successor Features, as detailed in (Zhu et al., 2023).

RL with neural networks. Recent advancements of theoretical analysis in RL with neural network approximation mainly include the Bellman Eluder dimension (Jiang et al., 2017; Russo & Van Roy, 2013), Neural Tangent Kernel (NTK) (Yang et al., 2020; Cai et al., 2019; Xu & Gu, 2020; Du et al., 2020), and Besov regularity (Suzuki, 2019; Ji et al., 2022; Nguyen-Tang et al., 2022). However, each of these frameworks has its limitations. The Eluder dimension exhibits exponential growth even for shallow neural networks (Dong et al., 2021), making it challenging to characterize sample complexity in real-world applications of DRL. The NTK framework linearizes DNNs to bypass the non-convexity derived from the non-linear activation function in neural networks. Nevertheless, it requires using computationally inefficient, extremely wide neural networks (Yang et al., 2020). Moreover, the NTK approach falls short in explaining the advantages of utilizing nonlinear neural networks over linear function approximation (Liu et al., 2022; Fan et al., 2020). The Besov space framework (Ji et al., 2022; Nguyen-Tang et al., 2022; Liu et al., 2022; Fan et al., 2020) requires sparsity on neural networks and makes the impractical assumption that the algorithm can effectively identify the global optimum, which is unfeasible for non-convex optimization involving NNs.

Theory of generalization in deep learning. The theory of generalization in deep learning has been extensively developed in supervised learning, where labeled data is available throughout training. Generalization in learned models necessitates low training error and small generalization gap. However, in DNNs, training errors and generalization gaps are analyzed separately due to their non-convex nature. To ensure bounded generalization, it is common to focus on *one-hidden-layer* neural networks (Safran & Shamir, 2018) in convergence analysis. Existing theoretical analysis tools in supervised learning with generalization guarantees draw heavily from various frameworks, including the Neural Tangent Kernel (NTK) framework (Jacot et al., 2018; Du et al., 2018; Lee et al., 2018), model recovery techniques (Zhong et al., 2017; Ge et al., 2018; Bakshi et al., 2019; Soltanolkotabi et al., 2018; Zhang et al., 2020), and the analysis of structured data (Li & Liang, 2018; Shi et al., 2022; Brutzkus & Globerson, 2021; Allen-Zhu & Li, 2022; Karp et al., 2021; Wen & Li, 2021; Zhang et al., 2023b; Li et al., 2023; Chowdhury et al., 2023).

2. Preliminaries

In this paper, we address the learning problem involving multiple tasks $\{\mathcal{T}_i\}_{i=1}^n$ and aim to find the optimal policy $\pi_i^\star$ for each task $\mathcal{T}_i$. We begin by presenting the preliminaries for a single task and then elaborate on our algorithm for learning with multiple tasks in the following section.

Markov decision process and Q-learning. The Markov decision process (MDP) is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the set of possible actions. The transition operator $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ gives the probability of transitioning from the current state $\boldsymbol{s}$ and action a to the next state $\boldsymbol{s}'$. The function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [-R_{\max}, R_{\max}]$ measures the reward for a given state-action pair. The discount factor $\gamma \in [0, 1)$ determines the significance of future rewards.

For the i-th task, the goal of the agent is to find the optimal policy $\pi_i^\star$ with $a_t = \pi_i^\star(\boldsymbol{s}_t)$ at each time step t. The aim is to maximize the expected discounted sum of reward as $\sum_{t=0}^\infty \gamma^t \cdot r_i(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1})$, where r_i denotes the reward function for the i-th task. For any state-action pair $(\boldsymbol{s}, a)$, we define the action-value function Q_i^π given a policy π as

$$Q_i^\pi(\boldsymbol{s}, a) = \mathbb{E}_{\pi,\mathcal{P}}\big[\textstyle\sum_{t=0}^\infty \gamma^t r_i(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1}) \mid \boldsymbol{s}_0 = \boldsymbol{s}, a_0 = a\big].$$

The optimal Q-function, denoted as $Q_i^{\pi^\star}$ or $Q_i^\star$, satisfies

$$\begin{aligned} Q_i^\star(\boldsymbol{s}, a) &:= \max_\pi Q_i^\pi(\boldsymbol{s}, a) \\ &= \mathbb{E}_{\boldsymbol{s}'|\boldsymbol{s},a}\, r_i(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma \max_{a'} Q_i^{\pi^\star}(\boldsymbol{s}', a'), \end{aligned} \tag{1}$$

where (1) is also known as the Bellman equation. Through the optimal action-value function $Q_i^\star$, the agent can derive the optimal policy (Watkins & Dayan, 1992; Sutton & Barto, 2018) following

$$\pi_i^\star(\boldsymbol{s}) = \arg\max_a Q_i^\star(\boldsymbol{s}, a). \tag{2}$$

Deep Q-networks (DQNs). The DQN utilizes a DNN parameterized with weights ω, i.e., $Q_i(\boldsymbol{s}, a; \omega) : \mathbb{R}^d \to \mathbb{R}$ for the i-th task, to approximate the optimal Q-value function $Q_i^\star$ in (1). Specifically, given input feature $\boldsymbol{x} := \boldsymbol{x}(\boldsymbol{s}, a)$, the output of the L-hidden-layer DNN is defined as

$$Q_i(\boldsymbol{s}, a; \omega) := \omega_{L+1}^\top / K \cdot \sigma\big(\omega_L^\top \cdots \sigma(\omega_1^\top \boldsymbol{x})\big), \tag{3}$$

where $\sigma(\cdot)$ is the ReLU activation function, i.e., $\sigma(z) = \max\{0, z\}$.

Successor feature. For i-the task, suppose the expected one-step reward associated with the transition $(\boldsymbol{s}, a, \boldsymbol{s}')$ can be computed as

$$r_i(\boldsymbol{s}, a, \boldsymbol{s}') = \boldsymbol{\phi}(\boldsymbol{s}, a, \boldsymbol{s}')^\top \boldsymbol{w}_i^\star, \text{ with } \boldsymbol{\phi}, \boldsymbol{w}_i^\star \in \mathbb{R}^d, \tag{4}$$

where ϕ remains the same for all the tasks. With the reward function in (4), the Q-value function in (2) is reformulated

$$\begin{aligned} &Q_i^\pi(\boldsymbol{s}, a) \\ &= \mathbb{E}_{\pi,\mathcal{P}}\big[\textstyle\sum_{t=0}^\infty \gamma^t \boldsymbol{\phi}(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1}) \mid \boldsymbol{s}_0, a_0\big]^\top \boldsymbol{w}_i^\star \\ &:= \psi_i^\pi(\boldsymbol{s}, a)^\top \boldsymbol{w}_i^\star. \end{aligned} \tag{5}$$

Then, the optimal Q function satisfies

$$\begin{aligned} &Q_i^\star(\boldsymbol{s},a)\\ =&\mathbb{E}_{\pi_i^\star,\mathcal{P}}\big[\textstyle\sum_{i=0}^{\infty}\gamma^i\boldsymbol{\phi}(\boldsymbol{s}_i,a_i,\boldsymbol{s}_{i+1})\mid \boldsymbol{s}_0,a_0\big]^\top\boldsymbol{w}_i^\star \qquad (6)\\ :=&\,\psi_i^\star(\boldsymbol{s},a)^\top\boldsymbol{w}_i^\star. \end{aligned}$$

3. Problem Formulation and Algorithm

Problem formulation. Without loss of generality, the data is assumed to be collected from the tasks in the order of $\mathcal{T}_1$ to $\mathcal{T}_n$ during the learning process. The goal is to utilize collected data for the current task, e.g., $\mathcal{T}_j$, and the learned knowledge from previous tasks $\{\mathcal{T}_i\}_{i=1}^{j-1}$ to derive the optimal policy $\pi_j^\star$ for the current $\mathcal{T}_j$. These tasks share the same environment dynamic but the reward function changes across the task as shown in (4). For each task $\mathcal{T}_i$, we denote its reward as

$$r_i=\boldsymbol{\phi}\cdot\boldsymbol{w}_i^\star,\quad \textit{with}\quad \|\boldsymbol{\phi}\|_2\le\phi_{\max}, \qquad (7)$$

where $\boldsymbol{\phi}$ is the transition feature across all the tasks and $\boldsymbol{w}_i^\star$ is the reward mapping.

From (6), the learning of optimal Q-function for the i-th task is decomposed as two sub-tasks: learning SF $\psi_i^\star(\boldsymbol{s},a)$ and learning reward $\boldsymbol{w}_i^\star$.

Reward mapping. To find the optimal $\boldsymbol{w}_i^\star$, we utilize the information from $\boldsymbol{\phi}(\boldsymbol{s},a,\boldsymbol{s}')$ and $r_i(\boldsymbol{s},a,\boldsymbol{s}')$. The value of $\boldsymbol{w}_i^\star$ can be obtained by solving the optimization problem

$$\min_{\boldsymbol{w}_i}:\ \|r_i-\boldsymbol{\phi}\cdot\boldsymbol{w}_i\|_2. \qquad (8)$$

Successor features. We use ψ_i^π to denote the successor feature for the i-th task, and ψ_i^π satisfies

$$\psi_i^\pi(\boldsymbol{s},a)=\mathbb{E}_{\boldsymbol{s}'|\boldsymbol{s},a}\,\boldsymbol{\phi}(\boldsymbol{s},a,\boldsymbol{s}')+\gamma\cdot\psi_i^\pi\big(\boldsymbol{s}',\pi(\boldsymbol{s}')\big). \qquad (9)$$

The expression given by (9) aligns perfectly with the Bellman equation in (1), where $\boldsymbol{\phi}$ acts as the reward. Therefore, following DQNs, we utilize a function $\psi(\boldsymbol{s},a)$ parameterized using the DNN as

$$\psi_i(\Theta_i;\boldsymbol{s},a)=H\big(\Theta_i;\boldsymbol{x}(\boldsymbol{s},a)\big), \qquad (10)$$

where $\boldsymbol{x}:\mathcal{S}\times\mathcal{A}\longrightarrow\mathbb{R}^d$ is the feature mapping of the state-action pair. Without loss of generality, we assume $|\boldsymbol{x}(\boldsymbol{s},a)|\le 1$. Then, finding $\psi^\star$ is to minimize the mean squared Bellman error (MSBE)

$$\begin{aligned}\min_{\Theta_i}:&f(\Theta_i):=\mathbb{E}_{(\boldsymbol{s},a)\sim\pi^\star}\Big[\mathbb{E}_{\boldsymbol{s}'|\boldsymbol{s},a}\,\psi_i(\Theta_i;\boldsymbol{s},a)\\ &-\boldsymbol{\phi}(\boldsymbol{s},a,\boldsymbol{s}')-\gamma\cdot\psi_i\big(\Theta_i;\boldsymbol{s}',\pi^\star(s')\big)\Big]^2.\end{aligned} \qquad (11)$$

It is worth mentioning that although (11) and (8) appear to be independent of each other, the update of $\boldsymbol{w}_i$ does affect the update of ψ_i through the shift in data distribution. The collected data is estimated based on the policy depending on the current estimated values of ψ_i and $\boldsymbol{w}_i$, which shifts the distribution of the collected data away from $\pi_i^\star$. This, in turn, leads to a bias depending on the value of $\boldsymbol{w}_i$ in the calculation of the gradient of Θ_i in minimizing (11).

Generalized policy improvement (GPI). Suppose we have acquired knowledge about the optimal successor features for the previous n tasks, and we use $\hat{\psi}_i$ to denote the estimated successor feature function for the i-th task with $i\in[n]$. Now, let's consider a new task $\mathcal{T}_{n+1}$ with the reward function defined as $r_{n+1}=\boldsymbol{\phi}\boldsymbol{w}_{n+1}^\star$. Instead of training from scratch, we can leverage the knowledge acquired from previous tasks to improve our approach. We achieve this by deriving the policy as follows

$$\pi(a|\boldsymbol{s})=\arg\max_a\max_{1\le i\le n+1}\hat{\psi}_i(\boldsymbol{s},a)^\top\boldsymbol{w}_{n+1}^\star. \qquad (12)$$

This strategy tends to yield better performance than relying solely on $\hat{\psi}_{n+1}(\boldsymbol{s},a)^\top\boldsymbol{w}_{n+1}^\star$, especially when $\hat{\psi}_{n+1}$ has not yet converged to the optimal successor feature $\psi_{n+1}^\star$ during the early learning stage, while some task is closely related to the new tasks, i.e., some $\boldsymbol{w}_i^\star$ is close to $\boldsymbol{w}_{n+1}^\star$. This policy improvement operator is derived from Bellman's policy improvement theorem (Bertsekas & Tsitsiklis, 1996) and (1). When the reward is fixed across different policies, e.g., $\{\pi_i\}_{i=1}^n$, and given that the optimal Q-function represents the maximum across the entire policy space, the maximum of multiple Q-functions corresponding to different policies, $\max_{1\le i\le n}Q^{\pi_n}$, is expected to be closer to $Q^\star$ than any individual Q-function, Q^{π_i}. In this paper, the parameter ϕ in learning the successor feature is analogous to the reward in learning the Q-function. As ϕ remains the same for different tasks, this analogy has inspired the utilization of GPI in our setting, even where the rewards change.

3.1. Successor feature Deep Q-Network

The goal is to find $\boldsymbol{w}_i$ and Θ_i by solving the optimization problems in (8) and (11) for each task sequentially, and the optimization problems are solved by mini-batch stochastic gradient descent (mini-batch SGD). Algorithm 1 contains two loops, and the outer loop number n is the number of tasks and inner loop number T is the maximum number of iterations in solving (8) and (11) for each task. At the beginning, we initialize the parameters as $\Theta^{(0)}$ and $\boldsymbol{w}_i^{(0)}$ for task i with $1\le i\le n$. In t-th inner loop for the i-th task, let $\boldsymbol{s}_t$ be the current state, and θ_c be the learned weights for task c. The agent selects and executes actions according to

$$a=\pi_\beta(\max_{c\in[i]}\psi(\Theta_c;\boldsymbol{s}_t,a)^\top\boldsymbol{w}_i^{(t)}), \qquad (13)$$

where $\pi_\beta(Q(\boldsymbol{s}_t, a))$ is the policy operator based on the function $Q(\boldsymbol{s}_t, a)$, e.g., greedy, ε-greedy, and softmax. For example, if $\pi_\beta(\cdot)$ stands for greedy policy, then $a = \arg\max_a \max_{c\in[i]} \psi(\Theta_c; \boldsymbol{s}_t, a)^\top \boldsymbol{w}_i^{(t)}$. The collected data are stored in a replay buffer with size N. Then, we sample a mini-batch of samples from the replay buffer and denote the samples as $\mathcal{D}_t$.

Algorithm 1 Successor Feature Deep Q-Network (SF-DQN)

Input: Number of iterations T, and experience replay buffer size N, step size $\{\eta_t, \kappa_t\}_{t=1}^T$.
Initialize $\{\Theta_i^{(0)}\}_{i=1}^n$ and $\{\boldsymbol{w}_i^{(0)}\}_{i=1}^n$.
for Task $i = 1, 2, \cdots, n$ **do**
  for $t = 0, 1, 2, \cdots, T-1$ **do**
    Collect data and store in the experience replay buffer $\mathcal{D}_t$ following a behavior policy π_t in (13).
    Perform gradient descent steps on $\Theta_i^{(t)}$ and $\boldsymbol{w}^{(t)}$ following (14).
  end for
  Return $Q_i = \psi_i(\Theta_i^{(T)})^\top \boldsymbol{w}_i^{(T)}$ for $i = 1, 2, \cdots, n$.
end for

Next, denote the gradient as $g_{\boldsymbol{w}}(\boldsymbol{s}_m, \boldsymbol{a}_m, \boldsymbol{s}'_m; \boldsymbol{w}^{(t)}) = \big(\boldsymbol{\phi}(\boldsymbol{s}_m, a_m, \boldsymbol{s}'_m)^\top \boldsymbol{w}^{(t)} - r(\boldsymbol{s}_m, a_m, \boldsymbol{s}'_m)\big) \cdot \boldsymbol{\phi}(\boldsymbol{s}_m, a_m, \boldsymbol{s}'_m)$ and $g_\Theta(\boldsymbol{s}_m, \boldsymbol{a}_m, \boldsymbol{s}'_m; \Theta^{(t)}) = \big(\psi(\Theta_i^{(t)}; \boldsymbol{s}_m, a_m) - \boldsymbol{\phi}(\boldsymbol{s}_m, a_m, \boldsymbol{s}'_m) - \gamma \cdot \psi(\Theta_i^{(t)}; \boldsymbol{s}'_m, a')\big) \cdot \nabla_{\Theta_i} \psi(\Theta_i^{(t)}; \boldsymbol{s}_m, a_m)$, we update the current weights using a mini-batch gradient descent algorithm following

$$\begin{aligned} \boldsymbol{w}^{(t+1)} &= \boldsymbol{w}^{(t)} - \kappa_t \cdot \sum_{m\in\mathcal{D}_t} g_{\boldsymbol{w}}(\boldsymbol{s}_m, \boldsymbol{a}_m, \boldsymbol{s}'_m; \boldsymbol{w}^{(t)}) \\ \Theta_i^{(t+1)} &= \Theta_i^{(t)} - \eta_t \cdot \sum_{m\in\mathcal{D}_t} g_\Theta(\boldsymbol{s}_m, \boldsymbol{a}_m, \boldsymbol{s}'_m; \Theta^{(t)}), \end{aligned} \tag{14}$$

where η_t and κ_t are the step sizes, and $a' = \arg\max_a \max_{c\in[i]} \psi(\Theta_c; \boldsymbol{s}'_m, a)^\top \boldsymbol{w}_i^{(t)}$. The gradient for $\Theta_i^{(t)}$, as $g_\Theta(\boldsymbol{s}_m, \boldsymbol{a}_m, \boldsymbol{s}'_m; \Theta^{(t)})$ in (14), can be viewed as the gradient of

$$\sum_{m\in\mathcal{D}_t} \big[\psi_i(\Theta_i; \boldsymbol{s}_m, a_m) - \boldsymbol{\phi}(\boldsymbol{s}_m, a_m, \boldsymbol{s}'_m) - \mathbb{E}_{\boldsymbol{s}'_m|\boldsymbol{s}_m, a_m} \max_{a'_m} \psi_i(\Theta_i^{(t)}; \boldsymbol{s}'_m, a'_m)\big]^2, \tag{15}$$

which is the approximation to (11) via replacing $\max_{a'} \psi_i^\star$ with $\max_{a'} \psi_i(\Theta_i^{(t)})$.

4. Theoretical Results

4.1. Summary of Major Theoretical Findings

To the best of our knowledge, our results (formally presented in Section 4.3) provide the first theoretical characterization for SF-DQN with GPI, including a comparison with the conventional Q-learning under commonly used assumptions. Before formally presenting them, we summarize the highlights as follows.

Table 1: Important Notations

K	Number of neurons in the hidden layer.
L	Number of the hidden layers.
d	Dimension of the feature mapping of $(\boldsymbol{s}, a)$.
T	Number of iterations.
$\Theta_i^\star$, $\boldsymbol{w}_i^\star$	The global optimal to (11) and (8) for i-th task.
N	Replay buffer size.
ρ_1	The smallest eigenvalue of $\mathbb{E}\nabla\boldsymbol{\psi}_i(\Theta_i^\star)\nabla\boldsymbol{\psi}_i(\Theta_i^\star)^\top$.
ρ_2	The smallest eigenvalue of $\mathbb{E}\boldsymbol{\phi}(\boldsymbol{s}, a)\boldsymbol{\phi}(\boldsymbol{s}, a)^\top$.
$q^\star$	A variable indicates the relevance between current and previous tasks.
$C^\star$	A constant related to the distribution shift between the behavior and optimal policies.

(T1) Learned Q-function converges to the optimal Q-function at a rate of $1/T$ with generalization guarantees. We demonstrate that the learned parameters $\Theta_i^{(T)}$ and $\boldsymbol{w}_i^{(T)}$ converge towards their respective ground truths, $\Theta_i^\star$ and $\boldsymbol{w}_i^\star$, indicating that SF-DQN converges to optimal Q-function at a rate of $1/T$ as depicted in (22) (Theorem 1). Moreover, the generalization error of the learned Q-function scales on the order of $\frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^\star\|_2}{1-\gamma-\Omega(N^{-1/2})-\Omega(C^\star)} \cdot \frac{1}{T}$. By employing a large replay buffer N, minimizing the data distribution shift factor $C^\star$, and improving the estimation of task-specific reward weights $\boldsymbol{w}^{(0)}$, we can achieve a lower generalization error.

(T2) GPI enhances the generalization of the learned model with respect to the task relevance factor $q^\star$. We demonstrate that, when GPI is employed, the learned parameters exhibit improved estimation error with a reduction rate at $\frac{1-c}{1-c\cdot q^\star}$ for some constant $c < 1$ (Theorem 2), where $q^\star$ is defined in (23). From (23), it is clear that $q^\star$ decreases as the distances between task-specific reward weights, denoted as $\|\boldsymbol{w}_j^\star - \boldsymbol{w}_i^\star\|_2$, become smaller. This indicates a close relationship between the previous tasks and the current task, resulting in a smaller $q^\star$ and, consequently, a larger improvement through the usage of GPI.

(T3) SF-DQN achieves a superior performance over conventional DQN by a factor of γ for the estimation error of the optimal Q-function. When we directly transfer the learned knowledge of the Q-function to a new task without any additional training, our results demonstrate that SF-DQN always outperforms its conventional counterpart, DQN, by a factor of γ (Theorems 3 and 4). As γ ap-

proaches one, we raise the emphasis on long-term rewards, making the accumulated error derived from the incorrect Q-function more significant. Consequently, this leads to reduced transferability between the source tasks and the target task. Conversely, when γ is small, indicating substantial potential for transfer learning between the source and target tasks, we observe a more significant improvement when using SF-DQN.

4.2. Assumptions

In this section, we propose the assumptions in deriving our major theoretical results. These assumptions are commonly used in existing RL and neural network learning theories.

Assumption 1. *There exists a deep neural network with weights $\Theta_i^\star$ such that it minimizes* (11) *for the i-th task, i.e, $f(\Theta_i^\star) = 0$.*

Assumption 1 assumes a substantial expressive power of the deep neural network, allowing it to effectively represent $\psi^\star$ in the presence of an unknown ground truth $\Theta^\star$.

Assumption 2. *At any fixed outer iteration t, the behavior policy π_t and its corresponding transition kernel $\mathcal{P}_t$ satisfy*

$$\sup_{\boldsymbol{s}\in\mathcal{S}} \; d_{TV}\big(\mathbb{P}(\boldsymbol{s}_\tau \in \cdot) \mid \boldsymbol{s}_0 = \boldsymbol{s}), \mathcal{P}_t\big) \le \lambda \nu^\tau, \quad (16)$$

for some constant $\lambda > 0$ and $\nu \in (0,1)$, where d_{TV} denotes the total-variation distance.

Assumption 2 assumes the Markov chain $\{\boldsymbol{s}_n, a_n, \boldsymbol{s}_{n+1}\}$ induced by the behavior policy is uniformly ergodic with the corresponding invariant measure $\mathcal{P}_t$. This assumption is standard in Q-learning (Xu & Gu, 2020; Zou et al., 2019; Bhandari et al., 2018), where the data are non-i.i.d.

Assumption 3. *Let $Q^\star$ and Q_t be the optimal and estimated Q-function, respectively. We assume the greedy policy π_t, i.e., $\pi_t(a|\boldsymbol{s}) = \arg\max_{a'} Q_t(\boldsymbol{s}, a')$, satisfies*

$$\begin{aligned} &\big|\pi_t(a|\boldsymbol{s}) - \pi^\star(a|\boldsymbol{s})\big| \\ &\le C \cdot \sup_{(\boldsymbol{s},a)} \|Q_t(\boldsymbol{s},a) - Q^\star(\boldsymbol{s},a)\|_F, \end{aligned} \quad (17)$$

where C is a positive constant. Equivalently, when $Q_t = \psi(\Theta_i^{(t)})^\top \boldsymbol{w}_i^{(t)}$, we have

$$\begin{aligned} &\big|\pi_t(a|\boldsymbol{s}) - \pi^\star(a|\boldsymbol{s})\big| \\ &\le C \cdot \big(\|\Theta_i^{(t)} - \Theta_i^\star\|_2 + \|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_i^\star\|_2\big). \end{aligned} \quad (18)$$

Assumption 3 indicates the policy difference between the behavior policy and the optimal policy. Moreover, (18) can be considered as a more relaxed variant of condition (2) in (Zou et al., 2019) as (18) only requires the holding for the distance of an arbitrary function from the ground truth, rather than the distance between two arbitrary functions.

4.3. Main Theoretical Findings

4.3.1. Convergence analysis of SF-DQN

Theorem 1 demonstrates that the learned Q function converges to the optimal Q function when using SF-DQN for Task 1. Notably, GPI is not employed for the initial task, as we lack prior knowledge about the environment. Specifically, given conditions (i) the initial weights for ψ are close to the ground truth as shown in (19), (ii) the replay buffer is large enough as in (20), and (iii) the distribution shift between the behavior policy and optimal policy is bounded (as shown in Remark), the learned parameters from Algorithm (1) for task 1, $\psi_1(\Theta_1)$ and $\boldsymbol{w}_1$, converge to the ground truth $\psi_1^\star$ and $\boldsymbol{w}_1^\star$ as in (21), indicating that the learned Q function converges to the optimal Q function as in (22).

Theorem 1 (Convergence analysis of SF-DQN without GPI)**.** *Suppose the assumptions in Section 4.2 hold and the initial neuron weights of the SF of task* 1 *satisfy*

$$\frac{\|\Theta_1^{(0)} - \Theta_1^\star\|_F}{\|\Theta_1^\star\|_F} \le (1 - c_N) \cdot \frac{\rho_1}{K^2}, \quad (19)$$

for some positive c_N. When we select the step size as $\eta_t = \frac{1}{t+1}$, and the size of the replay buffer is

$$N = \Omega(c_N^{-2}\rho_1^{-1} \cdot K^2 \cdot L^2 d \log q). \quad (20)$$

Then, with the high probability of at least $1 - q^{-d}$, the weights $\theta^{(T)}$ from Algorithm 1 satisfy

$$\begin{aligned} \|\Theta_1^{(T)} - \Theta_1^\star\|_2 &\le \frac{C_1 + C^\star \cdot \|\boldsymbol{w}_1^{(0)} - \boldsymbol{w}_1^\star\|_2}{(1-\gamma-c_N)(1-\gamma)\rho_1 - C^\star} \cdot \frac{\log^2 T}{T}, \\ \|\boldsymbol{w}_1^{(T)} - \boldsymbol{w}_1^\star\|_2 &\le \Big(1 - \frac{\rho_2}{\boldsymbol{\phi}_{\max}}\Big)^T \|\boldsymbol{w}_1^{(0)} - \boldsymbol{w}_1^\star\|_2, \end{aligned} \quad (21)$$

where $C_1 = (2+\gamma) \cdot R_{\max}$, and $C^\star = |\mathcal{A}| \cdot R_{\max} \cdot (1 + \log_\nu \lambda^{-1} + \frac{1}{1-\nu}) \cdot C$. Specifically, the learned Q-function satisfies

$$\begin{aligned} &\max_{\boldsymbol{s},a} \Big|Q_1^{(T)} - Q^\star\Big| \\ &\le \frac{C_1 + \|\boldsymbol{w}_1^{(0)} - \boldsymbol{w}_1^\star\|_2}{(1-\gamma-c_N)(1-\gamma)\rho_1 - 1} \cdot \frac{\log^2 T}{T} \\ &\quad + \frac{\|\boldsymbol{w}_1^{(0)} - \boldsymbol{w}_1^\star\|_2 R_{\max}}{1-\gamma}\Big(1 - \frac{\rho_2}{\boldsymbol{\phi}_{\max}}\Big)^T. \end{aligned} \quad (22)$$

Remark 1 (**upper bound of** C): To ensure the meaningfulness of the upper bound in (22), specifically that the denominator needs to be greater than 0, C has an explicit upper bound as $C \le \frac{(1-\gamma-c_N)(1-\gamma)\rho_1}{|\mathcal{A}| \cdot R_{\max}}$. Considering the definition of C in Assumption 3, it implies that the difference between the behavior policy and the optimal policy is

bounded. In other words, the fraction of bad tuples [1] in the collected samples is constrained.

Remark 2 (**Initialization**): Note that (19) requires a good initialization. Firstly, it is still a state-of-the-art practice in analyzing Q-learning via deep neural network approximation. Secondly, according to the NTK theory (Jacot et al., 2018), there always exist some good local minima, which is almost as good as the global minima, near some random initialization. Finally, such a good initialization can also be adapted from some pre-trained models.

4.3.2. Improved performance with GPI.

Theorem 2 establishes that the estimated Q function converges towards the optimal solution with the implementation of GPI as shown in (24), leveraging the prior knowledge learned from previous tasks. The enhanced performance associated with GPI finds its expression as $q^\star$ defined in (23). Notably, when tasks i and j exhibit a higher degree of correlation, meaning that the distance between $\boldsymbol{w}_i^\star$ and $\boldsymbol{w}_j^\star$ for tasks i and j is smaller, we can observe a more substantial enhancement by employing GPI in transferring knowledge from task i to task j from (24).

Theorem 2 (Convergence analysis of SF-DQN with GPI)**.** *Let us define*

$$q^\star = \frac{2\gamma \cdot R_{\max}}{1-\gamma} \cdot \frac{\min_{1\le i\le j-1} \|\boldsymbol{w}_i^\star - \boldsymbol{w}_j^\star\|_2}{\|\Theta_j^{(0)} - \Theta_j^\star\|_2}. \tag{23}$$

Then, with the probability of at least $1-q^{-d}$, the neuron weights $\Theta_j^{(T)}$ for the j-th task satisfy

$$\begin{aligned}&\|\Theta_j^{(T)} - \Theta_j^\star\|_2 \\ \le &\frac{C_1 + C^\star \|\boldsymbol{w}_j^{(0)} - \boldsymbol{w}_j^\star\|_2}{(1-\gamma-c_N)(1-\gamma)\rho_1 - \min\{q^\star, 1\}\cdot C^\star} \cdot \frac{\log^2 T}{T}.\end{aligned} \tag{24}$$

Remark 3 (**Improvement via GPI**): Utilizing GPI enhances the convergence rate from in the order of $\frac{1}{1-C^\star}\cdot\frac{1}{T}$ to in the order of $\frac{1}{1-q^\star\cdot C^\star}\cdot\frac{1}{T}$. When the distance between the source task and target tasks is small, $q^\star$ can approach zero, indicating an improved generalization error by a factor of $1-C^\star$, where $C^\star$ is proportional to the fraction of bad tuples. The improvement achieved through GPI is derived from the reduction of the distance between the behavior policy and the optimal policy, subsequently decreasing the fraction of bad tuples in the collected data. Here, $C^\star$ is proportional to the fraction of bad tuples without using GPI, and $q^\star \cdot C^\star$ is proportional to the fraction of bad tuples when GPI is employed.

4.3.3. Improved Performance with the Knowledge Transfer

Using our proposed SF-DQN, we have estimated $Q_i^{\pi_i^\star}$ for task i. When the reward changes to $r_{n+1}(\boldsymbol{s}, a, \boldsymbol{s}') = \boldsymbol{\phi}^\top(\boldsymbol{s}, a, \boldsymbol{s}')\boldsymbol{w}_{n+1}^\star$ for a new task $\mathcal{T}_{n+1}$, and once $\boldsymbol{w}^\star n+1$ is estimated, we can calculate the estimated Q-value function for $\mathcal{T}_{n+1}$ by setting

$$Q_{n+1}^{\pi_{n+1}}(\boldsymbol{s}, a) = \max_{1\le j\le n} \psi(\Theta_j^{(T)}; \boldsymbol{s}, a)\boldsymbol{w}_{n+1}^\star. \tag{25}$$

As $\boldsymbol{w}_{n+1}^{(t)}$ experiences linear convergence to its optimal $\boldsymbol{w}^\star$, which is significantly faster than the sub-linear convergence of $\Theta_{n+1}^{(t)}$, as shown in (21), this derivation of Q_{n+1} in (25) simplifies the computation of $\Theta_{n+1}^\star$ into a much more manageable supervised setting for approximating $w_{n+1}^\star$ with only a modest performance loss as shown in (26). This is demonstrated in the following Theorem 3.

Theorem 3 (Transfer learning via SF-DQN)**.** *For the $(n+1)$-th task with $r_{n+1} = \boldsymbol{\phi}^\top w_{n+1}^\star$, suppose the Q-value function is derived based on* (25)*, we have*

$$\begin{aligned}&\max_{\boldsymbol{s},a} |Q_{n+1}^{\pi_{n+1}}(\boldsymbol{s}, a) - Q_{n+1}^\star(\boldsymbol{s}, a)| \\ \le &\frac{2\gamma}{1-\gamma}\phi_{\max} \min_{j\in[n]} \|\boldsymbol{w}_j^\star - \boldsymbol{w}_{n+1}^\star\|_2 + \frac{\|\boldsymbol{w}_{n+1}^\star\|_2}{(1-\gamma)\cdot T}.\end{aligned} \tag{26}$$

Remark 4 (Connection with existing works of SF in tabular cases): The second term of the upper bound in (26), $\frac{\|\boldsymbol{w}_{n+1}^\star\|_2}{(1-\gamma)\cdot T}$, characterizes the value of ϵ assumed in (Barreto et al., 2017), which results from the approximation error of the optimal Q-functions in the previous tasks [2].

Without the SF decomposition as shown in (6), one can apply a similar strategy in (25) for DQN as

$$Q_{n+1}^{\pi'_{n+1}}(s, a) = \max_{1\le j\le n} Q(\omega_j^{(T)}; s, a). \tag{27}$$

In Theorem 4, (28) illustrates the performance of (27) through DQN. Compared to Theorem 3, transfer learning via DQN is worse than that via SF-DQN by a factor of $\frac{1+\gamma}{2}$ when comparing the estimation error of the optimal function $Q_{n+1}^\star$ in (26) and (28), indicating the advantages of using SFs in transfer reinforcement learning.

[1] A "bad tuple" refers to the data $(\boldsymbol{s}, a)$ collected based on behavior policy $a = \pi_t(\boldsymbol{s})$ that differs from the optimal policy $a = \pi^\star(\boldsymbol{s})$. Intuitively, we can clearly see that the fraction of "bad tuples" is positively related to the distance between the behavior policy and the optimal policy (the motivation of Assumption 3). In fact, similar assumptions can be found in many theoretical frameworks when analyzing Q-learning with function approximation (Zou et al., 2019) to guarantee that there is a certain fraction of collected data that is useful for estimating the ground-truth Q-value.

[2] Our upper bound in (26) differs from the one in (Barreto et al., 2017) in the first term. This distinction arises from our improvement in Lemma 9 compared to Lemma 1 in (Barreto et al., 2017). See Appendix G for the proof of Lemma 9.

Theorem 4 (Transfer learning via DQN). *For the* $(n+1)$*-th task with* $r_{n+1} = \phi \cdot w^{\star}_{n+1}$*, suppose the Q-value function is derived based on* (27)*, we have*

$$\begin{aligned} &\max_{(\boldsymbol{s},a)} : |Q^{\pi'_{n+1}}_{n+1}(\boldsymbol{s},a) - Q^{\star}_{n+1}(\boldsymbol{s},a)| \\ \leq &\frac{2}{1-\gamma}\phi_{\max} \cdot \min_{j\in[n]} \|\boldsymbol{w}^{\star}_{j} - \boldsymbol{w}^{\star}_{n+1}\|_2 + \frac{\|\boldsymbol{w}^{\star}_{n+1}\|_2}{(1-\gamma)\cdot T}. \end{aligned} \tag{28}$$

Remark 5 (Improvement by a factor of $\frac{1+\gamma}{2}$): Transfer learning performance in SF-DQN is influenced by the knowledge gap between previous and current tasks, primarily attributed to differences in rewards and data distribution. In SF-DQN, the impact of reward differences is relatively small since ϕ that plays the role of reward remains fixed. The parameter γ affects the influence of data distribution differences. A small γ prioritizes immediate rewards, thereby the impact of data distribution on the knowledge gap is not significant. With a small γ, the impact of reward difference dominates, resulting in a high gap between SF-DQN and DQN in transfer learning.

4.4. Technical Challenges, and Comparison with Existing Works

Beyond deep learning theory: challenges in deep reinforcement learning. The proof of Theorem 1 is inspired by the convergence analysis of one-hidden-layer neural networks within the semi-supervised learning (Zhong et al., 2017; Zhang et al., 2022) and a recent theoretical framework in analyzing DQN (Zhang et al., 2023a). This proof tackles *two primary objectives*: (**i**) the first objective involves characterizing the local convex region of the objective functions presented in (11) and (8); (**ii**) the second objective focuses on quantifying the distance between the gradient defined in (14) and the gradient of the objective functions in (11) and (8).

However, extending this approach from the semi-supervised learning setting to the deep reinforcement learning domain introduces *additional challenges*. First, we expand our proof beyond the scope of one-hidden-layer neural networks to encompass multi-layer neural networks. This extension requires new technical tools for characterizing the Hessian matrix and concentration bounds, as outlined in Appendix F.1. Second, the approximation error bound deviates from the supervised learning scenarios due to several factors: the non-i.i.d. of the collected data, the distribution shift between the behavior policy and the optimal policy, and the approximation error incurred when utilizing (15) to estimate (11). Addressing these challenges requires developing supplementary tools, as mentioned in Lemma 7. Notably, this approximation does not exhibit scaling behavior proportional to $\|\Theta_i - \Theta^{\star}_i\|_2$, resulting in a sublinear convergence rate.

Beyond DQN: challenges in GPI. The major challenges in proving Theorems 2-4 centers on deriving the improved performance by utilizing GPI. The intuition is as follows. Imagine we have two closely related tasks, labeled as i and j, with their respective optimal weight vectors, $\boldsymbol{w}^{\star}_i$ and $\boldsymbol{w}^{\star}_j$, being close to each other. This closeness suggests that these tasks share similar rewards, leading to a bounded distributional shift in the data, which, in turn, implies that their optimal Q-functions should exhibit similarity. To rigorously establish this intuition, we aim to characterize the distance between these optimal Q-functions, denoted as $|Q^{\star}_i - Q^{\star}_j|$, in terms of the Euclidean distance between their optimal weight vectors, $||\boldsymbol{w}^{\star}_i - \boldsymbol{w}^{\star}_j||_2$ (See details in Appendix G). Furthermore, we can only estimate the optimal Q-function for previous tasks during the learning process, and such an estimation error accumulates in the temporal difference learning, e.g., the case of the SF learning of $\psi^{\star}$. We developed novel analytical tools to quantify the error accumulating in the temporal difference learning (see details in Appendix C), which is not a challenge for previous works in the supervised learning setting.

5. Experiments

This section summarizes empirical validation for the theoretical results obtained in Section 4 using a synthetic RL environment. The experiment setup and additional experimental results for real-world RL benchmarks are summarized in Appendix E.

Convergence of SF-DQN with varied initialization. Figure 1 shows the performance of Algorithm 1 with different initial $\boldsymbol{w}^{(0)}_1$ to the ground truth $\boldsymbol{w}^{\star}_1$. When the initialization is close to the ground truth, we observe an increased accumulated reward, which verifies our theoretical findings in (22) that the estimation error of the optimal Q-function reduces as $\|\boldsymbol{w}^{(0)}_1 - \boldsymbol{w}^{\star}\|_2$ decreases.

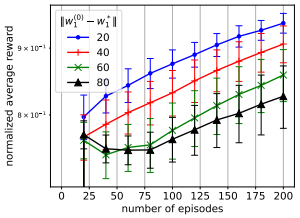


Figure 1: Performance of SF-DQN presented in Algorithm 1 on Task 1.

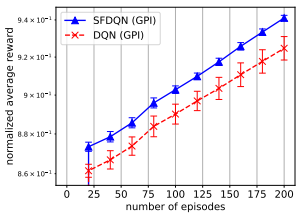


Figure 2: Transfer comparison for SF-DQN and DQN (with GPI)

Performance of SF-DQN with GPI when adapting to tasks with varying relevance. We conducted experiments to investigate the impact of GPI with varied task relevance. Since the difference in reward mapping impacts data distribution shift, rewards, and consequently the optimal Q-

Table 2: Normalized average reward for SF-DQN with and without GPI.

$\|w_1^* - w_2^*\|$	$= 0.01$	$= 0.1$	$= 1$	$= 10$
SF-DQN (w/ GPI)	$\mathbf{0.986} \pm 0.007$	$\mathbf{0.965} \pm 0.007$	$\mathbf{0.827} \pm 0.008$	$\mathbf{0.717} \pm 0.012$
SF-DQN (w/o GPI)	0.942 ± 0.004	0.911 ± 0.013	0.813 ± 0.009	0.707 ± 0.011

function, we utilize the metric $\|\boldsymbol{w}_1^\star - \boldsymbol{w}_2^\star\|_2$ to measure the task irrelevance. The results summarized in Table 2 demonstrate that when tasks are similar (i.e., small $\|\boldsymbol{w}_1^\star - \boldsymbol{w}_2^\star\|$), SF-DQN with GPI consistently outperforms its counterpart without GPI. However, when tasks are dissimilar (i.e., large $\|\boldsymbol{w}_1^\star - \boldsymbol{w}_2^\star\|$), both exhibit the same or similar performance, indicating that GPI is ineffective when two tasks are irrelevant. The observations in Table 2 validate our theoretical findings in (24), showing a more significant improvement in using GPI as $\|\boldsymbol{w}_1^\star - \boldsymbol{w}_2^\star\|_2$ decreases.

Comparison of the SF-DQN agent and DQN agent. From Figure 2, it is evident that the SF-DQN agent consistently achieves a higher average reward (task 2) than the DQN when starting training on task 2, where transfer learning occurs. These results strongly indicate the improved performance of the SF-DQN agent over the DQN, aligning with our findings in (26) and (28). SF-DQN benefits from reduced estimation error of the optimal Q-function compared to DQN when engaging in transfer reinforcement learning for relevant tasks.

6. Conclusions

This paper analyzes the transfer learning performance of SF & GPI, with SF being learned using deep neural networks. Theoretically, we present a convergence analysis of our proposed SF-DQN with generalization guarantees and provide theoretical justification for its superiority over DQN without using SF in transfer reinforcement learning. We further verify our theoretical findings through numerical experiments conducted in both synthetic and benchmark RL environments. Future directions include exploring the possibility of learning ϕ using a DNN approximation and exploring the combination of successor features with other deep reinforcement learning algorithms.

Acknowledgment

Part of this work was done when Shuai Zhang was a postdoc at Rensselaer Polytechnic Institute (RPI). This work was supported by AFOSR FA9550-20-1-0122, ARO W911NF-21-1-0255, NSF 1932196, NSF CAREER project 2047177, Cisco Research Award, and IBM through the IBM-Rensselaer Future of Computing Research Collaboration. We thank all anonymous reviewers for their constructive comments.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Allen-Zhu, Z. and Li, Y. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022.

Bacon, P.-L., Harb, J., and Precup, D. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Bakshi, A., Jayaram, R., and Woodruff, D. P. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory*, pp. 195–268. PMLR, 2019.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Barreto, A., Borsa, D., Quan, J., Schaul, T., Silver, D., Hessel, M., Mankowitz, D., Zidek, A., and Munos, R. Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *International Conference on Machine Learning*, pp. 501–510. PMLR, 2018.

Bertsekas, D. and Tsitsiklis, J. N. *Neuro-dynamic programming*. Athena Scientific, 1996.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pp. 1691–1692. PMLR, 2018.

Bhatia, R. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Brutzkus, A. and Globerson, A. An optimization and generalization analysis for max-pooling networks. In *Uncertainty in Artificial Intelligence*, pp. 1650–1660. PMLR, 2021.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. *Advances in Neural Information Processing Systems*, 32, 2019.

Chowdhury, M. N. R., Zhang, S., Wang, M., Liu, S., and Chen, P.-Y. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6074–6114. PMLR, 2023.

Coronato, A., Naeem, M., De Pietro, G., and Paragliola, G. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964, 2020.

Dayan, P. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624, 1993.

Dong, K., Yang, J., and Ma, T. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *Advances in Neural Information Processing Systems*, 34:26168–26182, 2021.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

Du, S. S., Lee, J. D., Mahajan, G., and Wang, R. Agnostic q-learning with function approximation in deterministic systems: Near-optimal bounds on approximation error and sample complexity. *Advances in Neural Information Processing Systems*, 33:22327–22337, 2020.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Ge, R., Lee, J. D., and Ma, T. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BkwHObbRZ.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

Ji, X., Chen, M., Wang, M., and Zhao, T. Sample complexity of nonparametric off-policy evaluation on low-dimensional manifolds using deep networks. *arXiv preprint arXiv:2206.02887*, 2022.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pp. 651–673. PMLR, 2018.

Karp, S., Winston, E., Li, Y., and Singh, A. Local signal adaptivity: Provable feature learning in neural networks beyond kernels. *Advances in Neural Information Processing Systems*, 34:24883–24897, 2021.

Kulkarni, T. D., Narasimhan, K., Saeedi, A., and Tenenbaum, J. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016a.

Kulkarni, T. D., Saeedi, A., Gautam, S., and Gershman, S. J. Deep successor reinforcement learning. *arXiv preprint arXiv:1606.02396*, 2016b.

Lazaric, A. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning: State-of-the-Art*, pp. 143–173. Springer, 2012.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Li, H., Wang, M., Liu, S., and Chen, P.-Y. A theoretical understanding of vision transformers: Learning, generalization, and sample complexity. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jClGv3Qjhb.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018.

Liu, F., Viano, L., and Cevher, V. Understanding deep neural function approximation in reinforcement learning via ϵ-greedy exploration. *arXiv preprint arXiv:2209.07376*, 2022.

Mitrophanov, A. Y. Sensitivity and convergence of uniformly ergodic markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Nguyen-Tang, T., Gupta, S., Tran-The, H., and Venkatesh, S. On sample complexity of offline reinforcement learning with deep reLU networks in besov spaces. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=LdEm0umNcv.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pp. 4430–4438, 2018.

Schwarting, W., Alonso-Mora, J., and Rus, D. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:187–210, 2018.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Shi, Z., Wei, J., and Liang, Y. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. In *International Conference on Learning Representations*, 2022.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., Precup, D., and Singh, S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Suzuki, T. Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

Taylor, M. E. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2018.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Watkins, C. J. and Dayan, P. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Wen, Z. and Li, Y. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pp. 11112–11122. PMLR, 2021.

Xu, P. and Gu, Q. A finite-time analysis of q-learning with neural network function approximation. In *International Conference on Machine Learning*, pp. 10555–10565. PMLR, 2020.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. I. On function approximation in reinforcement learning: optimism in the face of large state spaces. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 13903–13916, 2020.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

Zhang, J., Springenberg, J. T., Boedecker, J., and Burgard, W. Deep reinforcement learning with successor features for navigation across similar environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2371–2378. IEEE, 2017.

Zhang, S., Wang, M., Xiong, J., Liu, S., and Chen, P.-Y. Improved linear convergence of training cnns with generalizability guarantees: A one-hidden-layer case. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis. In *International Conference on Learning Representations*, 2022.

Zhang, S., Li, H., Wang, M., Liu, M., Chen, P.-Y., Lu, S., Liu, S., Murugesan, K., and Chaudhury, S. On the convergence and sample complexity analysis of deep q-networks with epsilon-greedy exploration. *Advances in Neural Information Processing Systems*, 36, 2023a.

Zhang, S., Wang, M., Chen, P.-Y., Liu, S., Lu, S., and Liu, M. Joint edge-model sparse learning is provably efficient for graph neural networks. *The Eleventh International Conference on Learning Representations*, 2023b.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4140–4149. JMLR. org, https://arxiv.org/abs/1706.03175, 2017.

Zhu, Z., Lin, K., Jain, A. K., and Zhou, J. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for sarsa with linear function approximation. *Advances in Neural Information Processing Systems*, 32, 2019.

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 (11) and (8) (Lemma 6), (**ii**) Characterization of the difference between the empirical gradient in (14) 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 (61), 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 (25) 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 (25) 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 (25) 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

$$f_{\pi^\star}(\theta):=\mathbb{E}_{(\boldsymbol{s},a)\sim\pi^\star}\big\|\psi(\theta;\boldsymbol{s},a)-\mathbb{E}_{\boldsymbol{s}'|(\boldsymbol{s},a),a'\sim\pi^\star(\boldsymbol{s}')}\big(\phi(\boldsymbol{s},a,\boldsymbol{s}')+\gamma\cdot\psi(\theta^\star;\boldsymbol{s}',a')\big)\big\|_2^2. \tag{29}$$

We can see that $\theta^\star$ is the global minimal to (29) with Assumption 1. For the convenience of presentation, we simplify $f_{\pi^\star}$ as f in the supplementary materials.

Then, the gradient of (29) is

$$\begin{aligned}&\nabla f_{\pi^\star}(\theta)\\&=\mathbb{E}_{(\boldsymbol{s},a)\sim\pi^\star,\boldsymbol{s}'|(\boldsymbol{s},a)\sim\mathcal{P},a'\sim\pi^\star}\big(\psi(\theta;\boldsymbol{s},\boldsymbol{a})-\phi(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')-\gamma\cdot\psi(\theta^\star;\boldsymbol{s}',a')\big)\cdot\nabla\psi(\theta;\boldsymbol{s},a).\end{aligned} \tag{30}$$

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,

$$\nabla_\ell f(\theta^\star):=\nabla_{\theta_\ell}f(\theta^\star)=\mathbf{0},\qquad\forall\ell\in[L]. \tag{31}$$

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, $\nabla_{\mathrm{Vec}(\theta)}H$.

Let n denote the dimension of $\boldsymbol{W}$ defined in (2). We denote n_l as the dimension of the vectorized neuron weights in the ℓ-th layer, namely, $n_\ell=\dim(\mathrm{vec}(\theta_\ell))$.

Then, the gradient in updating θ as

$$\begin{aligned}&g^{(t)}(\theta^{(t)};\mathcal{D}_t)\\&=\sum_{m\in\mathcal{D}_t}\Big(\psi(\theta^{(t)};\boldsymbol{s}_m,a_m)-\phi(\boldsymbol{s}_m,a_m,\boldsymbol{s}'_m)-\gamma\cdot\psi(\theta^{(t)};\boldsymbol{s}'_m,a'_m)\Big)\cdot\nabla_\theta\psi(\theta^{(t)};\boldsymbol{s}_m,a_m)\end{aligned} \tag{32}$$

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

$$\theta^{(t+1)}=\theta^{(t)}-\eta_t\cdot g^{(t)}(\theta^{(t)};\mathcal{D}_t). \tag{33}$$

Similar to (32), we define the gradient

$$l^{(t)}(\boldsymbol{w}^{(t)};\mathcal{D}_t)=\sum_{m\in\mathcal{D}_t}\Big(\phi(\boldsymbol{s}_m,a_m,\boldsymbol{s}'_m)^\top\boldsymbol{w}^{(t)}-r(\boldsymbol{s}_m,a_m,\boldsymbol{s}'_m)\Big)\cdot\phi(\boldsymbol{s}_m,a_m,\boldsymbol{s}'_m). \tag{34}$$

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

$$\begin{aligned}\|\nabla_\ell H\|_2&=\max_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}^\top\nabla_\ell H\|_2=\max_{\boldsymbol{\alpha}}\Big|\sum_{j=1}^K\boldsymbol{\alpha}_j^\top\frac{\partial H}{\partial\boldsymbol{w}_{\ell,j}}\Big|,\\ \|\nabla_\ell^2 H\|_2&=\max_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}^\top\nabla_\ell^2\,H\,\boldsymbol{\alpha}\|_2=\max_{\boldsymbol{\alpha}}\Big(\sum_{j=1}^K\boldsymbol{\alpha}_j^\top\frac{\partial H}{\partial\boldsymbol{w}_{\ell,j}}\Big)^2.\end{aligned} \tag{35}$$

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

$$\boldsymbol{h}^{(\ell)}(\theta;\boldsymbol{s},a)=\sigma(\theta_{\ell-1}^\top\boldsymbol{h}^{(\ell-1)})=\cdots=\sigma\Big(\theta_\ell^\top\sigma\big(\theta_{\ell-1}\cdots\sigma(\theta_1^\top\boldsymbol{x}(\boldsymbol{s},a))\big)\Big). \tag{36}$$

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

$$\psi(\theta;\boldsymbol{s},a)=\frac{\mathbf{1}^\top}{K_L}\sum_{k=1}^{K_L}\sigma(\theta_{L,k}^\top\boldsymbol{h}^{(L)})=\frac{\mathbf{1}^\top}{K_L}\sigma\big(\theta_L^\top\sigma(\theta_{L-1}^\top\boldsymbol{h}^{(L-1)})\big), \tag{37}$$

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

$$\begin{aligned}&\mathcal{J}_\ell(\theta)\\ &=\begin{cases}\big[\mathbf{1}^\top\sigma'(\theta_L^\top\boldsymbol{h}^{(L)})\theta_L^\top\cdot\sigma'(\theta_{L-1}^\top\boldsymbol{h}^{(L-1)})\theta_{L-1}^\top\cdots\sigma'(\theta_{\ell+1}^\top\boldsymbol{h}^{(\ell+1)})\theta_{\ell+1}^\top\big]^\top & \text{if}\quad \ell>1\\ \mathbf{1}\quad\text{if}\quad\ell=1\end{cases}.\end{aligned} \tag{38}$$

Then, the gradient of ψ can be represented as

$$\frac{\partial\psi}{\partial\theta_{\ell,k}}(\theta)=\frac{1}{K_\ell}\mathcal{J}_{\ell,k}(\theta)\sigma'\big(\theta_{\ell,k}^\top\boldsymbol{h}^{(\ell)}(\theta)\big)\boldsymbol{h}^{(\ell)}(\theta), \tag{39}$$

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.

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 (32) at point $\theta^{(t)}$ with respect to the dataset $\mathcal{D}_t$.
$f_{\pi^\star}$ or f	The population risk function defined in (29).
$\nabla_\ell H(\hat{\theta})$	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 (38).
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.

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*

$$|\lambda_i(\boldsymbol{B}) - \lambda_i(\boldsymbol{A})| \leq \|\boldsymbol{E}\|_2, \quad \forall \quad i \in [m]. \tag{40}$$

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*

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

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

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

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*

$$|\mathcal{S}_\xi| \leq (1 + 2/\xi)^d. \tag{41}$$

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*

$$\|\boldsymbol{A}\|_2 \leq (1-\xi)^{-1} \max_{\boldsymbol{\alpha}_1 \in \mathcal{S}_\xi(d_1), \boldsymbol{\alpha}_2 \in \mathcal{S}_\xi(d_2)} |\boldsymbol{\alpha}_1^T \boldsymbol{A} \boldsymbol{\alpha}_2|. \tag{42}$$

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. Definitions 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*

$$(\mathbb{E}|X|^p)^{1/p} \leq c_1\sqrt{p} \tag{43}$$

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

$$\mathbb{E}e^{s(X - \mathbb{E}X)} \leq e^{c_2\|X\|_{\psi_2}^2 s^2} \tag{44}$$

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*

$$(\mathbb{E}|X|^p)^{1/p} \leq c_3 p \tag{45}$$

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

$$\mathbb{E}e^{s(X - \mathbb{E}X)} \leq e^{c_4\|X\|_{\psi_1}^2 s^2} \tag{46}$$

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*

$$\|\theta - \theta^\star\|_2 \lesssim \frac{c_N \cdot \sigma_K}{\rho_1 \cdot K} \tag{47}$$

for some constant $c_N \in (0,1)$. Then, for the $f_{\pi^\star}$ defined in (29)*, we have*

$$\frac{(1-c_N)\rho_1}{K^2} \preceq \nabla^2_\ell f_{\pi^\star}(\theta) \preceq \frac{7}{K}. \tag{48}$$

Lemma 7 (Upper bound of the error gradient)**.** *Let $f_{\pi^\star}$ be the function defined in* (29)*. Let g_t be the function defined in* (32)*. Then, with probability at least $1 - q^{-K_{\ell-1}}$, we have*

$$\begin{aligned}\Big\|\nabla_\ell f_{\pi^\star}(\theta) - g_\ell(\theta^{(t)}; \mathcal{D}_t)\Big\|_2 \lesssim & \frac{1}{K_\ell} \cdot \|\theta - \theta^\star\|_2 \cdot \sqrt{\frac{K_{\ell-1}\log q}{|\mathcal{D}_t|}} + \frac{\gamma}{K_\ell} \cdot \|\theta^{(t)} - \theta^\star\|_2 \\ & + \frac{R_{\max}}{1-\gamma} \cdot (1+\gamma)\tau^\star \cdot \eta_{t-\tau^\star} \\ & + |\mathcal{A}| \cdot \frac{R_{\max}}{1-\gamma} \cdot (1 + \log_\nu \lambda^{-1} + \frac{1}{1-\nu}) \cdot C_t,\end{aligned} \tag{49}$$

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*

$$\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^\star\|_2 \le \Big(1 - \frac{\rho - c_N}{\phi_{\max}}\Big) \cdot \|\boldsymbol{w}^{(t)} - \boldsymbol{w}^\star\|_2. \tag{50}$$

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

$$\begin{aligned}\theta^{(t+1)} =& \theta^{(t)} - \eta_t \cdot g^{(t)}(\theta^{(t)}; \mathcal{D}_t) \\ =& \theta^{(t)} - \eta_t \cdot \nabla f(\theta^{(t)}) + \eta_t \cdot \big(\nabla f(\theta^{(t)}) - g^{(t)}(\theta^{(t)}; \mathcal{D}_t)\big).\end{aligned} \tag{51}$$

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

$$\begin{aligned}\nabla f(\theta^{(t)}) =& \nabla f(\theta^{(t)}) - \nabla f(\theta^\star) \\ =& \int_0^1 \nabla^2 f\Big(\theta^{(t)} + u \cdot (\theta^{(t)} - \theta^\star)\Big) du \cdot (\theta^{(t)} - \theta^\star),\end{aligned} \tag{52}$$

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

$$\boldsymbol{A}^{(t)} := \int_0^1 \nabla^2 f\Big(\theta^{(t)} + u \cdot (\theta^{(t)} - \theta^\star)\Big) du. \tag{53}$$

Then, we have

$$\begin{aligned}\|\theta^{(t+1)} - \theta^\star\|_2 \le & \|\boldsymbol{I} - \eta_t \boldsymbol{A}^{(t)}\|_2 \cdot \|\theta^{(t)} - \theta^\star\|_2 + \eta_t \cdot \|\nabla f(\theta^{(t)}) - g^{(t)}(\theta^{(t)}; \mathcal{D}_t)\|_2 \\ \le & \|\boldsymbol{I} - \eta_t \boldsymbol{A}^{(t)}\|_2 \cdot \|\theta^{(t)} - \theta^\star\|_2 + \eta_t \cdot \sum_{\ell=1}^{L} \Big\|\nabla_\ell f(\theta^{(t)}) - g^{(t)}(\theta_\ell^{(t)}; \mathcal{D}_t)\Big\|_2.\end{aligned} \tag{54}$$

From Lemma 6, we have

$$\|\boldsymbol{I} - \eta_t \boldsymbol{A}^{(t)}\|_2 \le 1 - \eta_t \cdot \frac{(1-c_N)\cdot\rho_1}{K^2}. \tag{55}$$

From Lemma 7, we have

$$\begin{aligned}\Big\|\nabla_\ell f_{\pi^\star}(\theta^{(t)}) - g_\ell(\theta^{(t)};\mathcal{D}_t)\Big\|_2 \lesssim& \frac{1}{K_\ell}\cdot\|\theta^{(t)}-\theta^\star\|_2\cdot\sqrt{\frac{K_{\ell-1}\log q}{|\mathcal{D}_t|}} + \frac{\gamma}{K_\ell}\cdot\|\theta^{(t)}-\theta^\star\|_2 \\ &+ \frac{R_{\max}}{1-\gamma}\cdot(1+\gamma)\tau^\star\cdot\eta_{t-\tau^\star} \\ &+ |\mathcal{A}|\cdot\frac{R_{\max}}{1-\gamma}\cdot(1+\log_\nu\lambda^{-1}+\frac{1}{1-\nu})\cdot C_t.\end{aligned} \tag{56}$$

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

$$|\mathcal{D}_t| \gtrsim\ c_N^{-2}\cdot\rho_1^{-1}\cdot\big(\sum_{\ell=1}^{L}K_\ell\sqrt{K_{\ell-1}}\big)^2\cdot\log q, \tag{57}$$

(54) can be simplified as

$$\|\theta^{(t+1)}-\theta^\star\|_2 \le (1-\eta_t\cdot\xi)\cdot\|\theta^{(t)}-\theta^\star\|_2 + \eta_t\cdot\Delta_t + \eta_t\cdot C^\star\|\boldsymbol{w}^{(t)}-\boldsymbol{w}^\star\|_2. \tag{58}$$

where

$$\begin{aligned}C^\star &= |\mathcal{A}|\cdot\frac{R_{\max}}{1-\gamma}\cdot(1+\log_\nu\lambda^{-1}+\frac{1}{1-\nu})\cdot C \\ \xi &= \frac{(1-\gamma-c_N)\rho_1}{K^2} - C^\star \\ \Delta_t &= \frac{R_{\max}}{1-\gamma}\cdot(1+\gamma)\tau^\star\cdot\eta_{t-\tau^\star}.\end{aligned} \tag{59}$$

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

$$(t+1)\cdot\|\theta^{(t+1)}-\theta^\star\|_2 \le t\cdot\|\theta^{(t)}-\theta^\star\|_2 + \xi^{-1}\cdot\Delta_t + \xi^{-1}\cdot C^\star\|\boldsymbol{w}^{(t)}-\boldsymbol{w}^\star\|_2. \tag{60}$$

Next, we have

$$\begin{aligned}&\sum_{t=0}^{T-1}(t+1)\cdot\|\theta^{(t+1)}-\theta^\star\|_2 - t\cdot\|\theta^{(t)}-\theta^\star\|_2 \\ \le&\sum_{t=0}^{T-1}\xi^{-1}\cdot(\Delta_t + C^\star\|\boldsymbol{w}^{(t)}-\boldsymbol{w}^\star\|_2).\end{aligned} \tag{61}$$

With the definition of Δ_t in (59), we have

$$\begin{aligned}\sum_{t=0}^{T-1}\Delta_t \le&\sum_{t=0}^{\tau^\star}\Delta_t + \sum_{t=\tau^\star}^{T-1}\lambda^{-1}\cdot\Delta_t \\ \le&\sum_{t=0}^{\tau^\star}\tau^\star\cdot\frac{R_{\max}}{1-\gamma} + \sum_{t=\tau^\star}^{T-1}\cdot\frac{R_{\max}\cdot(1+\gamma)}{1-\gamma}\cdot\tau^\star\cdot\frac{1}{T-\tau^\star+1} \\ \lesssim&\frac{R_{\max}\cdot\log^2 T}{1-\gamma} + \frac{R_{\max}\cdot(1+\gamma)\cdot\log^2 T}{1-\gamma}.\end{aligned} \tag{62}$$

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

$$\sum_{t=0}^{T-1}\|\boldsymbol{w}^{(t)}-\boldsymbol{w}^\star\|_2 \lesssim \|\boldsymbol{w}^{(0)}-\boldsymbol{w}^\star\|_2. \tag{63}$$

By multiplying $1/T$ on both sides of (61), we have

$$\|\theta^{(T)} - \theta^\star\|_2 \leq \frac{(2+\gamma)\cdot R_{\max}\cdot \log^2 T + C^\star\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^\star\|_2}{(1-\gamma-c_N)\rho_1 K^{-2} - C^\star}\cdot\frac{1}{T}. \tag{64}$$

□

C. Proofs of Theorems 3 and 4

Lemma 9. *We have*

$$\left|Q_i^{\pi_i^\star}(\boldsymbol{s},a) - Q_i^{\pi_j^\star}(\boldsymbol{s},a)\right| \leq \frac{2\gamma}{1-\gamma}\cdot\max_{\boldsymbol{s},a}|r_i(\boldsymbol{s},a) - r_j(\boldsymbol{s},a)|. \tag{65}$$

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

$$\begin{aligned} Q_{n+1}^{\pi_{n+1}}(\boldsymbol{s},a) - Q_{n+1}^{\pi_j}(\boldsymbol{s},a) &= \max_{i\in[n]} Q_{n+1}^{\pi_i^\star}(\boldsymbol{s},a) - Q_{n+1}^{\pi_j}(\boldsymbol{s},a) \\ &\geq Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a) - Q_{n+1}^{\pi_j}(\boldsymbol{s},a) \\ &= \big(\psi_j(\Theta_j^\star) - \psi_j(\Theta_j^{(T)})\big)\cdot\boldsymbol{w}_{n+1}^\star. \end{aligned} \tag{66}$$

According to Theorem 1, we have

$$\|\psi_j(\Theta_j^\star) - \psi_j(\Theta_j^{(T)})\|_2 \leq \frac{(2+\gamma)\cdot R_{\max}\cdot\log^2 T + C^\star\|\boldsymbol{w}_j^{(0)} - \boldsymbol{w}_j^\star\|_2}{(1-\gamma-c_N)\rho_1 K^{-2} - C^\star}\cdot\frac{1}{T} := \frac{C_3}{T} \tag{67}$$

Then, we have

$$\mathcal{T}^\pi Q_{n+1}^{\pi_j}(\boldsymbol{s},a) \geq Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma\cdot\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}. \tag{68}$$

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

$$\begin{aligned} Q_{n+1}^{\pi_j}(\boldsymbol{s},a) =& \lim_{k\to\infty}(\mathcal{T}^\pi)^k Q_{n+1}^{\pi_j}(\boldsymbol{s},a) \\ \geq& \lim_{k\to\infty}(\mathcal{T}^\pi)^{k-1}\Big(Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}\Big) \\ =& \lim_{k\to\infty}(\mathcal{T}^\pi)^{k-2}\cdot\mathcal{T}^\pi\Big(Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}\Big) \\ =& \lim_{k\to\infty}(\mathcal{T}^\pi)^{k-2}\cdot\Big(\mathcal{T}^\pi Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma^2\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}\Big) \\ =& \lim_{k\to\infty}(\mathcal{T}^\pi)^{k-2}\Big(Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \gamma\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} - \gamma^2\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}\Big) \\ =& Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \sum_{k=1}^{\infty}\gamma^k\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} \\ =& Q_{n+1}^{\pi_j}(\boldsymbol{s},a) - \frac{\gamma}{1-\gamma}\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} \\ \geq& Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a) - \frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} - \frac{\gamma}{1-\gamma}\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T} \\ =& Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a) - \frac{1}{1-\gamma}\frac{C_3\|\boldsymbol{w}_{n+1}^\star\|_2}{T}. \end{aligned} \tag{69}$$

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

$$Q_{n+1}^\star(\boldsymbol{s},a) - Q_{n+1}^{\pi_{n+1}}(\boldsymbol{s},a) = \big(Q_{n+1}^\star(\boldsymbol{s},a) - Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a)\big) + \big(Q_{n+1}^{\pi_j^\star}(\boldsymbol{s},a) - Q_{n+1}^{\pi_j}(\boldsymbol{s},a)\big). \tag{70}$$

From Lemma 9, we have

$$Q^{\star}_{n+1}(\boldsymbol{s},a)-Q^{\pi^{\star}_j}_{n+1}(\boldsymbol{s},a)\leq\frac{2\gamma}{1-\gamma}\cdot\max_{\boldsymbol{s},a}|r_{n+1}(\boldsymbol{s},a)-r_j(\boldsymbol{s},a)|, \tag{71}$$

and (69) suggests that

$$Q^{\pi^{\star}_j}_{n+1}(\boldsymbol{s},a)-Q^{\pi_j}_{n+1}(\boldsymbol{s},a)\leq\frac{C_3\|\boldsymbol{w}^{\star}_{n+1}\|_2}{(1-\gamma)T}. \tag{72}$$

Therefore, (70) can be upper bounded as

$$\begin{aligned}
&Q^{\star}_{n+1}(\boldsymbol{s},a)-Q^{\pi_{n+1}}_{n+1}(\boldsymbol{s},a)\\
&=\big(Q^{\star}_{n+1}(\boldsymbol{s},a)-Q^{\pi^{\star}_j}_{n+1}(\boldsymbol{s},a)\big)+\big(Q^{\pi^{\star}_j}_{n+1}(\boldsymbol{s},a)-Q^{\pi_j}_{n+1}(\boldsymbol{s},a)\big)\\
&\leq\frac{2\gamma}{1-\gamma}\cdot\max_{\boldsymbol{s},a}|r_{n+1}(\boldsymbol{s},a)-r_j(\boldsymbol{s},a)|+\frac{C_3\|\boldsymbol{w}^{\star}_{n+1}\|_2}{(1-\gamma)T}\\
&\leq\frac{2\gamma\cdot\phi_{\max}}{1-\gamma}\|\boldsymbol{w}_{n+1}-\boldsymbol{w}_j\|_2+\frac{C_3\|\boldsymbol{w}^{\star}_{n+1}\|_2}{(1-\gamma)T}.
\end{aligned} \tag{73}$$

Since (73) holds for any j, we have

$$|Q^{\star}_{n+1}(\boldsymbol{s},a)-Q^{\pi_{n+1}}_{n+1}(\boldsymbol{s},a)|\leq\frac{2\gamma\cdot\phi_{\max}}{1-\gamma}\min_{j\in[n]}\|\boldsymbol{w}_{n+1}-\boldsymbol{w}_j\|_2+\frac{C_3\|\boldsymbol{w}^{\star}_{n+1}\|_2}{(1-\gamma)T}. \tag{74}$$

□

Proof of Theorem 4. Let π'_{n+1} be generalized policy with DQN via GPI. Similar to (66), 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

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

□

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

$$Q_i^\star(\boldsymbol{s},a) - Q_i^{\pi_j}(\boldsymbol{s},a) \le \frac{2\gamma\cdot\phi_{\max}}{1-\gamma}\|\boldsymbol{w}_j-\boldsymbol{w}_i\|_2 + \frac{C_3\|\boldsymbol{w}_i^\star\|_2}{(1-\gamma)T}. \tag{77}$$

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

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

Therefore, at iteration t for task i, we have

$$\begin{aligned} C_t =& |Q_i^\star(\boldsymbol{s},a) - Q_i^{\pi_i^{(t)}}| \\ \le& \min\left\{\frac{2\gamma\cdot\phi_{\max}}{1-\gamma}\min_{1\le j\le i}\|\boldsymbol{w}_j-\boldsymbol{w}_i\|_2 + \frac{C_3\|\boldsymbol{w}_i^\star\|_2}{(1-\gamma)T}, \|\Theta_i^{(t)}-\Theta_i^\star\|_2\cdot\|\boldsymbol{w}_i^\star\|_2\right\} \\ \lesssim& \min\left\{\frac{2\gamma\cdot\phi_{\max}}{1-\gamma}\min_{1\le j\le i}\|\boldsymbol{w}_j-\boldsymbol{w}_i\|_2, \|\Theta_i^{(t)}-\Theta_i^\star\|_2\cdot\|\boldsymbol{w}_i^\star\|_2\right\} \text{ (As } T \text{ is sufficiently large)} \\ =& \min\{q_t,1\}\cdot\|\Theta_i^{(t)}-\Theta_i^\star\|_2, \end{aligned} \tag{79}$$

where

$$q_t = \frac{2\gamma\cdot R_{\max}}{1-\gamma}\cdot\frac{\min_{1\le i\le j-1}\ \|\boldsymbol{w}_i^\star-\boldsymbol{w}_j^\star\|_2}{\|\Theta_j^{(t)}-\Theta_j^\star\|_2}. \tag{80}$$

Following similar steps in (58) in the proof of Theorem 1, with C_t satisfying (79), we have

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

□

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.

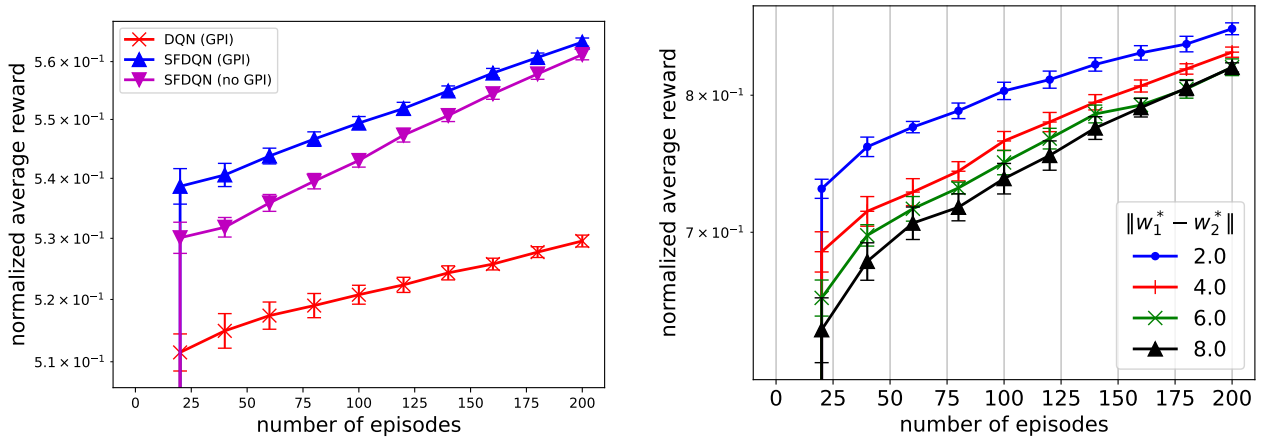
E.1. Synthetic data settings

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 (9) 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. Additional experiments on synthetic RL benchmarks

Comparison for transfer from multiple source tasks. In addition to the single source task case discussed in Section 5, we also investigate the transfer performance of SFDQN (with and without GPI) and DQN (GPI) agents when trained on multiple source tasks. For this purpose, we generate ϕ as described in the previous section, and generate additional source tasks and a target task by pertubing $\boldsymbol{w}_1^\star$. Thus, we obtain $\boldsymbol{w}_2^\star$, $\boldsymbol{w}_2^\star$, $\boldsymbol{w}_2^\star$, the reward vectors for three additional source tasks. The norm of all weight vectors is set to 1 to make sure the reward scales are similar across multiple source tasks. Then we train each learning agent on the four source tasks and apply transfer using GPI. Note that while we also test the case for SFDQN without GPI (thus no transfer), this agent leverages the similarity of source tasks and the target task. The results are shown in Figure 3a. It can be seen that the SFDQN agent performs the best, which can leverage the task closeness due to the proximity of source task weight vectors to that of the target task, and the transition dynamics information via GPI. SFDQN agent without GPI on the other hand can only leverage the task closeness due to the proximity of source weight vectors. DQN-GPI agent cannot leverage the task closeness information or the transition dynamic information explicitly as the SFDN agent, and hence performs worse.

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 3b. 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.



(a) Comparison of four different source task transfer performance for SF-DQN (with and without GPI) and DQN with GPI.

(b) Effect $\|w_1^\star - w_2^\star\|$ on the convergence of SF-DQN agent when training on task 2 with GPI

Figure 3: Additional experiments on synthetic environment

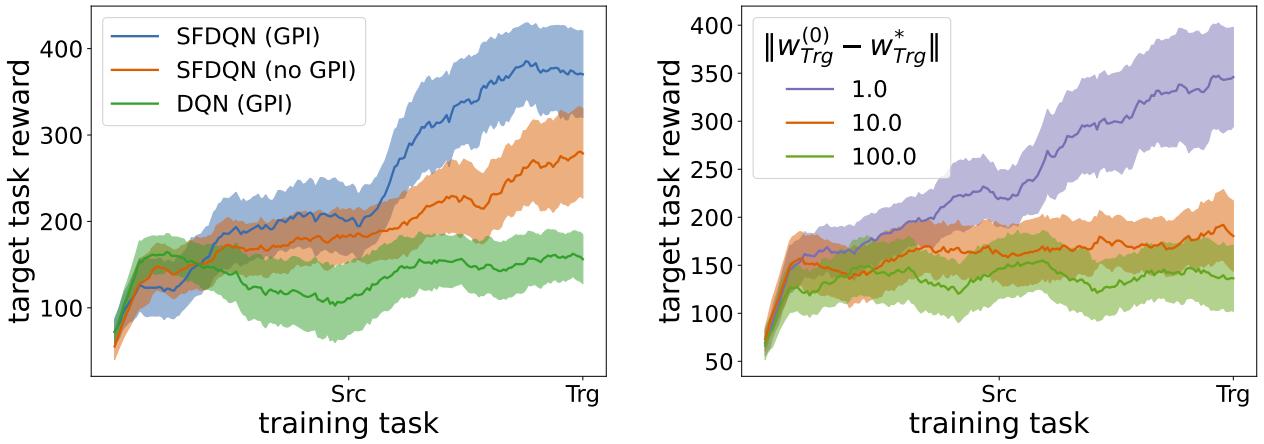
E.3. Real Data: Reacher environment

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 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 are 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`.

Comparison of SF-DQN (with and without GPI) and DQN (GPI). We first provide a comparison of the performance of SF-DQN with GPI, SF-DQN without GPI, and DQN with GPI, in Figure 4a. 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.

Effect of $\|w_{Trg}^{(0)} - w_{Trg}^\star\|$. 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 4b. It can be seen that when the reward mappings are initialized far away from the true reward mappings, the convergence of the SF-DQN



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

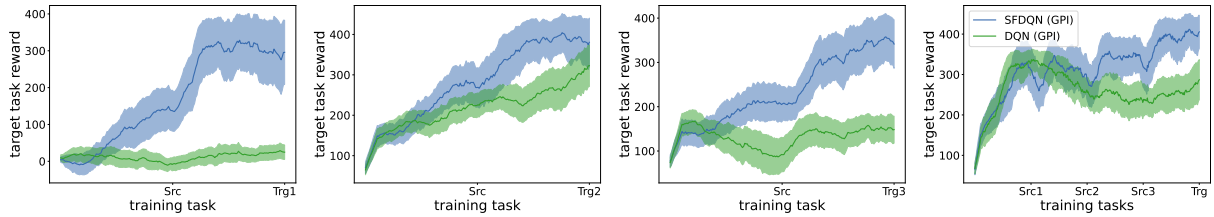
(b) Comparison of different initializations for target task reward mapping

Figure 4: Single source to single target task transfer experiments on Reacher environment

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.

Single source task to multiple target tasks transfer learning. Next we compare the performance of SFDQN and DQN with GPI for transferring knowledge from single source tasks to multiple target tasks. The results are given in Figures 5a, 5b, and 5c. It can be seen that SFDQN outperforms DQN significantly for most target tasks, which shows the efficacy of knowledge transfer in SFDQN with GPI. The gap of performance seems to be different for different target tasks, suggesting that the performance gain for SFDQN with GPI can vary depending on the source and target task relationship.

Multiple source tasks to single target task transfer learning. Next we investigate the effect of GPI for transferring knowledge from multiple source tasks to single target task. The results are given in Figure 3a. It can be seen that SFDQN outperforms DQN significantly, which shows the efficacy of knowledge transfer in SFDQN with GPI.



(a) Source to target task 1 transfer performance

(b) Source to target task 2 transfer performance

(c) Source to target task 3 transfer performance

(d) Multiple source to target task transfer performance

Figure 5: Multple source/target tasks transfer experiments on Reacher environment

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 (29). According to Weyl's inequality in Lemma 1, the eigenvalues of $\nabla^2_\ell f(\cdot)$ at any fixed point θ can be bounded in the form of (86). Therefore, we first provide the lower and upper bounds for $\nabla^2_\ell f$ at the desired ground truth $\theta^\star$. Then, the bounds for $\nabla^2_\ell f$ at any other point θ is bounded through (29) 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* (29)*. If θ is close to $\theta^\star$ such that*

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

we have

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

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 $\mathbb{E}_{\boldsymbol{h}}\|\boldsymbol{h}\|_2^2 \leq +\infty$.*

Let $\boldsymbol{\alpha} \in \mathbb{R}^{K_\ell K_{\ell-1}}$ be the unit vector defined in (35)*, 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* (36)*. 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 $\nabla^2_\ell f(\theta)$ at θ, respectively. Then, from Lemma 1, we have

$$\begin{aligned} \lambda_{\max}(\theta) &\leq \lambda_{\max}(\theta^\star) + \|\nabla^2_\ell f(\theta) - \nabla^2_\ell f(\theta^\star)\|_2, \\ \lambda_{\min}(\theta) &\geq \lambda_{\min}(\theta^\star) - \|\nabla^2_\ell f(\theta) - \nabla^2_\ell f(\theta^\star)\|_2. \end{aligned} \tag{86}$$

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

$$\begin{aligned} & \min_{\|\boldsymbol{\alpha}\|_2=1} \boldsymbol{\alpha}^\top \nabla^2_\ell f(\theta^\star) \boldsymbol{\alpha} \\ = \ & \frac{1}{K^2} \min_{\|\boldsymbol{\alpha}\|_2=1} \mathbb{E}_{\boldsymbol{h}\sim\mathcal{P}} \Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^\top \boldsymbol{h}^{(\ell)} \mathcal{J}_{\ell,k} \phi'(\theta_{\ell,j}^{\star\top} \boldsymbol{h}^{(\ell)}) \Big)^2 \\ = \ & \frac{1}{K^2} \min_{\|\boldsymbol{\alpha}\|_2=1} \int_{\mathbb{R}^{K_{\ell-1}}} \Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^\top \boldsymbol{h}^{(\ell)} \mathcal{J}_{\ell,k} \phi'(\theta_{\ell,j}^{\star\top} \boldsymbol{h}^{(\ell)}) \Big)^2 p_H(\boldsymbol{h}^{(\ell)}) \cdot d\mathbf{h}^{(\ell)} \\ = \ & \frac{1}{K^2} \min_{\|\boldsymbol{\alpha}\|_2=1} \int_{\{\boldsymbol{h}^{(\ell)} \mid \mathcal{J}_{\ell,k} \neq 0\}} \Big(\sum_{j=1}^K \boldsymbol{\alpha}_j^\top \boldsymbol{h}^{(\ell)} \phi'(\theta_{\ell,j}^{\star\top} \boldsymbol{h}^{(\ell)}) \Big)^2 p_H(\boldsymbol{h}^{(\ell)}) \cdot d\mathbf{h}^{(\ell)} \\ \gtrsim \ & \frac{\rho_1}{K^2}, \end{aligned} \tag{87}$$

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 (29). 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 (32), 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 (29), 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

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

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

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

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

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

From Lemma 3, we know that $|\mathcal{S}_{\frac{1}{2}}(d)|\leq 5^d$. 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

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

Then, we have

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

Therefore, we have

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

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 Lemma 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

Proof of Lemma 9. $\left|Q_i^{\pi_i^\star}(\boldsymbol{s}, a) - Q_i^{\pi_j^\star}(\boldsymbol{s}, a)\right|$ 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}(\boldsymbol{s}', a') - \max_{a'} Q_j^{\pi_j^\star}(\boldsymbol{s}', a')\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}(\boldsymbol{s}', a') - Q_j^{\pi_j^\star}(\boldsymbol{s}', a')\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}(\boldsymbol{s}', a') - Q_j^{\pi_j^\star}(\boldsymbol{s}', a')\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{129}$$

Let

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

and

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

Then, we have

$$\begin{aligned} I_5 =& \max_{\boldsymbol{s},a} \left| r_i + \gamma \cdot \sum_{\boldsymbol{s}'} p^a_{\boldsymbol{s},\boldsymbol{s}'} \max_{a'} Q_i^{\pi_i^\star}(\boldsymbol{s}', a') - r_j - \gamma \cdot \sum_{\boldsymbol{s}'} p^a_{\boldsymbol{s},\boldsymbol{s}'} \max_{a'} Q_j^{\pi_j^\star}(\boldsymbol{s}', a') \right| \\ \leq& \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s}, a) - r_j(\boldsymbol{s}, a)| + \gamma \max_{\boldsymbol{s},a} \sum_{\boldsymbol{s}'} p^a_{\boldsymbol{s},\boldsymbol{s}'} \cdot \max_{a'} |Q_i^{\pi_i^\star}(\boldsymbol{s}', a') - Q_j^{\pi_j^\star}(\boldsymbol{s}', a')| \\ \leq& \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s}, a) - r_j(\boldsymbol{s}, a)| + \gamma \cdot I_5. \end{aligned} \tag{130}$$

Therefore, we have

$$I_5 \leq \frac{1}{1-\gamma} \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s}, a) - r_j(\boldsymbol{s}, a)|. \tag{131}$$

Similar to (130), we have

$$\begin{aligned} I_6 \leq& \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s}, a) - r_j(\boldsymbol{s}, a)| + \gamma \max_{\boldsymbol{s},a} \sum_{\boldsymbol{s}'} p^a_{\boldsymbol{s},\boldsymbol{s}'} \cdot |Q_j^{\pi_j^\star}(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}')) - Q_i^{\pi_j^\star}(\boldsymbol{s}', \pi_j^\star(\boldsymbol{s}'))| \\ \leq& \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s}, a) - r_j(\boldsymbol{s}, a)| + \gamma \cdot I_6. \end{aligned} \tag{132}$$

Therefore, we have

$$I_6 \leq \frac{1}{1-\gamma} \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s}, a) - r_j(\boldsymbol{s}, a)|. \tag{133}$$

Therefore, we have

$$\left|Q_i^{\pi_i^\star}(\boldsymbol{s}, a) - Q_j^{\pi_i^\star}(\boldsymbol{s}, a)\right| \leq \gamma(I_5 + I_6) \leq \frac{2\gamma}{1-\gamma} \cdot \max_{\boldsymbol{s},a} |r_i(\boldsymbol{s}, a) - r_j(\boldsymbol{s}, a)|. \tag{134}$$

□

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 (136). 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 (29), we have

$$\begin{aligned} &\frac{\partial^2 f}{\partial \theta_{\ell,j_1} \partial \theta_{\ell,j_2}}(\theta^\star) = \frac{1}{K^2} \mathbb{E}_{\boldsymbol{x}} \mathcal{J}_{\ell,k} \sigma'(\theta_{j_1}^{\star\top} \boldsymbol{h}) \cdot \mathcal{J}_{\ell,k} \sigma'(\theta_{j_2}^{\star\top} \boldsymbol{h}) \cdot \boldsymbol{h}^\star \boldsymbol{h}^{\star\top}, \\ \text{and} \quad &\frac{\partial^2 f}{\partial \theta_{\ell,j_1} \partial \theta_{\ell,j_2}}(\theta) = \frac{1}{K^2} \mathbb{E}_{\boldsymbol{x}} \sigma' \mathcal{J}^\star_{\ell,k} (\theta_{\ell,j_1}^\top \boldsymbol{h}) \cdot \mathcal{J}^\star_{\ell,k} \sigma'(\theta_{\ell,j_2}^\top \boldsymbol{h}) \cdot \boldsymbol{h}\boldsymbol{h}^\top, \end{aligned} \tag{135}$$

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

Then, we have

$$\begin{aligned} &\frac{\partial^2 f}{\partial \theta_{\ell,j_1} \partial \theta_{\ell,j_2}}(\theta^*) - \frac{\partial^2 f}{\partial \theta_{\ell,j_1} \partial \theta_{\ell,j_2}}(\theta) \\ =& \frac{1}{K^2} \mathbb{E}_{\boldsymbol{x}} \big[\mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_1} \boldsymbol{h}^\star) \mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}^\star) \boldsymbol{h}^\star \boldsymbol{h}^{\star\top} \\ &\qquad - \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_1} \boldsymbol{h}) \mathcal{J}_{\ell,k} \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h}) \boldsymbol{h}\boldsymbol{h}^\top \big] \\ =& \frac{1}{K^2} \mathbb{E}_{\boldsymbol{x}} \big[\mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_1} \boldsymbol{h}^\star) \big(\mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}^\star) \boldsymbol{h}^\star \boldsymbol{h}^{\star\top} - \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h}) \boldsymbol{h}\boldsymbol{h}^\top \big) \\ &\qquad + \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h}) \big(\mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_1} \boldsymbol{h}) \boldsymbol{h}^\star \boldsymbol{h}^{\star\top} - \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_1} \boldsymbol{h}) \boldsymbol{h}\boldsymbol{h}^\top \big) \big] \\ :=& \frac{1}{K^2} (\boldsymbol{P}_1 + \boldsymbol{P}_2). \end{aligned} \tag{136}$$

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

$$\boldsymbol{a}^\top \boldsymbol{P}_1 \boldsymbol{a} = \mathbb{E}_{\boldsymbol{x}} \mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_1} \boldsymbol{h}^\star) \Big(\mathcal{J}^\star_{\ell,k} \sigma'(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}^\star)(\boldsymbol{a}^\top \boldsymbol{h}^\star)^2 - \mathcal{J}_{\ell,k} \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h})(\boldsymbol{a}^\top \boldsymbol{h})^2 \Big). \tag{137}$$

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{138}$$

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

$$\begin{aligned}
\theta^{\star \top}_{\ell,j_2} \boldsymbol{h} =& \theta^{\star \top}_{\ell,j_2} (z_1 \boldsymbol{a} + z_2 \boldsymbol{b} + z_3 \boldsymbol{c} + \cdots + z_d \boldsymbol{a}_d^\perp) \\
=& z_1 \theta^{\star \top}_{\ell,j_2} \boldsymbol{a} + z_2 \theta^{\star \top}_{\ell,j_2} \boldsymbol{b} + z_3 \theta^{\star \top}_{\ell,j_2} \boldsymbol{c} + \cdots + z_d \theta^{\star \top}_{\ell,j_2} \boldsymbol{a}_d^\perp \\
=& z_1 \theta^{\star \top}_{\ell,j_2} \boldsymbol{a} + z_2 \theta^{\star \top}_{\ell,j_2} \boldsymbol{b} + z_3 \theta^{\star \top}_{\ell,j_2} \boldsymbol{c} + 0 \\
=& \theta^{\star \top}_{\ell,j_2} (z_1 \boldsymbol{a} + z_2 \boldsymbol{b} + z_3 \boldsymbol{c}) \\
:=& \theta^{\star \top}_{\ell,j_2} \widetilde{\boldsymbol{h}}.
\end{aligned} \tag{139}$$

where $\widetilde{\boldsymbol{h}} = z_1 \boldsymbol{a} + z_2 \boldsymbol{b} + z_3 \boldsymbol{c}$. Similar to (139), 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

$$\begin{aligned}
I_4 :=& \mathbb{E}_{\boldsymbol{h}} \Big| \big(\sigma'(\theta^{\star \top}_{\ell,j_2} \boldsymbol{h}) - \sigma'(\theta^\top_{\ell,j_2} \boldsymbol{h}) \big) \cdot \big(\boldsymbol{a}^\top \boldsymbol{h} \big) \Big| \\
=& \int_{\mathcal{R}_{\boldsymbol{h}}} |\sigma'\big(\theta^\top_{\ell,j_2} \boldsymbol{h}\big) - \sigma'\big(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}\big)| \cdot |\boldsymbol{a}^\top \boldsymbol{h}|^2 \cdot f_H(\boldsymbol{h}) d\boldsymbol{h} \\
=& \int_{\mathcal{R}_{\boldsymbol{z}}} |\sigma'\big(\theta^\top_{\ell,j_2} \boldsymbol{h}\big) - \sigma'\big(\theta^{\star T}_{\ell,j_2} \boldsymbol{h}\big)| \cdot |\boldsymbol{a}^\top \boldsymbol{h}|^2 \cdot f_Z(\boldsymbol{z}) \cdot |\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{z})| d\boldsymbol{z}
\end{aligned} \tag{140}$$

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 (139), I_4 can be rewritten as

$$\begin{aligned}
I_4 =& \int_{\mathcal{R}_z} |\sigma'\big(\theta^\top_{\ell,j_2} \widetilde{\boldsymbol{h}}\big) - \sigma'\big(\theta^{\star T}_{\ell,j_2} \widetilde{\boldsymbol{h}}\big)| \cdot |\boldsymbol{a}^\top \widetilde{\boldsymbol{h}}|^2 \cdot f_Z(\boldsymbol{z}) d\boldsymbol{z} \\
=& \int_{\mathcal{R}_z} |\sigma'\big(\theta^\top_{\ell,j_2} \widetilde{\boldsymbol{h}}\big) - \sigma'\big(\theta^{\star T}_{\ell,j_2} \widetilde{\boldsymbol{h}}\big)| \cdot |\boldsymbol{a}^\top \widetilde{\boldsymbol{h}}|^2 \cdot f_Z(z_1, z_2, z_3) dz_1 dz_2 dz_3
\end{aligned} \tag{141}$$

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

$$I_4 = \int |\sigma'\big(\theta_{\ell,j_2}^{\top}\widetilde{\boldsymbol{h}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star\top}\widetilde{\boldsymbol{h}}\big)| \cdot |R\cos\sigma_1|^2 \cdot f_Z(R,\sigma_1,\sigma_2) \cdot R^2\sin\sigma_1 \cdot dRd\sigma_1 d\sigma_2. \tag{142}$$

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

$$f_Z(R,\sigma_1,\sigma_2) = f_Z(R). \tag{143}$$

Then, we have

$$\begin{aligned} I_4 &= \int |\sigma'\big(\theta_{\ell,j_2}^{\top}(\widetilde{\boldsymbol{h}}/R)\big) - \sigma'\big(\theta_{\ell,j_2}^{\star T}(\widetilde{\boldsymbol{h}}/R)\big)| \cdot |R\cos\sigma_1|^2 \cdot f_Z(R)R^2\sin\sigma_1 dRd\sigma_1 d\sigma_2 \\ &= \int_0^{\infty} R^4 f_z(R)dR \int_0^{\psi_1(R)}\int_0^{\psi_2(R)} |\cos\sigma_1|^2 \cdot \sin\sigma_1 \\ &\quad \cdot |\sigma'\big(\theta_{\ell,j_2}^{\top}(\widetilde{\boldsymbol{h}}/R)\big) - \sigma'\big(\theta_{\ell,j_2}^{\star T}(\widetilde{\boldsymbol{h}}/R)\big)| d\sigma_1 d\sigma_2 \\ &\le \int_0^{\infty} R^4 f_z(R)dR \int_0^{\pi}\int_0^{2\pi} \sin\sigma_1 \cdot |\sigma'\big(\theta_{\ell,j_2}^{\top}\bar{\boldsymbol{x}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star T}\bar{\boldsymbol{x}}\big)| d\sigma_1 d\sigma_2, \end{aligned} \tag{144}$$

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 $F_z(R) \ge 1 - 2e^{-\frac{R^2}{\sigma^2}}$ for some constant $\sigma > 0$. Then, the integration of R can be represented as

$$\begin{aligned} \int_0^{\infty} R^4 f_Z(R)dR &= \int_0^{\infty} R^4 d\big(1 - F_z(R)\big) \\ &\le \int_0^{\infty} 4R^3\big(1 - F_z(R)\big)dR \\ &\le \int_0^{\infty} 8R^3 e^{-\frac{R^2}{\sigma^2}} dR \\ &\le \frac{32}{\sqrt{2\pi}}\sigma \int_0^{\infty} R^2 e^{-\frac{R^2}{\sigma^2}} dR \\ &= 32\sigma^2 \int_0^{\infty} R^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{R^2}{\sigma^2}} dR, \end{aligned} \tag{145}$$

where the last inequality comes from the calculation that

$$\begin{aligned} \int_0^{\infty} 2R^2 e^{-\frac{R^2}{\sigma^2}} dR &= \sqrt{2\pi}\sigma^3, \\ \int_0^{\infty} 2R^3 e^{-\frac{R^2}{\sigma^2}} dR &= 4\sigma^4. \end{aligned} \tag{146}$$

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

$$\begin{aligned} I_4 &\le 32\sigma^2 \cdot \int_0^{\infty} R^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{R^2}{\sigma^2}} dR \int_0^{\pi}\int_0^{2\pi} \sin\sigma_1 \cdot |\sigma'\big(\theta_{\ell,j_2}^{\top}\bar{\boldsymbol{x}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star\top}\bar{\boldsymbol{x}}\big)| d\sigma_1 d\sigma_2 \\ &= 32\sigma^2 \cdot \mathbb{E}_{z_1,z_2,z_3}\big|\sigma'\big(\theta_{\ell,j_2}^{\top}\widetilde{\boldsymbol{x}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star\top}\widetilde{\boldsymbol{x}}\big)| \\ &\eqsim \mathbb{E}_{\widetilde{\boldsymbol{x}}}\big|\sigma'\big(\theta_{\ell,j_2}^{\top}\widetilde{\boldsymbol{x}}\big) - \sigma'\big(\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}}\big)|, \end{aligned} \tag{147}$$

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

$$|\sigma'(\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}})-\sigma'(\theta_{\ell,j_2}^{\star\top}\widetilde{\boldsymbol{x}})|=\begin{cases}1,\ \text{if}\ \widetilde{\boldsymbol{x}}\in\mathcal{A}_1\\ 0,\ \text{if}\ \widetilde{\boldsymbol{x}}\in\mathcal{A}_1^c\end{cases}. \tag{148}$$

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

$$|\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}}|\leq|\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}}-\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}}|\leq\|\theta_{\ell,j_2}^\star-\theta_{\ell,j_2}\|_2\cdot\|\widetilde{\boldsymbol{x}}\|_2. \tag{149}$$

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

$$\begin{aligned}\mathcal{A}_2=&\Big\{\widetilde{\boldsymbol{x}}\Big|\frac{|\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}}|}{\|\theta_{\ell,j_2}^*\|_2\|\widetilde{\boldsymbol{x}}\|_2}\leq\frac{\|\theta_{\ell,j_2}^*-\theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^*\|_2}\Big\}\\ =&\Big\{\theta_{\widetilde{\boldsymbol{x}},\theta_{\ell,j_2}^\star}\Big||\cos\theta_{\widetilde{\boldsymbol{x}},\theta_{\ell,j_2}^\star}|\leq\frac{\|\theta_{\ell,j_2}^\star-\theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2}\Big\}.\end{aligned} \tag{150}$$

Hence, we have that

$$\begin{aligned}\mathbb{E}_{\widetilde{\boldsymbol{x}}}|\sigma'(\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}})-\sigma'(\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}})|^2=&\mathbb{E}_{\widetilde{\boldsymbol{x}}}|\sigma'(\theta_{\ell,j_2}^\top\widetilde{\boldsymbol{x}})-\sigma'(\theta_{\ell,j_2}^{\star T}\widetilde{\boldsymbol{x}})|\\ =&\text{Prob}(\widetilde{\boldsymbol{x}}\in\mathcal{A}_1)\\ \leq&\text{Prob}(\widetilde{\boldsymbol{x}}\in\mathcal{A}_2).\end{aligned} \tag{151}$$

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

$$\begin{aligned}\text{Prob}(\widetilde{\boldsymbol{x}}\in\mathcal{A}_2)=\frac{\pi-\arccos\frac{\|\theta_{\ell,j_2}^\star-\theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2}}{\pi}\leq&\frac{1}{\pi}\tan(\pi-\arccos\frac{\|\theta_{\ell,j_2}^\star-\theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2})\\ =&\frac{1}{\pi}\cot(\arccos\frac{\|\theta_{\ell,j_2}^\star-\theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2})\\ \leq&\frac{2}{\pi}\frac{\|\theta_{\ell,j_2}^\star-\theta_{\ell,j_2}\|_2}{\|\theta_{\ell,j_2}^\star\|_2}\\ \leq&\|\theta_\ell^\star-\theta_\ell\|_2\end{aligned} \tag{152}$$

Hence, (144) and (152) suggest that

$$\begin{aligned}I_4\lesssim\|\theta_i-\theta_i^\star\|_2\cdot\|\boldsymbol{a}\|_2^2,\\ \text{and}\qquad\|\boldsymbol{P}_1\|_2\leq\|\theta-\theta^\star\|_2+I_4\lesssim\|\theta-\theta^\star\|_2,\end{aligned} \tag{153}$$

The same bound that is shown in (153) holds for $\boldsymbol{P}_2$ as well.

Therefore, we have

$$\begin{aligned}\|\nabla_\ell^2 f(\theta^\star) - \nabla_\ell^2 f(\theta)\|_2 =& \max_{\|\boldsymbol{\alpha}\|_2\le 1} \Big|\boldsymbol{\alpha}^\top \Big(\nabla_\ell^2 f(\theta^\star) - \nabla_\ell^2 f(\theta)\Big)\boldsymbol{\alpha}\Big| \\ \le& \frac{1}{K^2}\sum_{j_1=1}^K\sum_{j_2=1}^K \|\boldsymbol{P}_1+\boldsymbol{P}_2\|_2 \cdot \|\boldsymbol{\alpha}_{j_1}\|_2 \cdot \|\boldsymbol{\alpha}_{j_2}\|_2 \\ \lesssim& \frac{1}{K^2}\cdot\sum_{j_1=1}^K\sum_{j_2=1}^K \|\theta-\theta^\star\|_2 \cdot \|\boldsymbol{\alpha}_{j_1}\|_2\|\boldsymbol{\alpha}_{j_2}\|_2 \\ \lesssim& \frac{1}{K^2}\cdot\sum_{j_1=1}^K\sum_{j_2=1}^K \|\theta-\theta^\star\|_2 \cdot \Big(\frac{\|\boldsymbol{\alpha}_{j_1}\|_2^2+\|\boldsymbol{\alpha}_{j_2}\|_2^2}{2}\Big) \\ \lesssim& \frac{1}{K}\cdot\|\theta^\star-\theta\|_2,\end{aligned} \tag{154}$$

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 obvious 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

$$\langle f,g\rangle = \int_{\mathcal{R}} f(\boldsymbol{h})^\top g(\boldsymbol{h}) f_H(\boldsymbol{h})\cdot d\boldsymbol{h},\quad \forall f,g\in\mathcal{H}, \tag{155}$$

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

$$\boldsymbol{\alpha}^\top\boldsymbol{h}\sigma'(\theta_k^\top\boldsymbol{h})\neq 0 \quad \textit{almost everywhere}. \tag{156}$$

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 $\mathbb{E}_{\boldsymbol{h}\sim\mathcal{D}}|\psi_j(\boldsymbol{h})|^2\le\mathbb{E}_{\boldsymbol{h}\sim\mathcal{D}}|\boldsymbol{h}|^2<\infty$.

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

$$\lambda\big(\mathcal{X}_k\cap(\cup_{j\neq k}\mathcal{X}_k)\big)\le\sum_{j\neq k}\lambda(\mathcal{X}_k\cap\mathcal{X}_j)=0. \tag{157}$$

Therefore, we have

$$\lambda\big(\mathcal{X}_k/(\cup_{j\neq k}\mathcal{X}_k)\big)=\lambda(\mathcal{X}_k)-\lambda\big(\mathcal{X}_k\cap(\cup_{j\neq k}\mathcal{X}_k)\big)=\lambda(\mathcal{X}_k)>0. \tag{158}$$

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

$$\mathcal{B}(\boldsymbol{z}_k, r) \cap \mathcal{D}_j = \emptyset \quad \textit{with} \quad \forall r \leq r_0 \textit{ and } j \neq k, \tag{159}$$

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

$$\begin{aligned} \mathcal{B}_r^+ &= \mathcal{B}(\boldsymbol{z}_k, r) \cap \{\boldsymbol{h} \mid \theta_k^\top \boldsymbol{h} > 0\}, \\ \mathcal{B}_r^- &= \mathcal{B}(\boldsymbol{z}_k, r) \cap \{\boldsymbol{h} \mid \theta_k^\top \boldsymbol{h} < 0\}. \end{aligned} \tag{160}$$

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

$$\lim_{r\to 0_+} \frac{1}{\lambda(\mathcal{B}_r^+)} \int_{\mathcal{B}_r^+} \psi_k(\boldsymbol{h}) d\boldsymbol{h} = \lim_{r\to 0_-} \frac{1}{\lambda(\mathcal{B}_r^-)} \int_{\mathcal{B}_r^+} \psi_k(\boldsymbol{h}) d\boldsymbol{h} = \psi_k(\boldsymbol{z}_k). \tag{161}$$

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

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

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

$$\sum_{j=1}^K \alpha_j \psi_j(\boldsymbol{h}) \equiv 0, \quad \forall \boldsymbol{h}. \tag{163}$$

Then, we have

$$\begin{aligned} \lim_{r\to 0_+} \frac{1}{\lambda(\mathcal{B}_r^+)} \int_{\mathcal{B}_r^+} \sum_{j=1}^K \alpha_j \psi_j(\boldsymbol{h}) d\boldsymbol{h} &= 0 \\ \lim_{r\to 0_+} \frac{1}{\lambda(\mathcal{B}_r^-)} \int_{\mathcal{B}_r^+} \sum_{j=1}^K \alpha_j \psi_j(\boldsymbol{h}) d\boldsymbol{h} &= 0 \end{aligned} \tag{164}$$

Then, we have

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

where the last equality comes from (161) and (162).

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 (165), 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 (163) 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 (36), we have

$$\begin{aligned}
&\|\boldsymbol{h}^{(\ell)}(\theta)-\boldsymbol{h}^{(\ell)}(\theta^\star)\|_2\\
=&\|\sigma\big(\theta_{\ell-1}^\top\boldsymbol{h}^{(\ell-1)}(\theta)\big)-\sigma\big(\theta_{\ell-1}^{\star\top}\boldsymbol{h}^{(\ell-1)}(\theta^\star)\big)\|_2\\
=&\|\sigma\big(\theta_{\ell-1}^\top\boldsymbol{h}^{(\ell-1)}(\theta)\big)-\sigma\big(\theta_{\ell-1}^{\star\top}\boldsymbol{h}^{(\ell-1)}(\theta)\big)+\sigma\big(\theta_{\ell-1}^{\star\top}\boldsymbol{h}^{(\ell-1)}(\theta)\big)-\sigma\big(\theta_{\ell-1}^{\star\top}\boldsymbol{h}^{(\ell-1)}(\theta^\star)\big)\|_2\\
\le&\|\sigma\big(\theta_{\ell-1}^\top\boldsymbol{h}^{(\ell-1)}(\theta)\big)-\sigma\big(\theta_{\ell-1}^{\star\top}\boldsymbol{h}^{(\ell-1)}(\theta)\big)\|_2+\|\sigma\big(\theta_{\ell-1}^{\star\top}\boldsymbol{h}^{(\ell-1)}(\theta)\big)-\sigma\big(\theta_{\ell-1}^{\star\top}\boldsymbol{h}^{(\ell-1)}(\theta^\star)\big)\|_2\\
\le&\|\theta_{\ell-1}-\theta_{\ell-1}^\star\|_2\cdot\|\boldsymbol{h}^{(\ell-1)}(\theta)\|_2+\|\boldsymbol{h}^{(\ell-1)}(\theta)-\boldsymbol{h}^{(\ell-1)}(\theta^\star)\|_2.
\end{aligned}\tag{166}$$

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

$$\|\theta_i\|_2\le\|\theta_i^\star\|_2+\|\theta_i-\theta_i^\star\|_2\lesssim 1.\tag{167}$$

Therefore, we have

$$\|\boldsymbol{h}^{(i)}(\theta)\|_2\le\|\theta_i\|_2\cdots\|\theta_1\|_2\cdot\|\boldsymbol{x}\|_2\lesssim\|\boldsymbol{x}\|_2.\tag{168}$$

Then, we have

$$\begin{aligned}
&\|\boldsymbol{h}^{(\ell)}(\theta)-\boldsymbol{h}^{(\ell)}(\theta^\star)\|_2\\
\le&\|\theta_{\ell-1}-\theta_{\ell-1}^\star\|_2\cdot\|\boldsymbol{x}\|_2+\|\boldsymbol{h}^{(\ell-1)}(\theta)-\boldsymbol{h}^{(\ell-1)}(\theta^\star)\|_2\\
\le&\sum_{i=1}^{\ell-1}\|\theta_i-\theta_i^\star\|_2\cdot\|\boldsymbol{x}\|_2+\|\boldsymbol{h}^{(1)}(\theta)-\boldsymbol{h}^{(1)}(\theta^\star)\|_2\\
=&\sum_{i=1}^{\ell-1}\|\theta_i-\theta_i^\star\|_2\cdot\|\boldsymbol{x}\|_2+\|\boldsymbol{x}-\boldsymbol{x}\|_2\\
=&\sum_{i=1}^{\ell-1}\|\theta_i-\theta_i^\star\|_2\cdot\|\boldsymbol{h}^{(i-1)}(\theta)\|_2\\
\le&\|\theta-\theta^\star\|_2\cdot\|\boldsymbol{x}\|_2,
\end{aligned}\tag{169}$$

which completes the proof. □