Provably Efficient Offline Reinforcement Learning With Trajectory-Wise Reward

Tengyu Xu, Yue Wang, Shaofeng Zou, *Member, IEEE*, and Yingbin Liang, *Fellow, IEEE*

***Abstract*— The remarkable success of reinforcement learning (RL) heavily relies on observing the reward of every visited state-action pair. In many real world applications, however, an agent can observe only a score that represents the quality of the whole trajectory, which is referred to as the *trajectory-wise reward*. In such a situation, it is difficult for standard RL methods to well utilize trajectory-wise reward, and large bias and variance errors can be incurred in policy evaluation. In this work, we propose a novel offline RL algorithm, called Pessimistic vAlue iteRaTion with rEward Decomposition (PARTED), which decomposes the trajectory return into per-step proxy rewards via least-squares-based reward redistribution, and then performs pessimistic value iteration based on the learned proxy reward. To ensure the value functions constructed by PARTED are always pessimistic with respect to the optimal ones, we design a new penalty term to offset the uncertainty of the proxy reward. We first show that our PARTED achieves an $\tilde{\mathcal{O}}(dH^3/\sqrt{N})$ suboptimality for linear MDPs, where d is the dimension of the feature, H is the episode length, and N is the size of the offline dataset. We further extend our algorithm and results to general large-scale episodic MDPs with neural network function approximation. To the best of our knowledge, PARTED is the first offline RL algorithm that is provably efficient in general MDP with trajectory-wise reward.**

***Index Terms*— Linear Markov decision processes (MDPs), neural networks, function approximation, reward redistribution, pessimistic principle.**

I. Introduction

REINFORCEMENT learning (RL) aims at searching for an optimal policy in an unknown environment [1]. To achieve this goal, an instantaneous reward is typically required at every step so that RL algorithms can maximize the cumulative reward of a Markov Decision Process (MDP). In recent years, RL has achieved remarkable empirical success with a high quality reward function [2], [3], [4], [5], [6]. However, in many real-world scenarios, instantaneous rewards are hard or impossible to be obtained. For example, in the autonomous driving task [7], it is very costly and time consuming to score every state-action pair that the agent (car) visits. In contrast, it is fairly easy to score the entire trajectory after the agent completing the task [8]. Therefore, in practice, it becomes more reasonable to adopt trajectory-wise reward schemes, in which only a return signal that represents the quality of the entire trajectory is revealed to the agent in the end. In recent years, trajectory-wise rewards have become prevalent in many real-world applications [9], [10], [11], [12], [13].

Although trajectory-wise rewards are convenient to be obtained, it is often challenging for standard RL algorithms to utilize such a type of rewards well due to the high bias and variance it can introduce in the policy evaluation process [14], which leads to unsatisfactory policy optimization results. To address such an issue, [8], [15] proposed to encode the whole trajectory and search for a non-Markovian trajectory-dependent optimal policy using the contextual bandit method. Although this type of approaches have promising theoretical guarantees, they are difficult to be implemented in practice due to the difficulty of searching the large trajectory-dependent policy space whose dimension increases exponentially with the horizon length. Another type of approaches widely adopted in practice is called *reward redistribution*, which learns a reward function by allocating the trajectory-wise reward to every visited state-action pairs based on their contributions [14], [16], [17], [18], [19]. Since the reward function in reward redistribution is typically learned via solving a supervised learning problem, such an approach is sample-efficient and can be integrated into the existing RL frameworks easily. However, most of existing reward redistribution approaches do not have theoretical performance guarantee. So far, only [19] proposes a provably efficient reward redistribution algorithm, but is only applicable to tabular episodic MDP and requires both reward and transition kernel to be horizon-independent.

Despite the superior performance of the reward redistribution method, all previous algorithms considered only the *online* setting, which are not applicable to many critical domains where offline sampling is preferred (or can be required), as interactive data collection could be very costly and risky [7], [20]. *How to design reward redistribution in offline RL for trajectory-wise rewards is an important but fully unexplored problem.* For such a problem, designing reward redistribution algorithms can be hard due to the insufficient sample coverage issue [21] in offline RL. Further challenges can be encountered

Manuscript received 12 April 2023; revised 19 January 2024; accepted 30 June 2024. Date of publication 12 July 2024; date of current version 20 August 2024. The work of Shaofeng Zou was supported by the U.S. National Science Foundation under Grant CCF-2007783, Grant CCF-2106560, and Grant ECCS-2337375 (CAREER). The work of Yingbin Liang was supported in part by the U.S. National Science Foundation under Grant RINGS-2148253 and Grant CNS-2112471. *(Corresponding author: Shaofeng Zou.)*

Tengyu Xu was with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA. He is now with GenAI Meta AI Team, Meta Platforms, Menlo Park, CA 94025 USA (e-mail: xu.3260@buckeyemail.osu.edu).

Yue Wang is with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816 USA (e-mail: yue.wang@ucf.edu).

Shaofeng Zou is with the Department of Electrical Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14228 USA (e-mail: szou3@buffalo.edu).

Yingbin Liang is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: liang.889@osu.edu).

Communicated by A. Krishnamurthy, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2024.3427141

when we try to design *provably* efficient reward distribution algorithms for general MDPs with large state space and horizon-dependent rewards and transition kernels, which has not been studied in online setting.

The goal of this work is to design an offline RL algorithm with reward redistribution for trajectory-wise rewards, which has provable efficiency guarantee for general episodic MDPs.

A. Main Contributions

In this paper, we consider episodic MDP with possibly infinite state space and horizon-dependent reward function and transition kernel. The trajectory-wise reward adopts a standard sum-form as considered previously in [18], [19], [22], [23], [24], and [25], in which only the summation of rewards over the visited state-action pairs is revealed at the end of each episode.

We propose a novel Pessimistic vAlue iteRaTion with rEward Decomposition (PARTED) algorithm for offline RL with trajectory-wise rewards, which incorporates a least-square-based reward redistribution into the pessimistic value iteration (PEVI) algorithm [26], [27], [28], [29], [30]. Differently from the standard PEVI with instantaneous reward, in which reward and value function can be learned together by solving a single regression problem, in PARTED, reward need to be learned separately from the value function by training a regression model to decompose the trajectory return into per-step proxy rewards. In order to capture the reward and value function for a large state space, we adopt over-parameterized neural networks for function approximation. Moreover, to offset the estimation error of proxy rewards, we design a penalty function by transferring the uncertainty from the covariance matrix of trajectory features to step-wise proxy rewards via an "one-block-hot" vector, which is new in the literature.

We show that our proposed new penality term ensures that the value functions constructed by PARTED are always pessimistic with respect to the optimal ones. We then show that PARTED achieves an $\tilde{\mathcal{O}}(dH^3/\sqrt{N})$ suboptimality in the linear MDP setting, where d is the feature dimension, H is the time horizon of MDPs, and N is amount of offline data. We further extend such results to general MDPs with large scales. We show that with over-parameterized neural network function approximation, PARTED achieves an $\tilde{\mathcal{O}}(D_{\text{eff}}H^2/\sqrt{N})$ suboptimality, where D_{eff} is the effective dimension of neural tangent kernel matrix and generalizes the feature dimension in linear MDPs when $D_{\text{eff}} = dH$. To the best of our knowledge, PARTED is the first-known offline RL algorithm that is provably efficient in general episodic MDPs with trajectory-wise rewards.

B. Related Works

1) Trajectory-Wise Reward RL: Policy optimization with trajectory-rewards is extremely difficult. A variety of practical strategies have been proposed to resolve this technical challenge by redistributing trajectory rewards to step-wise rewards. RUDDER [14] trains a return predictor of state-action sequence with LSTM [31], and the reward at each horizon is then assigned by the difference between the predications of two adjacent sub-trajectories. Later, [16] improves RUDDER and utilizes a Transformer [32] for better reward learning. IRCR [17] assigns the proxy reward of a state-action pair as the normalized value of trajectory returns that contain the correspondingly state-action pair. RRD [18] learns a proxy reward function by solving a supervised learning problem together with a Monte-Carlo sampling strategy. Although those methods have achieved great empirical success, they all lack overall theoretical performance guarantee.

Differently from empirical studies, existing theoretical works of trajectory-wise reward RL are rare and focus only on the online setting. One line of research assumes trajectory reward being non-Markovian, and thus focuses on searching for a non-Markovian, trajectory-dependent optimal policy. [8] assumes that trajectory-wise reward is a binary signal generated by a logistic classifier with trajectory embedding as the input. In this setting, the policy optimization problem is reduced to a linear contextual bandit problem in which the trajectory embedding is the contextual vector. Reference [15] considers a similar setting as [8] but assumes only having access to a binary preference score between two trajectories instead of an absolute reward of a trajectory. Another line of research assumes that the trajectory-wise reward is the summation of underlying step-wise Markovian rewards. The goal of this line of work is to search for an optimal Markovian policy. Reference [33] adopted a mirror descent approach so that the summation of rewards alone is sufficient to perform the policy optimization. This approach relies on the on-policy unbiased sampling of trajectory rewards, and can hardly be extended to the offline setting. Reference [19] proposed to recover the reward by solving a least-squared regression problem that fits the summation of reward estimation toward the trajectory reward.

To our best knowledge, offline RL with trajectory-wise rewards (where no interaction with the environment is allowed) has not been studied before, and our work develops the first-known algorithm for such a setting with provable sample efficiency guarantee. Further, although our reward redistribution approach applies the least-square based method, which has also been adopted in [19], our algorithm is designed for general MDPs with possibly infinite state and horizon-dependent rewards and transition kernels, which is very different from that in [19] designed for tabular MDPs with time-independent rewards and transition kernels.

2) Offline RL: The major challenge in offline RL is the insufficient sample coverage in the pre-collected dataset, which arises from the lack of exploration [21], [34]. To address such a challenge, two types of algorithms have been studied: (1) regularized approaches, which prevent the policy from visiting states and actions that are less covered by the dataset [35], [36], [37], [38], [39]; (2) pessimistic approach, which penalize the estimated values of the less-covered state-action pairs [40], [41]. So far, a number of provably efficient pessimistic offline RL algorithm have been proposed in both tabular MDP setting [28], [30], [42], [43], [44], [45], [46], [47], [48] and linear MDP setting [21], [26], [29], [49], [50], [51], [52]. However, the efficiency of all those works relies on both the availability of instantaneous reward and special structures of MDP, which

can hardly be satisfied in practical settings. In this work, we take a first step towards relaxing those two assumptions by proposing PARTED, which is provably efficient in general episodic MDPs with trajectory-wise rewards.

II. Preliminary and Problem Formulation

A. Episodic Markov Decision Process

An episodic Markov decision process (MDP) is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $H > 0$ is the length of each episode, and $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ and $r = \{r_h\}_{h\in[H]}$ are the transition kernel and reward, respectively, where $[n] = \{1, 2, \cdots, n\}$ for integer $n \geq 1$. We assume $\mathcal{S}$ is a measurable space of possibly infinite cardinality and $\mathcal{A}$ is a finite set. For each $h \in [H]$, $\mathbb{P}_h(\cdot|s, a)$ denotes the transition probability when action a is taken at state s at timestep h, and $r_h(s, a) \in [0, 1]$ is a **random** reward that is observed with state-action pair (s, a) at timestep h. We denote the mean of the reward as $R_h(s, a) = \mathbb{E}[r_h(s, a)|s, a]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. For any policy $\pi = \{\pi_h\}_{h\in[H]}$, we define the state value function $V_h^\pi(\cdot) : \mathcal{S} \to \mathbb{R}$ and state-action value function $Q_h^\pi(\cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ at each timestep h as

$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=h}^{H} r_t(s_t, a_t) \middle| s_h = s \right],$$

$$Q_h^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=h}^{H} r_t(s_t, a_t) \middle| (s_h, a_h) = (s, a) \right],$$

where the expectation $\mathbb{E}_\pi$ is taken with respect to the randomness of the trajectory induced by policy π, which is obtained by taking action $a_t \sim \pi_t(\cdot|s_t)$ and transiting to the next state $s_{t+1} \sim \mathbb{P}_t(\cdot|s_t, a_t)$ at timestep $t \in [H]$. At each timestep $h \in [H]$, for any $f : \mathcal{S} \to \mathbb{R}$, we define the transition operator as $(\mathbb{P}_h f)(s, a) = \mathbb{E}[f(s_{h+1})|(s_h, a_h) = (s, a)]$ and the Bellman operator as $(\mathbb{B}_h f)(s, a) = R_h(s, a) + (\mathbb{P}_h f)(s, a)$. For episodic MDP $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$, we have

$$\begin{aligned} Q_h^\pi(s, a) &= (\mathbb{B}_h V_{h+1}^\pi)(s, a), \\ V_h^\pi(s) &= \langle Q_h^\pi(s, \cdot), \pi_h(\cdot|s) \rangle_{\mathcal{A}}, \\ V_{H+1}^\pi(s) &= 0, \end{aligned}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{A}}$ denotes the inner product over $\mathcal{A}$. We define the optimal policy π^* as the policy that yields the optimal value function, i.e., $V_h^{\pi^*}(s) = \sup_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$. For simplicity, we denote $V_h^{\pi^*}$ and $Q_h^{\pi^*}$ as V_h^* and Q_h^*, respectively. The Bellman optimality equation is given as follows

$$Q_h^*(s, a) = (\mathbb{B}_h V_h^*)(s, a), \tag{1}$$

$$V_h^*(s) = \operatorname*{argmax}_{a\in\mathcal{A}} Q_h^*(s, \cdot), \tag{2}$$

$$V_{H+1}^*(s) = 0. \tag{3}$$

The goal of reinforcement learning is to learn the optimal policy π^*. For any fixed π, we define the performance metric as

$$\text{SubOpt}(\pi, s) = V_1^*(s) - V_1^\pi(s),$$

which is the suboptimality of the policy π given the initial state $s_1 = s$.

B. Trajectory-Wise Reward and Offline RL

In the trajectory-wise reward setting, the transition of the environment is still Markovian and the agent can still observe and interact with the environment instantly as in standard MDPs. However, unlike standard MDPs in which the agent can receive an instantaneous reward $r_h(s, a)$ for every visited state-action pair x at each timestep h, in the trajectory-wise reward setting, only a reward that is associated with the whole trajectory can be observed at the end of the episode, i.e., $r(\tau)$ where $\tau = \{(s_1^\tau, a_1^\tau), \cdots, (s_H^\tau, a_H^\tau)\}$ denotes a trajectory and (s_h^τ, a_h^τ) is the h-th state-action pair in trajectory τ, which is called "trajectory reward" in the sequel. In this work, we consider the setting in which the trajectory reward is the summation of the underlying instantaneous reward in the trajectory of MDP $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$, i.e., $r(\tau) = \sum_{h=1}^H r_h(s_h^\tau, a_h^\tau)$. We denote the mean of the trajectory reward as $R(\tau) = \mathbb{E}[r(\tau)|\tau] = \sum_{h=1}^H R_h(s_h^\tau, a_h^\tau)$. Such a sum-form reward has been commonly considered in both theoretical [19] and empirical studies [18], [22], [23], [24], [25]. It models the situations where the agent's goal is captured by a certain metric with additive properties, e.g., the energy cost of a car during driving, the click rate of advertisements during a time interval, or the distance of a robot's running. Such a form of reward can be more general than the standard RL feedback and is expected to be more common in practical scenarios. Note that RL problems under trajectory-wise rewards is very challenging, as traditional policy optimization approach typically fails due the obscured feedback received from the environment, which causes large value function evaluation error [22].

We consider the offline RL setting, in which a learner has access only to a pre-collected dataset $\mathcal{D}$ consisting of N trajectories $\{\tau_i, r(\tau_i)\}_{i,h=1}^{N,H}$ rolled out from some possibly unknown behavior policy μ, where τ_i and $r(\tau_i)$ are the i-th trajectory and the observed trajectory reward of τ_i, respectively. Given this batch data $\mathcal{D}$ with only trajectory-wise rewards and a target accuracy ϵ, our goal is to find a policy π such that $\text{SubOpt}(\pi, s) \leq \epsilon$ for all $s \in \mathcal{S}$.

C. Linear MDPs

We mainly consider linear MDPs [53], [54] in this paper. An episodic MDP $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$ is a linear MDP with a known feature map $\phi(\cdot) : \mathcal{X} \to \mathbb{R}^d$ if for each $h \leq H$, there exist unknown vectors $w_h^*(s) \in \mathbb{R}^d$, $\forall s \in \mathcal{S}$ and an unknown vector $\theta_h^* \in \mathbb{R}^d$ such that

$$\mathbb{P}_h(s'|s, a) = \langle \phi(s, a), w_h^*(s') \rangle, \tag{4}$$

$$R_h(s, a) = \langle \phi(s, a), \theta_h^* \rangle, \tag{5}$$

for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Here we assume $\|\phi(x)\|_2 \leq 1$ for all $x \in \mathcal{X}$ and $\max\{\|w_h^*(\mathcal{S})\|_2, \|\theta_h^*\|_2\} \leq \sqrt{d}$ at each step $h \in [H]$, where $\|w_h^*(\mathcal{S})\|_2 = \int_{\mathcal{S}} \|w_h^*(s)\|_2 \, ds$.

In linear MDPs, it has been shown that both reward $R_h(\cdot)$ and transition value function $(\mathbb{P}_h \widehat{V}_{h+1})(\cdot)$ are linear functions with respect to $\phi(\cdot)$ [53], [55].

Notations: We use $\widetilde{\mathcal{O}}(X)$ to refer to a quality that is upper bounded by X, up to poly-log factors of d, H, N, m and $(1/\delta)$. Furthermore, we use $\mathcal{O}(X)$ to refer to a quantity that is upper

bounded by X up to constant multiplicative factors. We use I_d as the identity matrix in dimension d. Similarly, we denote by $\mathbf{0}_d \in \mathbb{R}^d$ as the vector whose components are zeros. For any square matrix M, we let $\|M\|_2$ denote the operator norm of M. Finally, for any positive definite matrix $M \in \mathbb{R}^{d\times d}$ and any vector $x \in \mathbb{R}^d$, we define $\|x\|_M = \sqrt{x^\top M x}$.

III. Algorithm

In this section, we propose a Pessimistic vAlue iteRaTion with rEward Decomposition (PARTED) algorithm for linear MDPs. PARTED shares a similar structure as that of pessimistic value iteration (PEVI) [26], [28], [47], but has a very different design due to trajectory-wise rewards. In PEVI, a pessimistic estimator of the value function is constructed from the dataset $\mathcal{D}$ and the Bellman optimality equation is then iterated based such an estimator. Since instantaneous rewards are available in PEVI, given a function class $\mathcal{G}$, PEVI constructs an estimated Bellman backup of value function $(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})$ by solving the following regression problem for all $h \in [H]$ in the backward direction:

$$(\widehat{\mathbb{B}}_h \widehat{V}_{h+1}) \tag{6}$$
$$= \operatorname*{argmin}_{g_h \in \mathcal{G}} L^h_{\text{PEVI}}(g_h) \tag{7}$$
$$= \sum_{\tau\in\mathcal{D}} \left(r_h(x_h^\tau) + \widehat{V}_{h+1}(s_{h+1}^\tau) - g_h(x_h^\tau)\right)^2 \tag{8}$$
$$+ \lambda \cdot \text{Reg}(g_h). \tag{9}$$

In eq. (9), $\widehat{V}_{h+1}(\cdot)$ is the pessimistic estimator of optimal value function constructed for horizon $h+1$, $\lambda > 0$ is a regularization parameter and $\text{Reg}(\cdot)$ is the regularization function. The optimal state-action value function can then be estimated as $\widehat{Q}_h(\cdot) = \min\{(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\cdot) - \Gamma_h(\cdot), H\}^+$, where $-\Gamma_h$ is a negative penalty used to offset the uncertainty in $(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(\cdot)$ and guarantee the pessimism of $\widehat{Q}_h$.

However, in the trajectory-wise reward setting, the absence of instantaneous reward functions $r_h(\cdot)$ renders previous algorithms, which relied on eq. (9), inapplicable. To overcome this issue, we propose to additionally construct estimations of the instantaneous reward r_h using the dataset. Leveraging these estimations and eq. (9), we can then derive estimations for the Bellman backups of value functions $(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})$ that form an integral part of our PARTED algorithm. Although straightforward, the constraints of a limited dataset and the stochastic nature of the MDPs further introduces uncertainty to the reward estimation. In response, we incorporate the concept of PEVI, constructing a penalty term to refine our estimations. We address the uncertainty arising from unknown instantaneous rewards by showing that the designed penalty term provides a pessimism guarantee for our estimations.

A. Reward Redistribution

We first construct our estimations of the instantaneous rewards from the trajectory-wise reward. Specifically, we estimate each $R_h(\cdot)$ for all $h \in [H]$ using a linear function $\langle \phi(s,a), \theta_h\rangle$, where $\theta_h \in \mathbb{R}^d$ is the estimation parameter. We aim to obtain the estimations $\Theta = [\theta_1^\top, \cdots, \theta_H^\top] \in \mathbb{R}^{dH}$ by minimizing the following loss function $L_r : \mathbb{R}^{dH} \to \mathbb{R}$:

$$L_r(\Theta) = \sum_{\tau\in\mathcal{D}} \left[\sum_{h=1}^H \langle \phi(x_h^\tau), \theta_h\rangle - r(\tau)\right]^2 \tag{10}$$
$$+ \lambda_1 \cdot \sum_{h=1}^H \|\theta_h - \theta_0\|_2^2, \tag{11}$$

where $\lambda_1 > 0$ is a regularization parameter. We then set the per-step proxy reward $\widehat{R}_h(\cdot)$ as

$$\widehat{R}_h(\cdot) = \langle \phi(\cdot), \widehat{\theta}_h\rangle, \tag{12}$$

where $\widehat{\Theta} = [\widehat{\theta}_1^\top, \cdots, \widehat{\theta}_H^\top] = \operatorname*{argmin}_{\Theta\in\mathbb{R}^{dH}} L_r(\Theta)$.

B. Transition Value Function Estimation

Similarly, we also use linear function $\langle \phi(s,a), w_h\rangle$ to estimate transition value functions $\{(\mathbb{P}_h \widehat{V}_{h+1})(\cdot,\cdot)\}_{h\in[H]}$ for all $h \in [H]$, where $w_h \in \mathbb{R}^d$ is a learnable parameter. For each $h \in [H]$, we define the loss function $L_v^h(w_h)$: $\mathbb{R}^d \to \mathbb{R}$ as

$$L_v^h(w_h) = \sum_{\tau\in\mathcal{D}} \left(\widehat{V}_{h+1}(s_{h+1}^\tau) - \langle \phi(x_h^\tau), w_h\rangle\right)^2$$
$$+ \lambda_2 \cdot \|w_h - w_0\|_2^2, \tag{13}$$

where $\lambda_2 > 0$ is a regularization parameter. We then define $(\widehat{\mathbb{P}}_h \widehat{V}_{h+1})(\cdot) : \mathcal{X} \to \mathbb{R}$ as

$$(\widehat{\mathbb{P}}_h \widehat{V}_{h+1})(\cdot) = \langle \phi(\cdot), \widehat{w}_h\rangle, \tag{14}$$

where

$$\widehat{w}_h = \operatorname*{argmin}_{w_h\in\mathbb{R}^d} L_v^h(w_h). \tag{15}$$

C. Penalty Term Construction

As discussed above, the estimations we obtained are uncertain due to limited dataset and randomness from the MDPs. We then construct the penalty term Γ_h to offset the uncertainties in them.

We first consider the penalty term that is used to offset the uncertainty raised from estimating the reward. For any trajectory $\tau \in \mathcal{D}$, we define the trajectory feature $\Phi(\tau) = [\phi(x_1^\tau), \cdots, \phi(x_H^\tau)] \in \mathbb{R}^{dH}$. Based on it, we define the trajectory feature covariance matrix $\Sigma(\Theta) \in \mathbb{R}^{dH\times dH}$ as

$$\Sigma = \lambda_1 \cdot I_{dH} + \sum_{\tau\in\mathcal{D}} \Phi(\tau)\Phi(\tau)^\top.$$

We further define an "one-block-hot" vector $\Phi_h(x) = [\mathbf{0}_d^\top, \cdots, \phi(x)^\top, \cdots, \mathbf{0}_d^\top]^\top \in \mathbb{R}^{dH}$ for all $x \in \mathcal{X}$, i.e., $[\Phi_h(x)]_{d(h-1)+1:dh} = \phi(x)$ and the rest entries are all zero.

The penalty term $b_{r,h}$ of the estimated reward is then defined as

$$b_{r,h}(x) = \left[\Phi_h(x)^\top \Sigma^{-1} \Phi_h(x)\right]^{1/2}. \tag{16}$$

Note that the reward penalty term is new and first proposed in this work. By constructing $b_{r,h}(x)$ in this way, we can capture the effect of uncertainty caused by solving the trajectory-wise regression problem in eq. (12), which is contained in the covariance matrix Σ, on the proxy reward $\hat{R}_h$ at each step $h \in [H]$, via the "one-block-hot" vector $\Phi_h(x)$.

Algorithm 1 Linear Pessimistic Value Iteration With Reward Decomposition (PARTED)

Input: Dataset $\mathcal{D} = \{\tau_i, r(\tau_i)\}_{i,h=1}^{N,H}$
Initialization: Set $\widehat{V}_{H+1}$ as zero function
Obtain $\widehat{R}_h$ and $\widehat{\Theta}$ according to eq. (12)
for $h = H, H-1, \cdot, 1$ **do**
  Obtain $\widehat{\mathbb{P}}_h\widehat{V}_{h+1}$ and $\widehat{w}_h$ according to eq. (14)
  Obtain $\Gamma_h(\cdot)$ according to eq. (18)
  $\widehat{Q}_h(\cdot) = \min\{\widehat{R}_h(\cdot) + \widehat{\mathbb{P}}_h\widehat{V}_{h+1}(\cdot) - \Gamma_h(\cdot), H-h+1\}^+$
  $\widehat{\pi}_h(\cdot|s) = \operatorname{argmax}_{\pi_h}\langle\widehat{Q}_h(s,\cdot), \pi_h(\cdot|s)\rangle$
  $\widehat{V}_h(\cdot) = \langle\widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot|\cdot)\rangle_{\mathcal{A}}$
end for

Next, we consider the penalty term that is used to offset the uncertainty raised from estimating the transition value function $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ for each $h \in [H]$. We define a matrix $\Lambda_h \in \mathbb{R}^{d\times d}$ as

$$\Lambda_h = \lambda_2 \cdot I_d + \sum_{\tau\in\mathcal{D}} \phi(x_h^\tau)\phi(x_h^\tau)^\top.$$

The penalty term $b_{v,h}$ of the estimated transition value function is then defined as

$$b_{v,h}(x) = \left[\phi(x)^\top \Lambda_h^{-1}\phi(x)^\top\right]^{1/2}. \tag{17}$$

Finally, the penalty term for the estimated Bellman operation $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}(\cdot)$ is obtained as Finally, the penalty term for the estimated Bellman operation $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}(\cdot)$ is obtained as

$$\Gamma_h(x) = \beta_1 b_{r,h}(x) + \beta_2 b_{v,h}(x), \tag{18}$$

where $\beta_1, \beta_2 > 0$ are constant factors and will be determined in algorithm design. The estimator of $Q_h(\cdot)$ and $V_h(\cdot)$ can then be obtained as

$$\widehat{Q}_h(\cdot) = \min\{\widehat{R}_h(\cdot) + (\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot) - \Gamma_h(\cdot), H-h+1\}^+, \tag{19}$$

$$\widehat{V}_h(\cdot) = \operatorname*{argmax}_{a\in\mathcal{A}} \widehat{Q}_h(\cdot, a). \tag{20}$$

We present our PARTED for linear MDPs in Algorithm 1 as follows.

While our PARTED algorithm shares similarities in form and concept with previous PEVI methods, both incorporating a penalty term, it is crucial to emphasize our major innovation lies in the construction of a penalty term specifically designed to offer pessimism guarantees for the estimation of instantaneous rewards. This issue is introduced by the trajectory-wise reward setting and cannot be effectively tackled by preceding methods.

We present our algorithm tailored for linear MDPs here for the sake of convenience and simplicity. As we elaborate later, we can further extend and generalize our methods to encompass general MDPs in conjunction with neural network function approximation.

IV. Main Results

A. Suboptimality of PARTED Under Linear MDPs

In this section, we illustrate our results for linear MDPs with trajectory-wise rewards. We adopt a standard assumption that is widely assumed in offline RL works, e.g., [26], [56], [57], [58], [59], [60], [61], [62], [63], [64], and [65]. It assumes that the dataset is collected by a well-explored behavior policy μ, which is also known as global data coverage assumption.

Assumption 1 (Well-Explored Dataset): The N trajectories in dataset $\mathcal{D}$ are independent and identically induced by a fixed behaviour policy μ. There exist absolute constants $C_\sigma > 0$ and $C_\varsigma > 0$ such that

$$\lambda_{\min}(\overline{M}(\Theta_0)) \geq C_\sigma$$
$$\text{and} \quad \lambda_{\min}(\overline{m}_h(w_0)) \geq C_\varsigma \quad \forall h \in [H],$$

where

$$\overline{M} = \mathbb{E}_\mu\left[\Phi(\tau)\Phi(\tau)^\top\right]$$
$$\text{and} \quad \overline{m}_h(w_0) = \mathbb{E}_\mu\left[\phi(x_h^\tau)\phi(x_h^\tau)^\top\right].$$

We recognize the global data coverage in the offline setting, acknowledging the possibility of relaxing it to partial coverage through the design of an alternative standard pessimistic penalty term, e.g., [26], [30], [34], [40], [44], [47], [48], [66], and [67]. However, it is important to underscore that the primary contribution of our work lies in addressing uncertainty arising from the *trajectory-wise reward*. This aspect stands in contrast to existing works that primarily focus on managing uncertainty inherent in the *dataset* itself and is orthogonal to them. While integrating their methods with ours is feasible for a broader application to a more general offline setting, opting for such integration may unnecessarily complicate the presentation. Therefore, we choose to present our results under the assumption of global coverage, with the aim of emphasizing the distinctive contribution of our approach in handling uncertainties associated with trajectory-wise rewards.

The following theorem characterizes the suboptimality of Algorithm 1.

Theorem 1: Consider Algorithm 1. Let $\lambda_1 = \lambda_2 = 1$ and $\beta_1 = \mathcal{O}(H\sqrt{dH\log(N/\delta)})$ and $\beta_2 = \mathcal{O}(dH^2\sqrt{\log(dH^3N^{5/2}/\delta)})$. Then, with probability at least $1-\delta$, we have

$$\text{SubOpt}(\widehat{\pi}, s) \leq \mathcal{O}\left(\frac{dH^3}{\sqrt{NC_\sigma C_\varsigma}}\sqrt{\log\left(\frac{dH^3N^{5/2}}{\delta}\right)}\right).$$

To highlight why trajectory-wise reward RL is more challenging than instantaneous reward RL, we observe that Theorem 1 with *trajectory-wise rewards* has an additional dependence on the horizon H, compared to the suboptimality $\widetilde{\mathcal{O}}(dH^2/\sqrt{N})$ [26, Corollary 4.5] of PEVI for linear MDP with *instantaneous rewards*. In contrast to the prior results obtained under the instantaneous reward setting, we have introduced an additional penalty term to mitigate uncertainty during the reward redistribution process. Specifically, PARTED is required to solve a trajectory-level regression problem with features $\Phi(\tau) \in \mathbb{R}^{dH}$, leading to increased uncertainty in

the regression solution utilized for constructing the per-step proxy reward. This supplementary penalty term introduces a higher level of pessimism compared to previous PEVI, resulting in an additional H dependence in the suboptimality. When assuming full knowledge of the instantaneous reward, we can set $\beta_1 = 0$ in the penalty term, and the suboptimality aligns with that of PEVI. Thus, we contend that this additional dependence is introduced by the trajectory-wise reward setting and is unavoidable.

1) Discussion of Proof of Theorem 1: Comparing to the analysis of PEVI for linear MDP with instantaneous reward, which has been extensively studied in offline RL [26], [27], [29], our analysis needs to address the following challenge. In instantaneous reward setting, both $R_h(\cdot)$ and $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ can be learned together by solving a single regression problem in per-step scale. However, in our trajectory-wise reward setting, $R_h(\cdot)$ and $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ need to be learned separately by solving two regression problems (eqs. (12) and (14)) in different scales, i.e., eq. (12) is in trajectory scale and eq. (14) is in per-step scale. In order to apply union concentrations to bound the Bellman estimation error $|(\mathbb{B}_h\widehat{V}_h)(\cdot) - (\widehat{\mathbb{B}}_h\widehat{V}_h)(\cdot)|$, we need to develop new techniques to handle the mismatch between eqs. (12) and (14) in terms of scale.

B. Suboptimality of PARTED for General MDPs

In this section, we extend our PARTED to general MDPs with large state space and present the main results. We provide a concise overview of our findings, with a more comprehensive and detailed discussion available in Section A and Section B. Our overarching objective is to validate two pivotal claims: (1) Our framework adeptly manages uncertainty arising from the trajectory-wise reward setting, even in the most general settings lacking latent structures; and (2) The suboptimality exhibited by PARTED in general MDPs corresponds with that observed in linear MDPs, signifying an extension of the preceding algorithm design without the introduction of additional sample complexity.

When the problem scale is large, especially for the general MDPs, we consider the function approximation setting, in which the state-action value function is approximated by a two-layer neural network. We denote $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ and view it as a subset of $\mathbb{R}^d$. We further assign a feature vector $x \in \mathcal{X}$ to represent a state-action pair (s, a). Without loss of generality, we assume that $\|x\|_2 = 1$ for all $x \in \mathcal{X}$. We also allow $x = 0$ to represent a null state-action pair. We now define a two-layer neural network $f(\cdot, b, w) : \mathcal{X} \to \mathbb{R}$ with $2m$ neurons and weights (b, w) as

$$f(x; b, w) = \frac{1}{\sqrt{2m}} \sum_{r=1}^{2m} b_j \cdot \sigma(w_r^\top x), \tag{21}$$

where $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$ is the activation function, $b_r \in \mathbb{R}$ and $w_r \in \mathbb{R}^d$ for all $r \in [2m]$, and $b = (b_1, \cdots, b_{2m})^\top \in \mathbb{R}^{2m}$ and $w = (w_1^\top, \cdots, w_{2m}^\top)^\top \in \mathbb{R}^{2md}$.

With the neural network approximation we introduced, we extend our PARTED to the general MDP setting. Although the algorithm design follows a similar framework as the linear MDP setting, unlike linear MDPs where both $R_h(\cdot)$ and $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ can be captured exactly by linear functions, for general MDPs with neural network approximation, we need to design new estimation and penalty terms to tackle the difficulties introduced by the approximation.

1) Reward Redistribution: We use a neural network $f(\cdot, \theta_h)$ given in eq. (21) to represent per-step mean reward $R_h(\cdot)$ for all $h \in [H]$, where the parameter $\theta_h \in \mathbb{R}^{2md}$ is obtained by minimizing the loss function

$$L_r(\Theta) = \sum_{\tau \in \mathcal{D}} \left[\sum_{h=1}^{H} f(x_h^\tau, \theta_h) - r(\tau) \right]^2 + \lambda_1 \cdot \sum_{h=1}^{H} \|\theta_h - \theta_0\|_2^2 . \tag{22}$$

Then, the per-step proxy reward $\widehat{R}_h(\cdot)$ is obtained as

$$\text{where} \quad \begin{aligned} &\widehat{R}_h(\cdot) = f(\cdot, \widehat{\theta}_h), \\ &\widehat{\Theta} = \operatorname*{argmin}_{\Theta \in \mathbb{R}^{2mdH}} L_r(\Theta) \\ &\text{and } \widehat{\Theta} = [\widehat{\theta}_1^\top, \cdots, \widehat{\theta}_H^\top]^\top. \end{aligned} \tag{23}$$

2) Transition Value Function Estimation: Similarly, we use H neural networks given in eq. (21) with parameter $\{w_h\}_{h\in[H]}$ to estimate $\{(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)\}_{h\in[H]}$, where $w_h \in \mathbb{R}^{2md}$ is the parameter of the h-th network. Specifically, for each $h \in [H]$, we define the loss function $L_v^h(w_h)$: $\mathbb{R}^{2md} \to \mathbb{R}$ as

$$L_v^h(w_h) = \sum_{\tau \in \mathcal{D}} \left(\widehat{V}_{h+1}(s_{h+1}^\tau) - f(x_h^\tau, w_h) \right)^2 + \lambda_2 \cdot \|w_h - w_0\|_2^2 , \tag{24}$$

where $\lambda_2 > 0$ is a regularization parameter and w_0 is the initialization shared by all neural networks. The estimated transition value function $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot) : \mathcal{X} \to \mathbb{R}$ can be obtained by solving the following optimization problem

$$(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot) = f(\cdot, \widehat{w}_h), \tag{25}$$

$$\text{where} \quad \widehat{w}_h = \operatorname*{argmin}_{w_h \in \mathbb{R}^{2md}} L_v^h(w_h). \tag{26}$$

3) Penality Term Construction: For any $\tau \in \mathcal{D}$ and $\Theta \in \mathbb{R}^{2mdH}$, we define a trajectory feature $\Phi(\tau, \Theta) = [\phi(x_1^\tau, \theta_1)^\top, \cdots, \phi(x_H^\tau, \theta_H)^\top]^\top$. Based on $\Phi(\tau, \Theta)$, the trajectory feature covariance matrix $\Sigma(\Theta) \in \mathbb{R}^{2mdH \times 2mdH}$ is then defined as

$$\Sigma(\Theta) = \lambda_1 \cdot I_{2mdH} + \textstyle\sum_{\tau \in \mathcal{D}} \Phi(\tau, \Theta)\Phi(\tau, \Theta)^\top.$$

We also define an "one-block-hot" vector $\Phi_h(x, \Theta) = [\mathbf{0}_{2md}^\top, \cdots, \phi(x, \theta_h)^\top, \cdots, \mathbf{0}_{2md}^\top]^\top$ for all $x \in \mathcal{X}$, where $\Phi_h(x, \Theta) \in \mathbb{R}^{2mdH}$ is a vector in which $[\Phi_h(x, \Theta)]_{2md(h-1)+1:2mdh} = \phi(x, \theta_h)$ and the rest entries are zero. The penalty term of reward for a given $\Theta \in \mathbb{R}^{2mdH}$ is defined as:

$$b_{r,h}(x, \Theta) = \left[\Phi_h(x, \Theta)^\top \Sigma^{-1}(\Theta) \Phi_h(x, \Theta)\right]^{1/2}. \tag{27}$$

Next, we consider the penalty of $(\widehat{\mathbb{P}}_h \widehat{V}_{h+1})(\cdot)$ for each $h \in [H]$. We define the per-step feature covariance matrix $\Lambda_h(w_h) \in \mathbb{R}^{2md \times 2md}$ as

$$\Lambda_h(w) = \lambda_2 \cdot I_{2md} + \sum_{\tau \in \mathcal{D}} \phi(x_h^\tau, w)\phi(x_h^\tau, w)^\top.$$

Then, the penality term of $(\widehat{\mathbb{P}}_h \widehat{V}_{h+1})(\cdot)$ for a given $w \in \mathbb{R}^{2md}$ is defined as

$$b_{v,h}(x, w) = \left[\phi(x, w)^\top \Lambda_h(w)^{-1} \phi(x, w)^\top\right]^{1/2}. \tag{28}$$

Finally, combining eqs. (27) and (28), the penalty term for $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}(\cdot)$ is constructed as

$$\Gamma_h(x, \Theta, w) = \beta_1 b_{r,h}(x, \Theta) + \beta_2 b_{v,h}(x, w), \tag{29}$$

where $\beta_1, \beta_2 > 0$ are parameters.

Based on the constructions of the estimation and penalty terms and our previous design of PARTED for linear MDPs, we design the PARTED algorithm for the general MDPs with neural network approximation in Algorithm 2. The algorithm follows the same framework as the linear PARTED, but with different estimation and penalty terms.

We then informally present the suboptimality of the policy $\widehat{\pi}$ obtained via our PARTED for general MDPs. The formal and accurate result can be found in Section B.

Theorem 2 (Informal): Assume that the function class defined by the neural network is reward realizable and Bellman complete, and further assume that the dataset is collected by a well-explored policy. Then with probability at least $1 - (N^2H^4)^{-1}$, the suboptimality of PARTED for general MDP is

$$\text{SubOpt}(\widehat{\pi}, s) \leq \widetilde{\mathcal{O}}\left(\frac{H \max\{\beta_1, \beta_2\}}{\sqrt{N C_\sigma C_\varsigma}}\right).$$

Theorem 2 shows that Algorithm 2 can find an ϵ-optimal policy with $\widetilde{\mathcal{O}}(H^2 \max\{\beta_1, \beta_2\}^2/\epsilon^2)$ episodes of offline data in the trajectory-wise reward setting up to some function approximation error, which vanishes as the neural network width increases.

Unlike linear MDPs where both $R_h(\cdot)$ and $(\mathbb{P}_h \widehat{V}_{h+1})(\cdot)$ can be captured exactly by linear functions, for general MDPs with neural network approximation, we need to develop new analysis to bound the estimation error that caused by the insufficient expressive power of neural networks in order to characterize the optimality of $\widehat{\theta}_h$ and $\widehat{w}_h$, respectively.

Note that linear function with feature $\Phi(\tau)$ and $\phi(x)$ can be viewed as a special case of our general MDPs, both of which belonging to a reproducing kernel Hilbert space (RKHS). We hence further extend our results to the RKHS setting under the finite spectrum NTK assumption as follows.

Corollary 1 (Informal): With probability at least $1 - (N^2H^4)^{-1}$, we have

$$\text{SubOpt}(\widehat{\pi}, s) = \widetilde{\mathcal{O}}\left(D_{\text{eff}} H^2/\sqrt{N C_\sigma C_\varsigma}\right),$$

where D_{eff} denotes the effective dimension of the reproducing kernel.

For the linear MDPs, the effective dimension satisfies $D_{\text{eff}} = dH$, which implies that the suboptimality in Corollary 1 (and Theorem 2) for general MDPs recover the suboptimality of linear MDPs in Theorem 1, which implies our results are self-consistent and we extend our results to the general MDP setting. A more detailed discussion can be found in Section B.

V. Conclusion

In this paper, we propose a novel offline RL algorithm, called PARTED, to handle the episodic RL problem with trajectory-wise rewards. PARTED uses a least-square-based reward redistribution method for reward estimation and incorporates a new penalty term to offset the uncertainty of proxy reward. We showed that for linear MDPs, PARTED achieves an $\widetilde{\mathcal{O}}(dH^3/\sqrt{N})$ suboptimality. We further extended our framework and method to general MDPs with neural network approximation, where we showed PARTED achieves an $\tilde{\mathcal{O}}(D_{\text{eff}} H^2/\sqrt{N})$ suboptimality, which matches the result of linear MDP when the effective dimension satisfies $D_{\text{eff}} = dH$. To the best of our knowledge, this is the first offline RL algorithm that is provably efficient in general episodic MDP setting with trajectory-wise rewards. As a future direction, it is interesting to incorporate the randomized return decomposition in [18] to improve the scalability of PARTED in the long horizon scenario.

Appendix A
Design of PARTED for General MDPs

In this section, we first provide our design of PARTED for general MDPs in details.

A. Reward Redistribution

In order to estimate the instantaneous rewards from the trajectory-wise reward, we use a neural network $f(\cdot, \theta_h)$ given in eq. (21) to represent per-step mean reward $R_h(\cdot)$ for all $h \in [H]$, where $\theta_h \in \mathbb{R}^{2md}$ is the parameter. We further assume, for simplicity, that all the neural networks share the same initial weights denoted by $\theta_0 \in \mathbb{R}^{2md}$. We define the following loss function $L_r(\cdot) : \mathbb{R}^{2mdH} \to \mathbb{R}$ for reward redistribution as

$$L_r(\Theta) = \sum_{\tau \in \mathcal{D}} \left[\sum_{h=1}^H f(x_h^\tau, \theta_h) - r(\tau)\right]^2 + \lambda_1 \cdot \sum_{h=1}^H \|\theta_h - \theta_0\|_2^2, \tag{30}$$

where $\Theta = [\theta_1^\top, \cdots, \theta_H^\top]^\top \in \mathbb{R}^{2mdH}$ and $\lambda_1 > 0$ is a regularization parameter. Then, the per-step proxy reward $\widehat{R}_h(\cdot)$ is obtained by solving the following optimization problem

$$\widehat{R}_h(\cdot) = f(\cdot, \widehat{\theta}_h), \quad \text{where} \quad \widehat{\Theta} = \underset{\Theta \in \mathbb{R}^{2mdH}}{\text{argmin}} \; L_r(\Theta) \quad \text{and} \quad \widehat{\Theta} = [\widehat{\theta}_1^\top, \cdots, \widehat{\theta}_H^\top]^\top. \tag{31}$$

B. Transition Value Function Estimation

Similarly, we use H neural networks given in eq. (21) with parameter $\{w_h\}_{h\in[H]}$ to estimate $\{(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)\}_{h\in[H]}$, where $w_h \in \mathbb{R}^{2md}$ is the parameter of the h-th network. Specifically, for each $h \in [H]$, we define the loss function $L_v^h(w_h)$: $\mathbb{R}^{2md} \to \mathbb{R}$ as

$$L_v^h(w_h) = \sum_{\tau\in\mathcal{D}} \left(\widehat{V}_{h+1}(s_{h+1}^\tau) - f(x_h^\tau, w_h)\right)^2 + \lambda_2 \cdot \|w_h - w_0\|_2^2, \tag{32}$$

where $\lambda_2 > 0$ is a regularization parameter and w_0 is the initialization shared by all neural networks. The estimated transition value function $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot) : \mathcal{X} \to \mathbb{R}$ can be obtained by solving the following optimization problem

$$(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot) = f(\cdot, \widehat{w}_h), \tag{33}$$

where

$$\widehat{w}_h = \operatorname*{argmin}_{w_h\in\mathbb{R}^{2md}} L_v^h(w_h). \tag{34}$$

C. Penalty Term Construction

It remains to construct the penalty term Γ_h to offset the uncertainties in $\widehat{R}_h$ and $(\widehat{\mathbb{P}}_h V_{h+1})$. First consider the penalty of $\widehat{R}_h(\cdot)$ for each $h \in [H]$. For any $\tau \in \mathcal{D}$ and $\Theta \in \mathbb{R}^{2mdH}$, we define a trajectory feature $\Phi(\tau, \Theta) = [\phi(x_1^\tau, \theta_1)^\top, \cdots, \phi(x_H^\tau, \theta_H)^\top]^\top$. Based on $\Phi(\tau, \Theta)$, the trajectory feature covariance matrix $\Sigma(\Theta) \in \mathbb{R}^{2mdH\times 2mdH}$ is then defined as

$$\Sigma(\Theta) = \lambda_1 \cdot I_{2mdH} + \sum_{\tau\in\mathcal{D}} \Phi(\tau, \Theta)\Phi(\tau, \Theta)^\top.$$

We also define an "one-block-hot" vector $\Phi_h(x, \Theta) = [\mathbf{0}_{2md}^\top, \cdots, \phi(x, \theta_h)^\top, \cdots, \mathbf{0}_{2md}^\top]^\top$ for all $x \in \mathcal{X}$, where $\Phi_h(x, \Theta) \in \mathbb{R}^{2mdH}$ is a vector in which $[\Phi_h(x, \Theta)]_{2md(h-1)+1:2mdh} = \phi(x, \theta_h)$ and the rest entries are zero. The penalty term of reward for a given $\Theta \in \mathbb{R}^{2mdH}$ is defined as:

$$b_{r,h}(x, \Theta) = \left[\Phi_h(x, \Theta)^\top \Sigma^{-1}(\Theta)\Phi_h(x, \Theta)\right]^{1/2}.$$

Note that the reward penalty term $b_{r,h}(x, \Theta)$ is new and first proposed in this work. By constructing $b_{r,h}(x, \Theta)$ in this way, we can capture the effect of uncertainty caused by solving the trajectory-wise regression problem in eq. (30), which is contained in the covariance matrix $\Sigma(\Theta)$, on the proxy reward $f(\cdot, \widehat{\theta}_h)$ at each step $h \in [H]$, via the "one-block-hot" vector $\Phi_h(\cdot, \Theta)$.

Next, we consider the penalty of $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ for each $h \in [H]$. We define the per-step feature covariance matrix $\Lambda_h(w_h) \in \mathbb{R}^{2md\times 2md}$ as

$$\Lambda_h(w) = \lambda_2 \cdot I_{2md} + \sum_{\tau\in\mathcal{D}} \phi(x_h^\tau, w)\phi(x_h^\tau, w)^\top.$$

Then, the penalty term of $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ for a given $w \in \mathbb{R}^{2md}$ is defined as

$$b_{v,h}(x, w) = \left[\phi(x, w)^\top \Lambda_h(w)^{-1}\phi(x, w)^\top\right]^{1/2}.$$

Finally, combining eq. (27) and (28), the penalty term for $\widehat{\mathbb{B}}_h\widehat{V}_{h+1}(\cdot)$ is constructed as

$$\Gamma_h(x, \Theta, w) = \beta_1 b_{r,h}(x, \Theta) + \beta_2 b_{v,h}(x, w), \tag{35}$$

where $\beta_1, \beta_2 > 0$ are parameters. The estimator of $Q_h(\cdot)$ and $V_h(\cdot)$ can then be obtained as

$$\begin{aligned}\widehat{Q}_h(\cdot) &= \min\{H, \\ &\quad \widehat{R}_h(\cdot) + (\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot) - \Gamma_h(\cdot, \widehat{\Theta}, \widehat{w}_h)\}^+, \\ \widehat{V}_h(\cdot) &= \operatorname*{argmax}_{a\in\mathcal{A}} \widehat{Q}_h(\cdot, \cdot).\end{aligned} \tag{36}$$

Furthermore, for any $h \in [H]$, we denote $\mathcal{V}_h(x, R_{\beta_1}, R_{\beta_2}, \lambda_1, \lambda_2)$ as the class of functions that takes the form $\overline{V}_h(\cdot) = \max_{a\in\mathcal{A}} \overline{Q}_h(\cdot, a)$, where

$$\begin{aligned}&\overline{Q}_h(x) \\ &= \min\{\langle\phi(x, \theta_0), \theta - \theta_0\rangle + \langle\phi(x, w_0), w - w_0\rangle \\ &\quad - \beta_1 \cdot \sqrt{\Phi_h(x, \theta_0)^\top \Sigma^{-1}\Phi_h(x, \theta_0)} \\ &\quad - \beta_2 \cdot \sqrt{\phi(x, w_0)^\top \Lambda^{-1}\phi(x, w_0)}, H\}^+,\end{aligned}$$

in which $\|\theta - \theta_0\|_2 \le H\sqrt{N/\lambda_1}$, $\|w - w_0\|_2 \le H\sqrt{N/\lambda_2}$, $\beta_1 \in [0, R_{\beta_1}]$, $\beta_2 \in [0, R_{\beta_2}]$, $\|\Sigma\|_2 \ge \lambda_1$ and $\|\Lambda\|_2 \ge \lambda_2$. To this end, for any $\epsilon > 0$, we define $\mathcal{N}_{\epsilon,h}^v$ as the $\epsilon-$covering number of $\mathcal{V}_h(x, R_{\beta_1}, R_{\beta_2}, \lambda_1, \lambda_2)$ with respect to the $\ell_\infty-$norm on $\mathcal{X}$, and we let $\mathcal{N}_\epsilon^v = \max_{h\in[H]}\{\mathcal{N}_{\epsilon,h}^v\}$.

With the estimation terms and the penalty terms constructed, we present our PARTED for general MDPs with neural network function approximation as follows.

Algorithm 2 Neural Pessimistic Value Iteration With Reward Decomposition (PARTED)

Input: Dataset $\mathcal{D} = \{\tau_i, r(\tau_i)\}_{i,h=1}^{N,H}$
Initialization: Set $\widehat{V}_{H+1}$ as zero function
Obtain $\widehat{R}_h$ and $\widehat{\Theta}$ according to eq. (31)
for $h = H, H-1, \cdot, 1$ **do**
  Obtain $\widehat{\mathbb{P}}_h\widehat{V}_{h+1}$ and $\widehat{w}_h$ according to eq. (25)
  Obtain $\Gamma_h(\cdot, \widehat{\Theta}, \widehat{w}_h)$ according to eq. (29)
  Obtain $\widehat{Q}_h(\cdot)$ and $\widehat{V}_h(\cdot)$ according to eq. (37) and let $\widehat{\pi}_h(\cdot|s) = \operatorname{argmax}_{\pi_h}\langle\widehat{Q}_h(s, \cdot), \pi_h(\cdot|s)\rangle$
end for

APPENDIX B
RESULTS FOR GENERAL MDPS WITH NEURAL NETWORK FUNCTION APPROXIMATION

In this section, we present the major results for the general MDPs.

We first make the following standard assumption on the activate function of the neural network, which can be satisfied by a number of activation functions such as ReLU and $\tanh(\cdot)$.

Assumption 2: For all $x \in \mathcal{X}$, we have $|\sigma'(x)| \le C_\sigma < +\infty$ and $\sigma'(0) = 0$.

We initialize b and w via a symmetric initialization scheme [68], [69]: for any $1 \le r \le m$ we set $b_{0,r} \sim \text{Unif}(\{-1, 1\})$ and $w_{0,r} \sim N(0, I_d/d)$, where I_d is the identity matrix in $\mathbb{R}^d$, and for any $m + 1 \le r \le 2m$, we set $b_{0,r} = -b_{0,r-m}$ and $w_{0,r} = w_{0,r-m}$. Under such an initialization, the initial neural network is a zero function, i.e.

$f(x;b_0,w_0)=0$ for all $x\in\mathcal{X}$, where $b_0=[b_{0,1},\cdots,b_{0,2m}]^\top$ and $w_0=[w_{0,1}^\top,\cdots,w_{0,2m}^\top]^\top$ are initialization parameters. During training, we fix the value of b at its initial value and only optimize w. To simplify the notation, we denote $f(x;b,w)$ as $f(x;w)$ and $\nabla_w f(x,w)$ as $\phi(x,w)$.

In the overparameterized scheme, the neural network width $2m$ is considered to be much larger than the number of trajectories N and horizon length H. Under such a scheme, the training process of neural networks can be captured by the framework of neural tangent kernel (NTK) [70]. Specifically, conditioning on the realization of w_0, we define a kernel $K(x,x'):\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ as

$$\begin{aligned}&K(x,x')=\langle\phi(x,w_0),\phi(x',w_0)\rangle\\&=\frac{1}{2m}\sum_{r=1}^{2m}\sigma'(w_{0,r}^\top x)\sigma'(w_{0,r}^\top x')x^\top x',\\&\forall(x,x')\in\mathcal{X}\times\mathcal{X},\end{aligned}$$

where $\sigma'(\cdot)$ is the derivative of the action function $\sigma(\cdot)$. It can be shown that $f(\cdot,w)$ is close to its linearization at w_0 when m is sufficiently large and w is not too far away from w_0, i.e.,

$$\begin{aligned}f(x,w)&\approx f_0(x,w)\\&=f(x,w_0)+\langle\phi(x,w_0),w-w_0\rangle\\&=\langle\phi(x,w_0),w-w_0\rangle,\quad\forall x\in\mathcal{X}.\end{aligned}$$

Note that $f_0(x,w)$ belongs to a reproducing kernel Hilbert space (RKHS) with kernel $K(\cdot,\cdot)$. Similarly, consider the sum of H neural networks $f(\tau,\Theta)=\sum_{h=1}^H f(x_h^\tau,\theta_h)$ with the same initialization θ_0 for each neural network, where $\tau=[x_1^\top,\cdots,x_H^\top]^\top$ and $\Theta=[\theta_1^\top,\cdots,\theta_H^\top]^\top$. If θ_h is not too far away from θ_0 for all $h\in[H]$ and m is sufficiently large, it can be shown that the dynamics of $f(\tau,\Theta)$ belong to a RKHS with kernel K_H defined as $K_H(\tau,\tau')=\sum_{h=1}^H K(x_h,x_h')$. We further define $\mathcal{H}_K$ and $\mathcal{H}_{K_H}$ as the RKHS induced by $K(\cdot,\cdot)$ and $K_H(\cdot,\cdot)$, respectively.

Based on the kernel $K(\cdot,\cdot)$ and $K_H(\cdot,\cdot)$, we define the Gram matrix $K_r,K_{v,h}\in\mathbb{R}^{N\times N}$ as

$$\begin{aligned}K_r&=[K_H(\tau_i,\tau_j)]_{i,j\in[N]},\\K_{v,h}&=[K(x_h^{\tau_i},x_h^{\tau_j})]_{i,j\in[N]}.\end{aligned}$$

We further define a function class as follows

$$\begin{aligned}&\mathcal{F}_{B_1,B_2}\\&=\Bigg\{f_\ell(x)=\int_{\mathbb{R}^d}\sigma'(w^\top x)\cdot x^\top\ell(w)dp(w):\\&\quad\sup_w\|\ell(w)\|_2\le B_1,\sup_w\frac{\|\ell(w)\|_2}{p(w)}\le B_2\Bigg\},\end{aligned}$$

where $\ell:\mathbb{R}^d\to\mathbb{R}^d$ is a mapping, B_1,B_2 are positive constants, and p is the density of $N(0,I_d/d)$. We then make the following assumption regarding the expressive power of the above function class.

Assumption 3: We assume that there exist $a_1,a_2,A_1,A_2>0$ such that $R_h(\cdot)\in\mathcal{F}_{a_1,a_2}$ and $(\mathbb{P}_h f)(\cdot)\in\mathcal{F}_{A_1,A_2}$ for any $f(\cdot):\mathcal{X}\to[0,H]$.

Assumbtion 3 ensures that both $R_h(\cdot)$ and $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ can be captured by an infinite width neural network. Note that Assumbtion 3 is mild since $\mathcal{F}_{B_1,B_2}$ is an expressive function class as shown in Lemma C.1 of [68]. Similar assumptions have also been adopted in many previous works that consider neural network function approximation [59], [71], [72], [73], [74]. Additionally, we assume that the data collection process explores the state-action space and trajectory space well. Note that similar assumptions have also been adopted in [26], [29], and [62]. We similarly assume the dataset is collected by a well-explored behavior policy as the linear MDP setting.

Assumption 4 (Well-Explored Dataset): Suppose the N trajectories in dataset $\mathcal{D}$ are independently and identically induced by a fixed behaviour policy μ. There exist absolute constants $C_\sigma>0$ and $C_\varsigma>0$ such that for all $h\in[H]$,

$$\begin{aligned}\lambda_{\min}(\overline{M}(\Theta_0))&\ge C_\sigma\\\lambda_{\min}(\overline{m}_h(w_0))&\ge C_\varsigma,\end{aligned}$$

where

$$\begin{aligned}\overline{M}(\Theta_0)&=\mathbb{E}_\mu\left[\Phi(\tau,\Theta_0)\Phi(\tau,\Theta_0)^\top\right]\\\overline{m}_h(w_0)&=\mathbb{E}_\mu\left[\phi(x_h^\tau,w_0)\phi(x_h^\tau,w_0)^\top\right].\end{aligned}$$

We can now formally present the suboptimality of the policy $\widehat{\pi}$ obtained via Algorithm 2.

Theorem 3: Consider Algorithm 2. Suppose Assumbtion 2-4 hold. Let $\lambda_1=\lambda_2=1+1/N$, $\beta_1=R_{\beta_1}$ and $\beta_2=R_{\beta_2}$, in which R_{β_1} and R_{β_2} satisfy

$$\begin{aligned}R_{\beta_1}&\ge H\Bigg(4a_2^2\lambda_1/d+2\log\det\left(I+K_r/\lambda_1\right)\\&\quad+10\log(NH^2)\Bigg)^{1/2},\\R_{\beta_2}&\ge H\left(8A_2^2\lambda_2/d+6C_\epsilon+16\log(NH^2\mathcal{N}_\epsilon^v)\right)^{1/2}\\&\quad+4\max_{h\in[H]}\{\log\det\left(I+K_{v,h}/\lambda_2\right)\},\end{aligned}$$

where $\epsilon=\sqrt{\lambda_2 C_\epsilon}H/(2NC_\phi)$, $C_\epsilon\ge1$ is an adjustable parameter, and $C_\phi>0$ is an absolute constant. In addition, let m be sufficiently large. Then, with probability at least $1-(N^2H^4)^{-1}$, we have

$$\text{SubOpt}(\widehat{\pi},s)\le\widetilde{\mathcal{O}}\left(\tfrac{H\max\{\beta_1,\beta_2\}}{\sqrt{N}}\right)+\varepsilon_1,$$

where

$$\begin{aligned}\varepsilon_1&=\max\{\beta_1H^{5/3},\beta_2H^{7/6}\}\widetilde{\mathcal{O}}\left(\frac{N^{1/12}}{m^{1/12}}\right)\\&\quad+\widetilde{\mathcal{O}}\left(\frac{H^{17/6}N^{5/3}}{m^{1/6}}\right).\end{aligned}$$

Theorem 2 shows that Algorithm 2 can find an ϵ-optimal policy with $\widetilde{\mathcal{O}}(H^2\max\{\beta_1,\beta_2\}^2/\epsilon^2)$ episodes of offline data in the trajectory-wise reward setting up to a function approximation error ε_1, which vanishes as the neural network width $2m$ increases. Note that the dependence of ε_1 on the network width, which is $\mathcal{O}(m^{-1/12})$, matches that of the approximation error in the previous work of value iteration algorithm with neural network function approximation [71].

A. Discussion of Proof of Theorem 2

Comparing to the analysis of PEVI for linear MDP with instantaneous reward, which has been extensively studied in offline RL [26], [27], [29], our analysis needs to address the following two new challenges: (1) In instantaneous reward setting, both $R_h(\cdot)$ and $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ can be learned together by solving a single regression problem in per-step scale. However, in our trajectory-wise reward setting, $R_h(\cdot)$ and $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ need to be learned separately by solving two regression problems (eqs. (31) and (25)) in different scales, i.e., eq. (31 is in trajectory scale and eq. (25) is in per-step scale. In order to apply union concentrations to bound the Bellman estimation error $|(\mathbb{B}_h\widehat{V}_h)(\cdot)-(\widehat{\mathbb{B}}_h\widehat{V}_h)(\cdot)|$, we need to develop new techniques to handle the mismatch between eqs. (31) and (25) in terms of scale. (2) In linear MDP, both $R_h(\cdot)$ and $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ can be captured exactly by linear functions. However, in the more general MDP that we consider, we need to develop new analysis to bound the estimation error that caused by the insufficient expressive power of neural networks in order to characterize the optimality of $\widehat{\theta}_h$ and $\widehat{w}_h$ in eqs. (31) and (25), respectively.

To obtain a more concrete suboptimality bound for Algorithm 2, we impose an assumption on the spectral structure of kernels K_H and K.

Assumption 5 (Finite Spectrum NTK [71]): Conditioned on the randomness of (b_0, w_0), let T_{K_H} and T_K be the integral operator induced by K_H and K (see Section J for definition of T_{K_H} and T_K), respectively, and let $\{\omega_j\}_{j\geq 1}$ and $\{\upsilon_j\}_{j\geq 1}$ be eigenvalues of T_{K_H} and T_K, respectively. We have $\omega_j = 0$ for all $j \geq D_1 + 1$ and $\upsilon_j = 0$ for all $\upsilon_j \geq D_2 + 1$, where D_1, D_2 are positive integers.

Assumbtion 5 implies that $\mathcal{H}_{K_H}$ and $\mathcal{H}_K$ are D_1-dimensional and D_2-dimensional, respectively. For concrete examples of neural networks that satisfy Assumbtion 5, please refer to Section B.3 in [71]. Note that such an assumption is in parallel to the "effective dimension" assumption in [75] and [76].

Corollary 2: Consider Algorithm 2. Suppose Assumbtion 2-5 hold. Let $\lambda_1 = \lambda_2 = 1 + 1/N$, $\beta_1 = \widetilde{\mathcal{O}}(HD_1)$ and $\beta_2 = \widetilde{\mathcal{O}}(H\max\{D_1, D_2\})$. Then, with probability at least $1-(N^2H^4)^{-1}$, we have

$$\text{SubOpt}(\widehat{\pi}, s) = \widetilde{\mathcal{O}}\left(D_{\text{eff}}H^2/\sqrt{N}\right) + \varepsilon_2,$$

where $D_{\text{eff}} = \max\{D_1, D_2\}$ denotes the effective dimension and

$$\varepsilon_2 = \max\left\{\sqrt{H}, \max\{D_1, D_2\}, \tfrac{H^{5/3}N^{19/12}}{m^{1/12}}\right\} \cdot \widetilde{\mathcal{O}}\left(\frac{H^{13/6}N^{1/12}}{m^{1/12}}\right).$$

Corollary 2 states that when β_1 and β_2 are chosen properly according to the dimension of $\mathcal{H}_{K_H}$ and $\mathcal{H}_K$, the suboptimality of the policy $\widehat{\pi}$ incurred by Algorithm 2 converges to an ϵ-optimal policy with $\widetilde{\mathcal{O}}(D_{\text{eff}}^2H^4/\epsilon^2)$ episodes of offline data up to a function approximation error ε_2.

APPENDIX C
PROOF FLOW OF THEOREM 2

In this section, we present the main proof flow of Theorem 2.

Theorem 4: Let $\lambda_1 = \lambda_2 = 1 + 1/N$, $\beta_1 = R_{\beta_1}$ and $\beta_2 = R_{\beta_2}$, in which R_{β_1} and R_{β_2} satisfy

$$R_{\beta_1} \geq H\big(4a_2^2\lambda_1/d + 2\log\det(I + K_r/\lambda_1) + 10\log(NH^2)\big)^{1/2},$$

$$R_{\beta_2} \geq H\Big(8A_2^2\lambda_2/d + 6C_\epsilon + 16\log(NH^2\mathcal{N}_\epsilon^v) + 4\max_{h\in[H]}\{\log\det(I + K_{v,h}/\lambda_2)\}\Big)^{1/2},$$

where $\epsilon = \sqrt{\lambda_2 C_\epsilon}H/(2NC_\phi)$, $C_\epsilon \geq 1$ is an adjustable parameter, and $C_\phi > 0$ is an absolute constant. In addition, let m be sufficiently large. Then, with probability at least $1-(N^2H^4)^{-1}$, we have

$$\text{SubOpt}(\widehat{\pi}, s) \leq \widetilde{\mathcal{O}}\left(\tfrac{H\max\{\beta_1,\beta_2\}}{\sqrt{N}}\right) + \varepsilon_1,$$

where

$$\varepsilon_1 = \max\{\beta_1 H^{5/3}, \beta_2 H^{7/6}\}\widetilde{\mathcal{O}}\left(\frac{N^{1/12}}{m^{1/12}}\right) + \widetilde{\mathcal{O}}\left(\frac{H^{17/6}N^{5/3}}{m^{1/6}}\right).$$

We first decompose the suboptimality $\text{SubOpt}(\pi, s)$, and then present the two main results of Lemma 1 and Lemma 2 to bound the evaluation error and summation of penality terms, respectively. The detailed proof of Lemma 1 and Lemma 2 can be found at Section F and Section G.

We define the evaluation error at each step $h \in [H]$ as

$$\delta_h(s,a) = (\mathbb{B}_h\widehat{V}_{h+1})(s,a) - \widehat{Q}_h(s,a), \tag{38}$$

where $\mathbb{B}_h$ is the Bellman operator defined in Section II-A and $\widehat{V}_h$ and $\widehat{Q}_h$ are estimation of state- and state-action value functions, respectively. To proceed the proof, we first decompose the suboptimality into three parts as follows via the standard technique (see Section A in [26]).

$$\begin{aligned}
&\textit{SubOpt}(\pi, s)\\
&= -\sum_{h=1}^{H}\mathbb{E}_\pi\left[\delta_h(s_h,a_h)\big|s_1 = s\right]\\
&\quad + \sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\delta_h(s_h,a_h)\big|s_1 = s\right]\\
&\quad + \sum_{h=1}^{H}\mathbb{E}_{\pi^*}\Big[\langle\widehat{Q}_h(s_h,\cdot), \pi_h^*(\cdot|s_h)\\
&\quad - \widehat{\pi}_h(\cdot|s_h)\rangle\big|s_1 = s\Big].
\end{aligned} \tag{39}$$

In Algorithm 2, the output policy at each horizon $\widehat{\pi}_h$ is greedy with respect to the estimated Q-value $\widehat{Q}_h$. Thus, we have for $\forall h \in [H], \forall s_h \in \mathcal{S}$

$$\langle\widehat{Q}_h(s_h,\cdot), \pi_h^*(\cdot|s_h) - \widehat{\pi}_h(\cdot|s_h)\rangle \leq 0.$$

According to eq. (39), we have the following holds for the suboptimality of $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$

$$\begin{aligned}&\text{SubOpt}(\widehat{\pi}, s)\\ &= -\sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}\left[\delta_h(s_h, a_h)\middle| s_1 = s\right]\\ &\quad + \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\delta_h(s_h, a_h)\middle| s_1 = s\right]. \end{aligned} \tag{40}$$

In the following lemma, we provide the first main technical result for the proof, which bounds the evaluation error $\delta_h(s, a)$. Recall that we use $\mathcal{X}$ to represent the joint state-action space $\mathcal{S} \times \mathcal{A}$ and use x to represent a state action pair (s, a).

Lemma 1: Let λ_1, $\lambda_2 = 1 + 1/N$. Suppose Assumbtion 3 holds. With probability at least $1 - \mathcal{O}(N^{-2}H^{-4})$, it holds for all $h \in [H]$ and $x \in \mathcal{X}$ that

$$\begin{aligned}&-\varepsilon_b \le \delta_h(x)\\ &\quad \le 2\left[\beta_1 \cdot b_{r,h}(x, \widehat{\Theta}) + \beta_2 \cdot b_{v,h}(x, \widehat{w}_h) + \varepsilon_b\right],\end{aligned}$$

where

$$\begin{aligned}\varepsilon_b &= \max\{\beta_1 H^{2/3}, \beta_2 H^{1/6}\}\mathcal{O}\left(\frac{N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right)\\ &\quad + \mathcal{O}\left(\frac{H^{17/6}N^{5/3}\sqrt{\log(N^2H^5m)}}{m^{1/6}}\right),\\ \beta_1 &= H\Bigg(\frac{4a_2^2\lambda_1}{d} + 2\log\det\left(I + \frac{K_N^r}{\lambda_1}\right)\\ &\quad + 10\log(NH^2)\Bigg)^{1/2},\\ \beta_2 &= H\Bigg(\frac{8A_2^2\lambda_2}{d} + 4\max\left\{\log\det\left(I + \frac{K_{N,h}^v}{\lambda_2}\right)\right\}\\ &\quad + 6C_\epsilon + 16\log(NH^2\mathcal{N}_\epsilon^v)\Bigg)^{1/2},\\ \epsilon &= \sqrt{\lambda_2 C_\epsilon}H/(2NC_\phi),\ \text{where } C_\epsilon \ge 1.\end{aligned}$$

Proof: The main technical development of the proof lies in handling the uncertainty caused by redistributing the trajectory-wise reward via solving a trajectory-level regression problem and analyzing the dynamics of neural network optimization. The detailed proof is provided in Section F. □

Applying Lemma 1 to eq. (40) yields

$$\begin{aligned}&\text{SubOpt}(\widehat{\pi}, s)\\ &= -\sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}\left[\delta_h(s_h, a_h)\middle| s_1 = s\right]\\ &\quad + \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\delta_h(s_h, a_h)\middle| s_1 = s\right]\\ &\le 3H\varepsilon_b + \max_x 2\beta_1 \cdot \sum_{h=1}^H b_{r,h}(x, \widehat{\Theta})\\ &\quad + \max_x 2\beta_2 \cdot \sum_{h=1}^H b_{v,h}(x, \widehat{w}_h).\end{aligned} \tag{41}$$

The following lemma captures the second main technical result for the proof, which bounds the summation of the penalty terms $\beta_1 \cdot \sum_{h=1}^H b_{r,h}(x, \widehat{\Theta}) + \beta_2 \cdot \sum_{h=1}^H b_{v,h}(x, \widehat{w}_h)$.

Lemma 2: Suppose Assumbtion 3&4 hold. We have the following holds with probability $1 - \mathcal{O}(N^{-2}H^{-4})$

$$\begin{aligned}&\beta_1 \cdot \sum_{h=1}^H b_{r,h}(x, \widehat{\Theta}) + \beta_2 \cdot \sum_{h=1}^H b_{v,h}(x, \widehat{w}_h)\\ &\le \left(\frac{\beta_1}{\sqrt{C_\sigma}} + \frac{\beta_2}{\sqrt{C_\varsigma}}\right)\frac{\sqrt{2}HC_\phi}{\sqrt{N}}\\ &\quad + \max\{\beta_1 H^{5/3}, \beta_2 H^{7/6}\}\\ &\quad \cdot \mathcal{O}\left(\frac{N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right).\end{aligned}$$

Proof: The proof develops new analysis to characterize the summation of the penality term $b_{r,h}$ constructed by trajectory features, which is unique in the trajectory-wise reward setting. The detailed proof is provided in Section G. □

Applying Lemma 2 to eq. (41), we have

$$\begin{aligned}&\text{SubOpt}(\widehat{\pi}, s)\\ &\le 3H\varepsilon_b + \left(\frac{\beta_1}{\sqrt{C_\sigma}} + \frac{\beta_2}{\sqrt{C_\varsigma}}\right)\frac{2\sqrt{2}HC_\phi}{\sqrt{N}}\\ &\quad + \max\{\beta_1 H^{5/3}, \beta_2 H^{7/6}\}\\ &\quad \cdot \mathcal{O}\left(\frac{N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right)\\ &\le 4H\varepsilon_b + \left(\frac{\beta_1}{\sqrt{C_\sigma}} + \frac{\beta_2}{\sqrt{C_\varsigma}}\right)\frac{2\sqrt{2}HC_\phi}{\sqrt{N}},\end{aligned} \tag{42}$$

which completes the proof.

Appendix D
Proof of Corollary 2

To provide a concrete bound for SubOpt$(\widehat{\pi}, s)$ defined in eq. (42), we first need to bound the penalty coefficients β_1, β_2 under Assumbtion 5. Recalling the properties of β_1, β_2 in Theorem 2, we have

$$\begin{aligned}&H\Bigg(\frac{4a_2^2\lambda_1}{d} + 2\log\det\left(I + \frac{K_N^r}{\lambda_1}\right)\\ &\quad + 10\log(NH^2)\Bigg)^{1/2}\\ &\le R_{\beta_1} = \beta_1,\end{aligned} \tag{43}$$

$$\begin{aligned}&H\Bigg(\frac{8A_2^2\lambda_2}{d} + 4\max_{h\in[H]}\left\{\log\det\left(I + \frac{K_{N,h}^v}{\lambda_2}\right)\right\}\\ &\quad + 6C_\epsilon + 16\log(NH^2\mathcal{N}_\epsilon^v)\Bigg)^{1/2}\\ &\le R_{\beta_2} = \beta_2.\end{aligned} \tag{44}$$

Recall that we use $\mathcal{X}$ to represent the joint state-action space $\mathcal{S} \times \mathcal{A}$ and use x to represent a state action pair (s, a). We define the maximal information gain associated with RHKS with kernels K_N^r and $K_{N,h}^v$ as follows

$$\begin{aligned}&\Gamma_{K_N^r}(N, \lambda_1)\\ &= \sup_{\mathcal{D}\subset\mathcal{D}_\tau}\{1/2 \cdot \log\det(I_{2dmH} + \lambda_1^{-1} \cdot K_N^r)\},\end{aligned} \tag{45}$$

$$\Gamma_{K^v_{N,h}}(N,\lambda_2) = \sup_{\mathcal{D}\subset\mathcal{D}_x}\{1/2\cdot\log\det(I_{2dm}+\lambda_2^{-1}\cdot K^v_{N,h})\}, \tag{46}$$

where $\mathcal{D}_x$ and $\mathcal{D}_\tau$ are discrete subsets of state-action pair $x\in\mathcal{X}$ and trajectory $\tau\in\mathcal{X}\times\cdots\times\mathcal{X}$ with cardinality no more than N, respectively. Applying Lemma 9 in Section J and Assumbtion 5, we have

$$\Gamma_{K^r_N}(N,\lambda_1)\le C_{K_1}\cdot D_1\cdot\log N \tag{47}$$

and

$$\Gamma_{K^v_{N,h}}(N,\lambda_2)\le C_{K_2}\cdot D_2\cdot\log N, \tag{48}$$

where C_{K_1}, C_{K_2} are absolute constants. Recall that $\mathcal{N}^v_{\epsilon,h}$ is the cardinality of the function class. Next, we proceed to bound the term $\mathcal{N}^v_\epsilon=\max_{h\in[H]}\{\mathcal{N}^v_{\epsilon,h}\}$.

$$\begin{aligned}&\mathcal{V}_h(x,R_\theta,R_w,R_{\beta_1},R_{\beta_2},\lambda_1,\lambda_2)\\&=\Big\{\max_{a\in\mathcal{A}}\{\overline{Q}_h(s,a,\theta,w,\beta_1,\beta_2,\Sigma,\Lambda)\}:\mathcal{S}\to[0,H],\\&\quad\|\theta-\theta_0\|_2\le R_\theta,\|w-w_0\|_2\le R_w,\beta_1\in[0,R_{\beta_1}],\\&\quad\beta_2\in[0,R_{\beta_2}],\|\Sigma\|_2\ge\lambda_1,\|\Lambda\|_2\ge\lambda_2\Big\},\end{aligned}$$

where $R_\theta=H\sqrt{N/\lambda_1}$, $R_w=H\sqrt{N/\lambda_2}$, and

$$\begin{aligned}&\overline{Q}_h(x,\theta,w,\beta_1,\beta_2,\Sigma,\Lambda)\\&=\min\{\langle\phi(x,\theta_0),\theta-\theta_0\rangle+\langle\phi(x,w_0),w-w_0\rangle\\&\quad-\beta_1\cdot\sqrt{\Phi_h(x,\theta_0)^\top\Sigma^{-1}\Phi_h(x,\theta_0)}\\&\quad-\beta_2\cdot\sqrt{\phi(x,w_0)^\top\Lambda^{-1}\phi(x,w_0)},H\}^+.\end{aligned}$$

Note that

$$\begin{aligned}&\Big|\max_{a\in\mathcal{A}}\{\overline{Q}_h(s,a,\theta,w,\beta_1,\beta_2,\Sigma,\Lambda)\}\\&\quad-\max_{a\in\mathcal{A}}\{\overline{Q}_h(s,a,\theta',w',\beta_1',\beta_2',\Sigma',\Lambda')\}\Big|\\&\le\max_{a\in\mathcal{A}}\Big|\overline{Q}_h(s,a,\theta,w,\beta_1,\beta_2,\Sigma,\Lambda)\\&\quad-\overline{Q}_h(s,a,\theta',w',\beta_1',\beta_2',\Sigma',\Lambda')\Big|\\&\overset{(i)}{\le}\max_{a\in\mathcal{A}}|\langle\phi(x,\theta_0),\theta-\theta'\rangle|\\&\quad+\max_{a\in\mathcal{A}}|\langle\phi(x,w_0),w-w'\rangle|\\&\quad+\max_{a\in\mathcal{A}}\Big|(\beta_1-\beta_1')\\&\quad\cdot\sqrt{\Phi_h(x,\theta_0)^\top\Sigma^{-1}\Phi_h(x,\theta_0)}\Big|\\&\quad+\max_{a\in\mathcal{A}}\Big|\beta_1'\cdot\Big[\sqrt{\Phi_h(x,\theta_0)^\top\Sigma^{-1}\Phi_h(x,\theta_0)}\\&\quad-\sqrt{\Phi_h(x,\theta_0)^\top\Sigma'^{-1}\Phi_h(x,\theta_0)}\Big]\Big|\\&\quad+\max_{a\in\mathcal{A}}\Big|(\beta_2-\beta_2')\cdot\sqrt{\phi(x,w_0)^\top\Lambda^{-1}\phi(x,w_0)}\Big|\\&\quad+\max_{a\in\mathcal{A}}\Big|\beta_2'\cdot\Big[\sqrt{\phi(x,w_0)^\top\Lambda^{-1}\phi(x,w_0)}\\&\quad-\sqrt{\phi(x,w_0)^\top\Lambda'^{-1}\phi(x,w_0)}\Big]\Big|\\&\overset{(ii)}{\le}\max_{a\in\mathcal{A}}|\langle\Phi_h(x,\Theta_0),\Theta-\Theta'\rangle|\\&\quad+\max_{a\in\mathcal{A}}|\langle\phi(x,w_0),w-w'\rangle|+\frac{C_\phi}{\sqrt{\lambda_1}}|\beta_1-\beta_1'|\\&\quad+\frac{C_\phi}{\sqrt{\lambda_2}}|\beta_2-\beta_2'|\\&\quad+R_{\beta_1}\max_{a\in\mathcal{A}}|\|\Phi_h(x,\theta_0)\|_{\Sigma^{-1}}-\|\Phi_h(x,\theta_0)\|_{\Sigma'^{-1}}|\\&\quad+R_{\beta_2}\max_{a\in\mathcal{A}}|\|\phi(x,w_0)\|_{\Lambda^{-1}}-\|\phi(x,w_0)\|_{\Lambda'^{-1}}|,\end{aligned}\tag{49}$$

where (i) follows from contractive properties of operators $\min\{\cdot,H\}^+$ and $\max_{a\in\mathcal{A}}\{\cdot\}$ and the triangle inequality, and (ii) follows from the fact that $\|\phi(x,w_0)\|_2,\|\Phi_h(x,\theta_0)\|_2\le C_\phi$.

Following arguments similar to those in the proof of Corollaries 4.8, Corollaries 4.4 and Section D.1 in [71], we have the followings hold for terms in the right hand side of eq. (49)

$$\begin{aligned}&|\langle\Phi_h(x,\Theta_0),\Theta-\Theta'\rangle|=|g_1(x)-g_2(x)|\\&\text{where }\|g_i\|_{\mathcal{H}_{K_H}}\le R_g=2H\sqrt{\Gamma_{K^r_N}(N,\lambda_1)},\\&|\langle\phi(x,w_0),w-w'\rangle|=|h_1(x)-h_2(x)|\\&\text{where }\|h_i\|_{\mathcal{H}_K}\le R_h=2H\sqrt{\Gamma_{K^v_{N,h}}(N,\lambda_2)},\\&|\|\Phi_h(x,\theta_0)\|_{\Sigma^{-1}}-\|\Phi_h(x,\theta_0)\|_{\Sigma'^{-1}}|\\&=|\|\Psi(x)\|_\Omega-\|\Psi(x)\|_{\Omega'}|,\\&|\|\phi(x,w_0)\|_{\Lambda^{-1}}-\|\phi(x,w_0)\|_{\Lambda'^{-1}}|\\&=|\|\psi(x)\|_\Upsilon-\|\psi(x)\|_{\Upsilon'}|,\end{aligned}\tag{50, 51}$$

where $g_1(\cdot),g_2(\cdot)$ are two functions in RKHS $\mathcal{H}_{K_H}$, $h_1(\cdot),h_2(\cdot)$ are two functions in RKHS $\mathcal{H}_K$, $\Psi(\cdot)$ and $\psi(\cdot)$ are feature mappings of RKHSs $\mathcal{H}_{K_H}$ and $\mathcal{H}_K$, respectively, $\Omega,\Omega':\mathcal{H}_{K_H}\to\mathcal{H}_{K_H}$ are self-adjoint operators with eigenvalues bounded in $[0,1/\lambda_1]$, and $\Upsilon,\Upsilon':\mathcal{H}_K\to\mathcal{H}_K$ are self-adjoint operators with eigenvalues bounded in $[0,1/\lambda_2]$. We define the following two function classes

$$\mathcal{F}_1=\{\|\Psi(\cdot)\|_\Omega:\|\Omega\|_2\le1/\lambda_1\},\tag{52}$$

and

$$\mathcal{F}_2=\{\|\psi(\cdot)\|_\Upsilon:\|\Upsilon\|_2\le1/\lambda_2\}.\tag{53}$$

For any $\epsilon>0$, we denote $\mathcal{N}(\epsilon,\mathcal{H},R)$ as the ϵ-covering of $\{f\in\mathcal{H}:\|f\|_\mathcal{H}\le R\}$, denote $\mathcal{N}(\epsilon,\mathcal{F}_1,\lambda_1)$ as the ϵ-covering number of $\mathcal{F}_1$ in eq. (52), denote $\mathcal{N}(\epsilon,\mathcal{F}_2,\lambda_2)$ as the ϵ-covering number of $\mathcal{F}_2$ in eq. (53), and denote $\mathcal{N}(\epsilon,R)$ as the ϵ-covering number of the interval $[0,R]$ with respect to the Euclidean distance. Note that eq. (49) implies

$$\begin{aligned}\mathcal{N}^v_{\epsilon,h}\le&\,\mathcal{N}(\epsilon/6,\mathcal{H}_{K^H_m},R_g)\cdot\mathcal{N}(\epsilon/(6R_{\beta_2}),\mathcal{F}_2,\lambda_2)\\&\cdot\mathcal{N}(\epsilon/6,\mathcal{H}_{K_m},R_h)\end{aligned}$$

$$\cdot \mathcal{N}(\epsilon/(6C_\phi), R_{\beta_1}) \cdot \mathcal{N}(\epsilon/(6C_\phi), R_{\beta_2}) \cdot \mathcal{N}(\epsilon/(6R_{\beta_1}), \mathcal{F}_1, \lambda_1). \tag{54}$$

Based on Corollary 4.1.13 in [77], we have the followings hold for $\mathcal{N}(\epsilon/(6C_\phi), R_{\beta_1})$ and $\mathcal{N}(\epsilon/(6C_\phi), R_{\beta_2})$, respectively

$$\mathcal{N}(\epsilon/(6C_\phi), R_{\beta_1}) \le 1 + 12C_\phi R_{\beta_1}/\epsilon \text{ and } \mathcal{N}(\epsilon/(6C_\phi), R_{\beta_2}) \le 1 + 12C_\phi R_{\beta_2}/\epsilon. \tag{55}$$

Moreover, as shown in Lemma D.2 and Lemma D.3 in [71], under the finite spectrum NTK assumption in Assumbtion 5, we have the followings hold

$$\log \mathcal{N}(\epsilon/6, \mathcal{H}_{K_m^H}, R_g) \le C_1 \cdot D_1 \cdot [\log(6R_g/\epsilon) + C_2], \tag{56}$$

$$\log \mathcal{N}(\epsilon/6, \mathcal{H}_{K_m}, R_h) \le C_3 \cdot D_2 \cdot [\log(6R_h/\epsilon) + C_4], \tag{57}$$

$$\log \mathcal{N}(\epsilon/(6R_{\beta_1}), \mathcal{F}_1, \lambda_1) \le C_5 \cdot D_1^2 \cdot [\log(6R_{\beta_1}/\epsilon) + C_6], \tag{58}$$

$$\log \mathcal{N}(\epsilon/(6R_{\beta_2}), \mathcal{F}_2, \lambda_2) \le C_7 \cdot D_2^2 \cdot [\log(6R_{\beta_2}/\epsilon) + C_8]. \tag{59}$$

where C_i ($i \in \{1, \cdots, 8\}$) are absolute constants that do not rely on N, H or ϵ. Then, substituting eq. (55)-(59) into eq. (54), we have

$$\begin{aligned}
&\log \mathcal{N}_{\epsilon,h}^v \\
&\le \mathcal{N}(\epsilon/6, \mathcal{H}_{K_m^H}, R_g) + \mathcal{N}(\epsilon/6, \mathcal{H}_{K_m}, R_h) \\
&\quad + \mathcal{N}(\epsilon/(6C_\phi), R_{\beta_1}) + \mathcal{N}(\epsilon/(6C_\phi), R_{\beta_2}) \\
&\quad + \mathcal{N}(\epsilon/(6R_{\beta_1}), \mathcal{F}_1, \lambda_1) + \mathcal{N}(\epsilon/(6R_{\beta_2}), \mathcal{F}_2, \lambda_2) \\
&\le \log(1 + 12C_\phi R_{\beta_1}/\epsilon) + \log(1 + 12C_\phi R_{\beta_2}/\epsilon) \\
&\quad + C_1 D_1[\log(6R_g/\epsilon) + C_2] \\
&\quad + C_3 D_2[\log(6R_h/\epsilon) + C_4] \\
&\quad + C_5 D_1^2[\log(6R_{\beta_1}/\epsilon) + C_6] \\
&\quad + C_7 D_2^2[\log(6R_{\beta_2}/\epsilon) + C_8],
\end{aligned} \tag{60}$$

We next proceed to show that there exists an absolute constant $R_{\beta_1} > 0$ such that eq. (43) holds. Substituting eq. (48) to eq. (43), we can obtain

$$\text{L.H.S of eq. (43)} \le H\left(\frac{4a_2^2\lambda_1}{d} + 4C_{K_1} D_1 \log N + 10\log(NH^2)\right)^{1/2}.$$

If we let

$$R_{\beta_1} = C_{\beta_1} H\sqrt{D_1 \log(NH^2)}, \tag{61}$$

in which C_{β_1} is a sufficiently large constant, then we have the following holds

$$\text{L.H.S of eq. (43)} \le R_{\beta_1}.$$

Note that eq. (60) directly implies that

$$\begin{aligned}
&\log \mathcal{N}_\epsilon^v \\
&= \max_{h\in[H]} \{\log \mathcal{N}_{\epsilon,h}^v\} \\
&\le \log(1 + 12C_\phi R_{\beta_1}/\epsilon) + \log(1 + 12C_\phi R_{\beta_2}/\epsilon) \\
&\quad + C_1 D_1[\log(6R_g/\epsilon) + C_2] \\
&\quad + C_3 D_2[\log(6R_h/\epsilon) + C_4] \\
&\quad + C_5 D_1^2[\log(6R_{\beta_1}/\epsilon) + C_6] \\
&\quad + C_7 D_2^2[\log(6R_{\beta_2}/\epsilon) + C_8] \\
&\overset{(i)}{\le} C_1' D_1^2 \log(R_{\beta_1}/\epsilon) + C_2' D_2^2 \log(R_{\beta_2}/\epsilon) \\
&\quad + C_3' D_1 \log(H\sqrt{D_1}/\epsilon) + C_4' D_2 \log(H\sqrt{D_2}/\epsilon) \\
&\overset{(ii)}{\le} C_1'' D_1^2 \log(NH^2\sqrt{D_1}/\epsilon) + C_2' D_2^2 \log(R_{\beta_2}/\epsilon) \\
&\quad + C_3' D_1 \log(H\sqrt{D_1}/\epsilon) \\
&\quad + C_4' D_2 \log(H\sqrt{D_2}/\epsilon),
\end{aligned} \tag{62}$$

where in (i) we let C_1', C_2', C_3' and C_4' be sufficiently large absolute constants, in (ii) we use eq. (61) and let C_1'' be sufficiently large. Then, we proceed to show that there exists an absolute constant $R_{\beta_2} > 0$ such that eq. (44) holds. Using eq. (48) and eq. (62), the left hand side of eq. (44) can be bounded as follows

$$\begin{aligned}
&\text{L.H.S of eq. (44)} \\
&\le H\Big(\frac{8A_2^2\lambda_2}{d} + 8C_{K_2} \cdot D_2 \cdot \log N \\
&\quad + 20\log(NH^2) + 6C_\epsilon + 16\log \mathcal{N}_\epsilon^v\Big)^{1/2} \\
&\overset{(i)}{\le} HC_{\beta_2,1}\sqrt{D_2 \log(NH^2)} + HC_{\beta_2,2}\sqrt{\log \mathcal{N}_\epsilon^v} \\
&\quad + HC_{\beta_2,3}\sqrt{C_\epsilon} \\
&\le HC_{\beta_2,1}\sqrt{D_2 \log(NH^2)} \\
&\quad + HC_{\beta_2,2}\Big[D_1\sqrt{C_1'' \log(NH^2\sqrt{D_1}/\epsilon)} \\
&\quad + D_2\sqrt{C_2' \log(R_{\beta_2}/\epsilon)} \\
&\quad + \sqrt{C_3' D_1 \log(H\sqrt{D_1}/\epsilon)} \\
&\quad + \sqrt{C_4' D_2 \log(H\sqrt{D_2}/\epsilon)}\Big] \\
&\quad + HC_{\beta_2,3}\sqrt{C_\epsilon}.
\end{aligned}$$

where (i) follows from the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $C_{\beta_2,1}$, $C_{\beta_2,2}$ and $C_{\beta_2,3}$ are sufficiently large constants. Clearly, if we let

$$R_{\beta_2} = C_{\beta_2} H \max\{D_1, D_2\} \cdot \log(NH^2 \max\{D_1, D_2\}/\epsilon), \tag{63}$$

where C_{β_2} is a sufficiently large absolute constant, then we have

$$\text{L.H.S of eq. (44)} \le R_{\beta_2}. \tag{64}$$

Finally, substituting the value of R_{β_1} in eq. (61) and value of R_{β_2} in eq. (63) into eq. (42) and letting $C_\epsilon = \max\{D_1, D_2\}^2$ (which implies $\epsilon = \sqrt{\lambda_2}\max\{D_1, D_2\}H/(2NC_\phi)$), we have

$$\begin{aligned}
&\text{SubOpt}(\widehat{\pi}, s) \\
&\le \left(\frac{\beta_1}{\sqrt{C_\sigma}} + \frac{\beta_2}{\sqrt{C_\varsigma}}\right)\frac{2\sqrt{2}HC_\phi}{\sqrt{N}} + 4H\varepsilon_b \\
&\le \max\{R_{\beta_1}, R_{\beta_2}\}\mathcal{O}\left(\frac{H}{\sqrt{N}}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\max\{R_{\beta_1}H^{5/3}, R_{\beta_2}H^{7/6}\}\\
&\cdot\mathcal{O}\left(\frac{N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right)\\
&+\mathcal{O}\left(\frac{H^{23/6}N^{5/3}\sqrt{\log(N^2H^5m)}}{m^{1/6}}\right)\\
\leq\ &\mathcal{O}\Bigg(\frac{H^2\max\{D_1,D_2\}}{\sqrt{N}}\\
&\cdot\log\left(\frac{NH^2\max\{D_1,D_2\}}{\epsilon}\right)\Bigg)\\
&+\max\left\{\sqrt{H},\max\{D_1,D_2\}\right\}\\
&\cdot\mathcal{O}\left(\frac{H^{13/6}N^{1/12}(\log m)^{1/4}}{m^{1/12}}\log\left(N^2H^2\right)\right)\\
&+\mathcal{O}\left(\frac{H^{23/6}N^{5/3}\sqrt{\log(N^2H^5m)}}{m^{1/6}}\right)\\
\overset{(i)}{\leq}\ &\mathcal{O}\left(\frac{H^2\max\{D_1,D_2\}}{\sqrt{N}}\log\left(2C_\phi N^2H\right)\right)\\
&+\max\left\{\sqrt{H},\max\{D_1,D_2\},\frac{H^{5/3}N^{19/12}}{m^{1/12}}\right\}\\
&\cdot\mathcal{O}\left(\frac{H^{13/6}N^{1/12}(\log m)^{1/4}}{m^{1/12}}\log\left(N^2H^5m\right)\right),
\end{aligned}
$$

where (i) follows from the definition of ϵ and the fact that $\lambda_2\geq 1$.

Appendix E
Linear MDP With Trajectory-Wise Reward

In this section, we present the full details of our study on the offline RL in the linear MDP setting with trajectory-wise rewards.

A. Linear MDP and Algorithm

We define the linear MDP [53] as follows, where the transition kernel and expected reward function are linear in a feature map. We use $\mathcal{X}$ to represent the joint state-action space $\mathcal{S}\times\mathcal{A}$ and use x to represent a state action pair.

Definition 1 (Linear MDP): We say an episodic MDP $(\mathcal{S},\mathcal{A},\mathbb{P},r,H)$ is a linear MDP with a known feature map $\phi(\cdot):\mathcal{X}\to\mathbb{R}^d$ if there exist an unknown vector $w_h^*(s)\in\mathbb{R}^d$ over $\mathcal{S}$ and an unknown vector $\theta_h^*\in\mathbb{R}^d$ such that

$$
\begin{aligned}
\mathbb{P}_h(s'|s,a)&=\langle\phi(s,a),w_h^*(s')\rangle,\\
R_h(s,a)&=\langle\phi(s,a),\theta_h^*\rangle,
\end{aligned}
\tag{65}
$$

for all $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$ at each step $h\in[H]$. Here we assume $\|\phi(x)\|_2\leq 1$ for all $x\in\mathcal{X}$ and $\max\{\|w_h^*(\mathcal{S})\|_2,\|\theta_h^*\|_2\}\leq\sqrt{d}$ at each step $h\in[H]$, where with an abuse of notation, we define $\|w_h^*(\mathcal{S})\|_2=\int_{\mathcal{S}}\|w_h^*(s)\|_2\,ds$.

We present our PARTED algorithm for linear MDPs with trajectory-wise rewards in Algorithm 3. Note that Algorithm 3 shares a structure similar to that of Algorithm 2. Specifically, we estimate each $R_h(\cdot)$ for all $h\in[H]$ using a linear function $\langle\phi(s,a),\theta_h\rangle$, where $\theta_h\in\mathbb{R}^d$ is a learnable parameter.

Algorithm 3 Linear Pessimistic Value Iteration With Reward Decomposition (PARTED)

Input: Dataset $\mathcal{D}=\{\tau_i,r(\tau_i)\}_{i,h=1}^{N,H}$
Initialization: Set $\widehat{V}_{H+1}$ as zero function
Obtain $\widehat{R}_h$ and $\widehat{\Theta}$ according to eq. (67)
for $h=H,H-1,\cdot,1$ **do**
  Obtain $\widehat{\mathbb{P}}_h\widehat{V}_{h+1}$ and $\widehat{w}_h$ according to eq. (69)
  Obtain $\Gamma_h(\cdot)$ according to eq. (72)
  $\widehat{Q}_h(\cdot)=\min\{\widehat{R}_h(\cdot)+\widehat{\mathbb{P}}_h\widehat{V}_{h+1}(\cdot)-\Gamma_h(\cdot),H-h+1\}^+$
  $\widehat{\pi}_h(\cdot|s)=\operatorname{argmax}_{\pi_h}\langle\widehat{Q}_h(s,\cdot),\pi_h(\cdot|s)\rangle$
  $\widehat{V}_h(\cdot)=\langle\widehat{Q}_h(\cdot,\cdot),\widehat{\pi}_h(\cdot|\cdot)\rangle_{\mathcal{A}}$
end for

We define the vector $\Theta=[\theta_1^\top,\cdots,\theta_H^\top]\in\mathbb{R}^{dH}$ and the loss function $L_r:\mathbb{R}^{dH}\to\mathbb{R}$ for reward learning as

$$
\begin{aligned}
L_r(\Theta)=\sum_{\tau\in\mathcal{D}}\left[\sum_{h=1}^{H}\langle\phi(x_h^\tau),\theta_h\rangle-r(\tau)\right]^2\\
+\lambda_1\cdot\sum_{h=1}^{H}\|\theta_h-\theta_0\|_2^2,
\end{aligned}
\tag{66}
$$

where $\lambda_1>0$ is a regularization parameter. We then define $\widehat{R}_h(\cdot)$ as

$$
\begin{aligned}
&\widehat{R}_h(\cdot)=\langle\phi(\cdot),\widehat{\theta}_h\rangle,\\
&\text{where}\quad\widehat{\Theta}=\operatorname*{argmin}_{\Theta\in\mathbb{R}^{2dmH}}L_r(\Theta)\\
&\text{and}\ \widehat{\Theta}=[\widehat{\theta}_1^\top,\cdots,\widehat{\theta}_H^\top]^\top.
\end{aligned}
\tag{67}
$$

Similarly, we also use linear function $\langle\phi(s,a),w_h\rangle$ to estimate transition value functions $\{(\mathbb{P}_h\widehat{V}_{h+1})(\cdot,\cdot)\}_{h\in[H]}$ for all $h\in[H]$, where $w_h\in\mathbb{R}^d$ is a learnable parameter. For each $h\in[H]$, we define the loss function $L_v^h(w_h)$: $\mathbb{R}^d\to\mathbb{R}$ as

$$
\begin{aligned}
L_v^h(w_h)=\sum_{\tau\in\mathcal{D}}\left(\widehat{V}_{h+1}(s_{h+1}^\tau)-\langle\phi(x_h^\tau),w_h\rangle\right)^2\\
+\lambda_2\cdot\|w_h-w_0\|_2^2,
\end{aligned}
\tag{68}
$$

where $\lambda_2>0$ is a regularization parameter. We then define $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot):\mathcal{X}\to\mathbb{R}$ as

$$
\begin{aligned}
&(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)=\langle\phi(\cdot),\widehat{w}_h\rangle,\\
&\text{where}\qquad\widehat{w}_h=\operatorname*{argmin}_{w_h\in\mathbb{R}^d}L_v^h(w_h).
\end{aligned}
\tag{69}
$$

It remains to construct the penalty term Γ_h. We first consider the penalty term that is used to offset the uncertainty raised from estimating the reward $R_h(\cdot)$ for each $h\in[H]$. We define the vectors $\Phi_h(x)=[\mathbf{0}_d^\top,\cdots,\phi(x)^\top,\cdots,\mathbf{0}_d^\top]^\top\in\mathbb{R}^{dH}$ and $\Phi(\tau)=[\phi(x_1^\tau),\cdots,\phi(x_H^\tau)]\in\mathbb{R}^{dH}$, where $\Phi_h(x)\in\mathbb{R}^{dH}$ is a vector in which $[\Phi_h(x)]_{d(h-1)+1:dh}=\phi(x)$ and the rest entries are all zero. We define a matrix $\Sigma(\Theta)\in\mathbb{R}^{dH\times dH}$ as

$$
\Sigma=\lambda_1\cdot I_{dH}+\sum_{\tau\in\mathcal{D}}\Phi(\tau)\Phi(\tau)^\top.
$$

The penalty term $b_{r,h}$ of the estimated reward is then defined as

$$
b_{r,h}(x)=\left[\Phi_h(x)^\top\Sigma^{-1}\Phi_h(x)\right]^{1/2}.
\tag{70}
$$

Next, we consider the penalty term that is used to offset the uncertainty raised from estimating the transition value function $(\mathbb{P}_h \widehat{V}_{h+1})(\cdot)$ for each $h \in [H]$. We define a matrix $\Lambda_h \in \mathbb{R}^{d\times d}$ as

$$\Lambda_h = \lambda_2 \cdot I_d + \sum_{\tau\in\mathcal{D}} \phi(x_h^\tau)\phi(x_h^\tau)^\top.$$

The penality term $b_{v,h}$ of the estimated transition value function is then defined as

$$b_{v,h}(x) = \left[\phi(x)^\top \Lambda_h^{-1} \phi(x)^\top\right]^{1/2}. \tag{71}$$

Finally, the penalty term for the estimated Bellman operation $\widehat{\mathbb{B}}_h \widehat{V}_{h+1}(\cdot)$ is obtained as

$$\Gamma_h(x) = \beta_1 b_{r,h}(x) + \beta_2 b_{v,h}(x), \tag{72}$$

where $\beta_1, \beta_2 > 0$ are constant factors.

B. Main Result

We consider the following dataset coverage assumption so that we can explicitly bound the suboptimality of Algorithm 3. Note that the following assumption has also been considered in [26].

Assumption 6 (Well-Explored Dataset): Suppose the N trajectories in dataset $\mathcal{D}$ are independent and identically induced by a fixed behaviour policy μ. There exist absolute constants $C_\sigma > 0$ and $C_\varsigma > 0$ such that $\forall h \in [H]$

$$\lambda_{\min}(\overline{M}(\Theta_0)) \geq C_\sigma, \lambda_{\min}(\overline{m}_h(w_0)) \geq C_\varsigma,$$

where

$$\overline{M} = \mathbb{E}_\mu\left[\Phi(\tau)\Phi(\tau)^\top\right],$$
$$\overline{m}_h(w_0) = \mathbb{E}_\mu\left[\phi(x_h^\tau)\phi(x_h^\tau)^\top\right].$$

We provide a formal statement of Theorem 1 as follows, which characterizes the suboptimality of Algorithm 3.

Theorem 5 (Formal Statement of Theorem 1): Consider Algorithm 3. Let $\lambda_1 = \lambda_2 = 1$ and $\beta_1 = \mathcal{O}(H\sqrt{dH\log(N/\delta)})$ and $\beta_2 = \mathcal{O}(dH^2\sqrt{\log(dH^3N^{5/2}/\delta)})$. Then, with probability at least $1-\delta$, we have

$$\text{SubOpt}(\widehat{\pi}, s) \leq \mathcal{O}\left(\frac{dH^3}{\sqrt{N}}\sqrt{\log\left(\frac{dH^3N^{5/2}}{\delta}\right)}\right).$$

C. Proof Flow of Theorem 1

In this section, we present the main proof flow of Theorem 1. Our main development is Lemma 3, the proof of which is presented in Section H.

Recalling the suboptimality of $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ in eq. (40), we have

$$\begin{aligned}&\text{SubOpt}(\widehat{\pi}, s)\\ &= -\sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}\left[\delta_h(s_h,a_h)\middle|s_1 = s\right]\\ &\quad + \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\delta_h(s_h,a_h)\middle|s_1 = s\right],\end{aligned}$$

where $\delta_h(\cdot)$ is the evaluation error defined as

$$\delta_h(s,a) = (\mathbb{B}_h\widehat{V}_{h+1})(s,a) - \widehat{Q}_h(s,a).$$

To characterize the suboptimality $\text{SubOpt}(\widehat{\pi}, s)$, we provide the following lemma to bound $\delta_h(\cdot)$ in the linear MDP setting.

Lemma 3: Let λ_1, $\lambda_2 = 1$, and let $\beta_1 = C_{\beta_1}H\sqrt{dH\log(N/\delta)}$ and $\beta_2 = C_{\beta_2}dH^2\sqrt{\log(dH^3N^{5/2}/\delta)}$, where C_{β_1}, C_{β_2} are two absolute constants. Suppose Assumbtion 3 holds. With probability at least $1-\delta/2$, it holds for all $h \in [H]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$ that

$$0 \leq \delta_h(x) \leq 2\left[\beta_1 \cdot b_{r,h}(x) + \beta_2 \cdot b_{v,h}(x)\right].$$

Proof: The main technical development here lies in handling additional challenges caused by the reward redistribution of trajectory-wise rewards, which are not present in linear MDPs with instantaneous rewards [26]. The detailed proof is provided in Section H. □

Applying Lemma 3 to eq. (40), we can obtain

$$\begin{aligned}&\text{SubOpt}(\widehat{\pi}, s)\\ &\leq 2\beta_1 \cdot \sum_{h=1}^H \max_x b_{r,h}(x)\\ &\quad + 2\beta_2 \cdot \sum_{h=1}^H \max_x b_{v,h}(x).\end{aligned} \tag{73}$$

Then, following steps similar to those in Section G, we have the followings hold with probability at least $1-\delta/2$

$$b_{r,h}(x) \leq \frac{C'}{\sqrt{N}} \quad \text{and} \quad b_{v,h}(x) \leq \frac{C''}{\sqrt{N}}, \tag{74}$$

where C' and C'' are absolute constants dependent only on C_σ, C_ς and $\log(1/\delta)$. Then, substituting eq. (74) into eq. (73), we have the following holds with probability $1-\delta$

$$\begin{aligned}&\text{SubOpt}(\widehat{\pi}, s)\\ &\leq 2C_{\beta_1}H^2\sqrt{dH\log(N/\delta)}\cdot\frac{C'}{\sqrt{N}}\\ &\quad + 2C_{\beta_2}dH^3\sqrt{\log(dH^3N^{5/2}/\delta)}\cdot\frac{C''}{\sqrt{N}}\\ &\leq \mathcal{O}\left(\frac{dH^3}{\sqrt{N}}\sqrt{\log\left(\frac{dH^3N^{5/2}}{\delta}\right)}\right),\end{aligned}$$

which completes the proof.

Appendix F
Proof of Lemma 1

Recall that we let (b_0, w_0) be the initial value of network parameters obtained via the symmetric initialization scheme, which makes $f(\cdot; w_0)$ a zero function. We denote $(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\cdot) = \widehat{R}_h(\cdot) + (\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ as the estimator of Bellman operator $(\mathbb{B}_h\widehat{V}_{h+1})(\cdot) = R_h(\cdot) + (\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$. To prove Lemma 1, we show that $(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\cdot) - \beta_1 b_{r,h}(\cdot, \widehat{\Theta}) - \beta_2 b_{v,h}(\cdot, \widehat{w})$ is approximately a pessimistic estimator of $(\mathbb{B}_h\widehat{V}_{h+1})(\cdot)$ up to a function approximation error. We consider m to be sufficiently large such that $m \geq NH^2$.

A. Uncertainty of Estimated Reward $\widehat{R}_h(\cdot)$

In this step, we aim to bound the estimation error $\left|\widehat{R}_h(\cdot) - R_h(\cdot)\right|$. Since $\widehat{\Theta}$ is the global minimizer of the loss function L_r defined in eq. (30), we have

$$\begin{aligned} L_r(\widehat{\Theta}) &= \sum_{\tau\in\mathcal{D}}\left[\sum_{h=1}^{H} f(x_h^\tau, \widehat{\theta}_h) - r(\tau)\right]^2 \\ &\quad + \lambda_1 \cdot \sum_{h=1}^{H}\left\|\widehat{\theta}_h - \theta_0\right\|_2^2 \\ &\le L_r(\Theta_0) \\ &= \sum_{\tau\in\mathcal{D}}\left[\sum_{h=1}^{H} f(x_h^\tau, \theta_0) - r(\tau)\right]^2 \\ &\overset{(i)}{=} \sum_{\tau\in\mathcal{D}}[r(\tau)]^2 \\ &\overset{(ii)}{\le} NH^2, \end{aligned} \tag{75}$$

where (i) follows from the fact that $f(x,\theta_0) = 0$ for all $x \in \mathcal{X}$ and (ii) follows from the fact that $r(\tau) \le H$ for any trajectory τ and we have total N trajectories in the offline sample set $\mathcal{D}$. We define the vector $\Theta_0 = [\theta_0^\top, \cdots, \theta_0^\top]^\top \in \mathbb{R}^{2mdH}$. Note that eq. (75) implies

$$\begin{aligned} &\left\|\widehat{\theta}_h - \theta_0\right\|_2^2 \\ &\le \left\|\widehat{\Theta} - \Theta_0\right\|_2^2 \\ &= \sum_{h=1}^{H}\left\|\widehat{\theta}_h - \theta_0\right\|_2^2 \\ &\le NH^2/\lambda_1, \quad \forall h \in [H]. \end{aligned} \tag{76}$$

Hence, each $\widehat{\theta}_h$ belongs to the Euclidean ball $\mathcal{B}_\theta = \{\theta \in \mathbb{R}^{2md} : \|\theta - \theta_0\|_2 \le H\sqrt{N/\lambda_1}\}$.

Since the radius of $\mathcal{B}_\theta$ does not depend on m, when m is sufficient large it can be shown that $f(\cdot,\theta)$ is close to its linearization at θ_0, i.e.,

$$f(\cdot,\theta) \approx \langle \phi(\cdot,\theta_0), \theta - \theta_0\rangle, \quad \forall \theta \in \mathcal{B}_\theta,$$

where $\phi(\cdot,\theta) = \nabla_\theta f(\cdot,\theta)$. Furthermore, according to Assumbtion 3, there exists a function $\ell_{a_1,a_2} : \mathbb{R}^d \to \mathbb{R}^d$ such that the mean of the true reward function $R_h(\cdot) = \mathbb{E}[r_h(\cdot)]$ satisfies

$$R_h(x) = \int_{\mathbb{R}^d} \sigma'(\theta^\top x)\cdot x^\top \ell_r(\theta) dp(\theta), \tag{77}$$

where $\sup_\theta \|\ell_r(\theta)\|_2 \le a_1$, $\sup_\theta(\|\ell_r(\theta)\|_2 / p(\theta)) \le a_2$ and p is the density of the distribution $N(0, I_d/d)$. We then proceed to bound the difference between $\widehat{R}_h(\cdot)$ and $R_h(\cdot)$.

Step I: In the first step, we show that with high probability the mean of the true reward $R_h(\cdot)$ can be well-approximated by a linear function with the feature vector $\phi(\cdot,\theta_0)$. Lemma 4 in Section I implies that that $R_h(\cdot)$ in eq. (77) can be well-approximated by a finite-width neural network, i.e., with probability at least $1 - N^{-2}H^{-4}$ over the randomness of initialization θ_0, for all $h \in [H]$, there exists a function $\widetilde{R}_h(\cdot) : \mathcal{X} \to \mathbb{R}$ satisfying

$$\begin{aligned} &\sup_{x\in\mathcal{X}}\left|\widetilde{R}_h(x) - R_h(x)\right| \\ &\le \frac{2(L_\sigma a_2 + C_\sigma^2 a_2^2)\sqrt{\log(N^2H^5)}}{\sqrt{m}}, \end{aligned} \tag{78}$$

where $\widetilde{R}_h(\cdot)$ can be written as

$$\widetilde{R}_h(x) = \frac{1}{\sqrt{m}}\sum_{r=1}^{m}\sigma'(\theta_{0,r}^\top x)\cdot x^\top \ell_r,$$

where $\|\ell_r\|_2 \le a_2/\sqrt{dm}$ for all $r \in [m]$ and $\theta_0 = [\theta_{0,1}, \cdots, \theta_{0,m}]$ is generated via the symmetric initialization scheme. We next proceed to show that there exists a vector $\tilde{\theta}_h \in \mathbb{R}^{2md}$ such that $\widetilde{R}_h(\cdot) = \langle \phi(\cdot,\theta_0), \tilde{\theta}_h - \theta_0\rangle$. Let $\tilde{\theta}_h = [\tilde{\theta}_{h,1}^\top, \cdots .\tilde{\theta}_{h,2m}^\top]^\top$, in which $\tilde{\theta}_{h,r}^\top = \theta_{0,r} + b_{0,r}\cdot \ell_r/\sqrt{2}$ for all $r \in \{1,\cdots,m\}$ and $\tilde{\theta}_{h,r}^\top = \theta_{0,r} + b_{0,r}\cdot \ell_{r-m}/\sqrt{2}$ for all $r \in \{m+1,\cdots,2m\}$. Then, we have

$$\begin{aligned} &\widetilde{R}_h(x) \\ &= \frac{1}{\sqrt{2m}}\sum_{r=1}^{m}\sqrt{2}(b_{0,r})^2\cdot\sigma'(\theta_{0,r}^\top x)\cdot x^\top \ell_r \\ &= \frac{1}{\sqrt{2m}}\sum_{r=1}^{m}\frac{1}{\sqrt{2}}(b_{0,r})^2\cdot\sigma'(\theta_{0,r}^\top x)\cdot x^\top \ell_r \\ &\quad + \frac{1}{\sqrt{2m}}\sum_{r=1}^{m}\frac{1}{\sqrt{2}}(b_{0,r})^2\cdot\sigma'(\theta_{0,r}^\top x)\cdot x^\top \ell_{r-m} \\ &= \frac{1}{\sqrt{2m}}\sum_{r=1}^{2m} b_{0,r}\cdot\sigma'(\theta_{0,r}^\top x)\cdot x^\top(\tilde{\theta}_{h,r} - \theta_{0,r}) \\ &= \phi(x,\theta_0)^\top(\tilde{\theta}_h - \theta_0). \end{aligned} \tag{79}$$

Thus, the true mean reward $R_h(\cdot)$ is approximately linear with the feature $\phi(\cdot,\theta_0)$. Since $\tilde{\theta}_{h,r} - \theta_{0,r} = b_{0,r}\cdot\ell_r/\sqrt{2}$ or $b_{0,r}\cdot\ell_{r-m}/\sqrt{2}$, we have

$$\left\|\tilde{\theta}_h - \theta_0\right\|_2 \le a_2\sqrt{2dm}.$$

Step II: In this step, we show that $\widehat{R}_h(\cdot)$ learned by neural network in Algorithm 2 can be well-approximated by its counterpart learned by a linear function with feature $\phi(\cdot,\theta_0)$.

Consider the following least-square loss function

$$\begin{aligned} &\bar{L}_r(\Theta) \\ &= \sum_{\tau\in\mathcal{D}}\left[\sum_{h=1}^{H}\langle\phi(x_h^\tau,\theta_0), \theta_h - \theta_0\rangle - r(\tau)\right]^2 \\ &\quad + \lambda_1\cdot\sum_{h=1}^{H}\|\theta_h - \theta_0\|_2^2 \\ &= \sum_{\tau\in\mathcal{D}}\left[\langle\Phi(\tau,\Theta_0), \Theta - \Theta_0\rangle - r(\tau)\right]^2 \\ &\quad + \lambda_1\cdot\|\Theta - \Theta_0\|_2^2. \end{aligned}$$

The global minimizer of $\bar{L}_r(\Theta)$ is defined as

$$\begin{aligned} \overline{\Theta} &= \operatorname*{argmin}_{\Theta\in\mathbb{R}^{2dmH}} \bar{L}_r(\theta), \\ \overline{\Theta} &= [\bar{\theta}_1^\top, \cdots, \bar{\theta}_H^\top]^\top. \end{aligned} \tag{80}$$

We define $\overline{R}_h(\cdot) = \langle \phi(\cdot, \theta_0), \bar{\theta}_h - \theta_0 \rangle$ for all $h \in [H]$. We then proceed to bound the term $\left| \widehat{R}_h(x) - \overline{R}_h(x) \right|$ as follows

$$\begin{aligned}
&\left| \widehat{R}_h(x) - \overline{R}_h(x) \right| \\
&= \left| f(x, \widehat{\theta}_h) - \langle \phi(x, \theta_0), \bar{\theta}_h - \theta_0 \rangle \right| \\
&= \left| f(x, \widehat{\theta}_h) - \langle \Phi_h(x, \Theta_0), \overline{\Theta} - \Theta_0 \rangle \right| \\
&= \Big| f(x, \widehat{\theta}_h) - \langle \Phi_h(x, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \\
&\quad + \langle \Phi_h(x, \Theta_0), \widehat{\Theta} - \overline{\Theta} \rangle \Big| \\
&\le \left| f(x, \widehat{\theta}_h) - \langle \Phi_h(x, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \right| \\
&\quad + \left| \langle \Phi_h(x, \Theta_0), \widehat{\Theta} - \overline{\Theta} \rangle \right| \\
&= \left| f(x, \widehat{\theta}_h) - \langle \phi_h(x, \theta_0), \widehat{\theta}_h - \theta_0 \rangle \right| \\
&\quad + \left| \langle \Phi_h(x, \Theta_0), \widehat{\Theta} - \overline{\Theta} \rangle \right| \\
&\le \left| f(x, \widehat{\theta}_h) - \langle \phi_h(x, \theta_0), \widehat{\theta}_h - \theta_0 \rangle \right| \\
&\quad + \| \Phi_h(x, \Theta_0) \|_2 \left\| \widehat{\Theta} - \overline{\Theta} \right\|_2 \\
&= \underbrace{\left| f(x, \widehat{\theta}_h) - \langle \phi_h(x, \theta_0), \widehat{\theta}_h - \theta_0 \rangle \right|}_{(i)} \\
&\quad + \underbrace{\| \phi(x, \theta_0) \|_2 \left\| \widehat{\Theta} - \overline{\Theta} \right\|_2}_{(ii)}.
\end{aligned}$$

According to Lemma 5 and the fact that $\left\| \widehat{\theta}_h - \theta_0 \right\|_2 \le H\sqrt{N/\lambda_1}$, we have the followings hold with probability at least $1 - N^{-2}H^{-4}$

$$(i) \le \mathcal{O}\left(C_\phi \left(\frac{N^2 H^4}{\lambda_1^2 \sqrt{m}} \right)^{1/3} \sqrt{\log m} \right), \tag{81}$$

$$(ii) \le C_\phi \left\| \widehat{\Theta} - \overline{\Theta} \right\|_2. \tag{82}$$

We then proceed to bound the term $\left\| \widehat{\Theta} - \overline{\Theta} \right\|_2$. Consider the minimization problem defined in eq. (31) and eq. (80). By the first order optimality condition, we have

$$\begin{aligned}
&\lambda_1 \left(\widehat{\Theta} - \Theta_0 \right) \\
&= \sum_{\tau \in \mathcal{D}} \left(r(\tau) - \sum_{h=1}^{H} f(x_h^\tau, \hat{\theta}_h) \right) \Phi(\tau, \widehat{\Theta})
\end{aligned} \tag{83}$$

$$\begin{aligned}
&\lambda_1 \left(\overline{\Theta} - \Theta_0 \right) = \sum_{\tau \in \mathcal{D}} \Phi(\tau, \Theta_0) \\
&\quad \cdot \left(r(\tau) - \langle \Phi(\tau, \Theta_0), \overline{\Theta} - \Theta_0 \rangle \right).
\end{aligned} \tag{84}$$

Note that eq. (84) implies

$$\Sigma(\Theta_0) \left(\overline{\Theta} - \Theta_0 \right) = \sum_{\tau \in \mathcal{D}} r(\tau) \Phi(\tau, \Theta_0). \tag{85}$$

Adding the term $\sum_{\tau \in \mathcal{D}} \langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \Phi(\tau, \Theta_0)$ on both sides of eq. (83) yields

$$\begin{aligned}
&\Sigma(\Theta_0) \left(\widehat{\Theta} - \Theta_0 \right) \\
&= \sum_{\tau \in \mathcal{D}} r(\tau) \Phi(\tau, \widehat{\Theta}) \\
&\quad + \sum_{\tau \in \mathcal{D}} \Big[\langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \Phi(\tau, \Theta_0) \\
&\quad - \left(\sum_{h=1}^{H} f(x_h^\tau, \hat{\theta}_h) \right) \Phi(\tau, \widehat{\Theta}) \Big].
\end{aligned} \tag{86}$$

Then, subtracting eq. (85) from eq. (86), we can obtain

$$\begin{aligned}
&\Sigma(\Theta_0)(\widehat{\Theta} - \overline{\Theta}) \\
&= \sum_{\tau \in \mathcal{D}} r(\tau) \left(\Phi(\tau, \widehat{\Theta}) - \Phi(\tau, \Theta_0) \right) \\
&\quad + \sum_{\tau \in \mathcal{D}} \Big[\langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \Phi(\tau, \Theta_0) \\
&\quad - \left(\sum_{h=1}^{H} f(x_h^\tau, \hat{\theta}_h) \right) \Phi(\tau, \widehat{\Theta}) \Big],
\end{aligned} \tag{87}$$

which implies

$$\begin{aligned}
&\left\| \Sigma(\Theta_0)(\widehat{\Theta} - \overline{\Theta}) \right\|_2 \\
&\le \sum_{\tau \in \mathcal{D}} r(\tau) \sqrt{\sum_{h=1}^{H} \left\| \phi(x_h^\tau, \theta_0) - \phi(x_h^\tau, \widehat{\theta}_h) \right\|_2^2} \\
&\quad + \sum_{\tau \in \mathcal{D}} \Big\| \langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \Phi(\tau, \Theta_0) \\
&\quad - \left(\sum_{h=1}^{H} f(x_h^\tau, \hat{\theta}_h) \right) \Phi(\tau, \widehat{\Theta}) \Big\|_2.
\end{aligned} \tag{88}$$

To bound the term $\| \langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \Phi(\tau, \Theta_0) - \left(\sum_{h=1}^{H} f(x_h^\tau, \hat{\theta}_h) \right) \Phi(\tau, \widehat{\Theta}) \|_2$, we proceed as follows

$$\begin{aligned}
&\langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \Phi(\tau, \Theta_0) \\
&\quad - \left(\sum_{h=1}^{H} f(x_h^\tau, \hat{\theta}_h) \right) \Phi(\tau, \widehat{\Theta}) \\
&= \langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle (\Phi(\tau, \Theta_0) - \Phi(\tau, \widehat{\Theta})) \\
&\quad - \left(\langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle - \sum_{h=1}^{H} f(x_h^\tau, \hat{\theta}_h) \right) \Phi(\tau, \widehat{\Theta}) \\
&= \langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle (\Phi(\tau, \Theta_0) - \Phi(\tau, \widehat{\Theta})) \\
&\quad - \left[\sum_{h=1}^{H} \left(\langle \phi(x_h^\tau, \theta_0), \widehat{\theta}_h - \theta_h^0 \rangle - f(x_h^\tau, \hat{\theta}_h) \right) \right] \Phi(\tau, \widehat{\Theta}),
\end{aligned}$$

which implies

$$\begin{aligned}
&\Big\| \langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \Phi(\tau, \Theta_0) \\
&\quad - \left(\sum_{h=1}^{H} f(x_h^\tau, \hat{\theta}_h) \right) \Phi(\tau, \widehat{\Theta}) \Big\|_2 \\
&\le \| \Phi(\tau, \Theta_0) \|_2 \left\| \widehat{\Theta} - \Theta_0 \right\|_2 \left\| \Phi(\tau, \Theta_0) - \Phi(\tau, \widehat{\Theta}) \right\|_2 \\
&\quad + \left[\sum_{h=1}^{H} \left| \langle \phi(x_h^\tau, \theta_0), \widehat{\theta}_h - \theta_h^0 \rangle - f(x_h^\tau, \hat{\theta}_h) \right| \right] \left\| \Phi(\tau, \widehat{\Theta}) \right\|_2
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\sum_{h=1}^{H} \|\phi(x_h^\tau, \theta_h^0)\|_2^2} \left\|\widehat{\Theta} - \Theta_0\right\|_2 \\
&\quad \cdot \sqrt{\sum_{h=1}^{H} \left\|\phi(x_h^\tau, \theta_0) - \phi(x_h^\tau, \widehat{\theta}_h)\right\|_2^2} \\
&\quad + \left[\sum_{h=1}^{H} \left|\langle \phi(x_h^\tau, \theta_0), \widehat{\theta}_h - \theta_h^0 \rangle - f(x_h^\tau, \hat{\theta}_h)\right|\right] \\
&\quad \cdot \sqrt{\sum_{h=1}^{H} \left\|\phi(x_h^\tau, \hat{\theta}_h)\right\|_2^2}, \qquad (89)
\end{aligned}
$$

where the last equality follows from the fact that $\|\Phi(\tau, \Theta)\|_2^2 = \sum_{h=1}^{H} \|\phi(x_h^\tau, \theta_h)\|_2^2$ for any $\Theta \in \mathbb{R}^{2mdH}$. According to Lemma 5 and the fact that $\left\|\widehat{\theta}_h - \theta_0\right\|_2 \le H\sqrt{N/\lambda_1}$, we have the followings hold with probability at least $1 - N^{-2}H^{-4}$ for all $h \in [H]$ and $\tau \in \mathcal{D}$

$$
\begin{aligned}
&\|\phi(x_h^\tau, \theta_0)\|_2 \le C_\phi, \\
&\left\|\phi(x_h^\tau, \hat{\theta}_h)\right\|_2 \le C_\phi, \qquad (90) \\
&\left\|\phi(x_h^\tau, \theta_0) - \phi(x_h^\tau, \widehat{\theta}_h)\right\|_2 \\
&\quad \le \mathcal{O}\left(C_\phi \left(\frac{H\sqrt{N/\lambda_1}}{\sqrt{m}}\right)^{1/3} \sqrt{\log m}\right), \qquad (91) \\
&\left|\langle \phi(x_h^\tau, \theta_0), \widehat{\theta}_h - \theta_0 \rangle - f(x_h^\tau, \hat{\theta}_h)\right| \\
&\quad \le \mathcal{O}\left(C_\phi \left(\frac{H^4 N^2/\lambda_1^2}{\sqrt{m}}\right)^{1/3} \sqrt{\log m}\right). \qquad (92)
\end{aligned}
$$

Substituting eq. (90), eq. (91) and eq. (92) into eq. (89), we have

$$
\begin{aligned}
&\left\|\langle \Phi(\tau, \Theta_0), \widehat{\Theta} - \Theta_0 \rangle \Phi(\tau, \Theta_0) \right. \\
&\quad \left. - \left(\sum_{h=1}^{H} f(x_h^\tau, \hat{\theta}_h)\right) \Phi(\tau, \widehat{\Theta})\right\| \\
&\le (H^2\sqrt{N/\lambda_1}) \\
&\quad \cdot \mathcal{O}\left(C_\phi^2 \left(\frac{H\sqrt{N/\lambda_1}}{\sqrt{m}}\right)^{1/3} \sqrt{\log m}\right) \\
&\quad + \mathcal{O}\left(C_\phi^2 H^{3/2} \left(\frac{H^4 N^2/\lambda_1^2}{\sqrt{m}}\right)^{1/3} \sqrt{\log m}\right) \\
&\le \mathcal{O}\left(\frac{C_\phi^2 H^{17/6} N^{2/3} \sqrt{\log(m)}}{m^{1/6} \lambda_1^{2/3}}\right). \qquad (93)
\end{aligned}
$$

Then, substituting eq. (93) into eq. (88), we have the following holds with probability at least $1 - N^{-2}H^{-4}$

$$
\begin{aligned}
&\left\|\Sigma(\Theta_0)(\widehat{\Theta} - \overline{\Theta})\right\|_2 \\
&\le NH \cdot \sqrt{H} \cdot \mathcal{O}\left(C_\phi \left(\frac{H\sqrt{N/\lambda_1}}{\sqrt{m}}\right)^{1/3} \sqrt{\log m}\right) \\
&\quad + N \cdot \mathcal{O}\left(\frac{C_\phi^2 H^{17/6} N^{2/3} \sqrt{\log(m)}}{m^{1/6} \lambda_1^{2/3}}\right) \\
&\le \mathcal{O}\left(\frac{C_\phi^2 H^{17/6} N^{5/3} \sqrt{\log(m)}}{m^{1/6} \lambda_1^{2/3}}\right),
\end{aligned}
$$

where we use the fact that $r(\tau) \le H$. We then proceed to bound the term $\left\|\widehat{\Theta} - \Theta_0\right\|_2$ as follows

$$
\begin{aligned}
&\left\|\widehat{\Theta} - \overline{\Theta}\right\|_2 \\
&= \left\|\Sigma^{-1}(\Theta_0)\Sigma(\Theta_0)(\widehat{\Theta} - \overline{\Theta})\right\|_2 \\
&\le \left\|\Sigma^{-1}(\Theta_0)\right\|_2 \left\|\Sigma(\Theta_0)(\widehat{\Theta} - \overline{\Theta})\right\|_2 \\
&\le \lambda_1^{-1} \left\|\Sigma(\Theta_0)(\widehat{\Theta} - \overline{\Theta})\right\|_2 \\
&\le \mathcal{O}\left(\frac{C_\phi^2 H^{17/6} N^{5/3} \sqrt{\log(m)}}{m^{1/6} \lambda_1^{5/3}}\right). \qquad (94)
\end{aligned}
$$

Substituting eq. (94) into eq. (82), we can bound (ii) as follows

$$
(ii) \le \mathcal{O}\left(\frac{C_\phi^3 H^{17/6} N^{5/3} \sqrt{\log(m)}}{m^{1/6} \lambda_1^{5/3}}\right). \qquad (95)
$$

Taking summation of the upper bounds of (i) in eq. (81) and (ii) in eq. (95), we have

$$
\begin{aligned}
&\left|\widehat{R}_h(x) - \overline{R}_h(x)\right| \\
&\le (i) + (ii) \\
&\le \mathcal{O}\left(C_\phi \left(\frac{N^2 H^4}{\lambda_1^2 \sqrt{m}}\right)^{1/3} \sqrt{\log m}\right) \\
&\quad + \mathcal{O}\left(\frac{C_\phi^3 H^{17/6} N^{5/3} \sqrt{\log(m)}}{m^{1/6} \lambda_1^{5/3}}\right) \\
&\le \mathcal{O}\left(\frac{C_\phi^3 H^{17/6} N^{5/3} \sqrt{\log(m)}}{m^{1/6} \lambda_1^{5/3}}\right). \qquad (96)
\end{aligned}
$$

Step III: In this step, we show that the bonus term $b_{r,h}(\cdot, \widehat{\Theta})$ in Algorithm 2 can be well approximated by $b_{r,h}(\cdot, \Theta_0)$. According to the definition of $b_{r,h}(\cdot, \Theta)$, we have

$$
\begin{aligned}
&\left|b_{r,h}(x, \widehat{\Theta}) - b_{r,h}(x, \Theta_0)\right| \\
&= \left| \left[\Phi_h(x, \widehat{\Theta})^\top \Sigma^{-1}(\widehat{\Theta}) \Phi_h(x, \widehat{\Theta})\right]^{1/2} \right. \\
&\quad \left. - \left[\Phi_h(x, \Theta_0)^\top \Sigma^{-1}(\Theta_0) \Phi_h(x, \Theta_0)\right]^{1/2} \right| \\
&\le \left| \Phi_h(x, \Theta_0)^\top \Sigma^{-1}(\Theta_0) \Phi_h(x, \Theta_0) \right. \\
&\quad \left. - \Phi_h(x, \widehat{\Theta})^\top \Sigma^{-1}(\widehat{\Theta}) \Phi_h(x, \widehat{\Theta}) \right|^{1/2}, \qquad (97)
\end{aligned}
$$

where the last inequality follows from the fact that $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x - y|}$. We then proceed to bound the term $|\Phi_h(x, \widehat{\Theta})^\top \Sigma^{-1}(\widehat{\Theta}) \Phi_h(x, \widehat{\Theta}) - \Phi_h(x, \Theta_0)^\top \Sigma^{-1}(\Theta_0) \Phi_h(x, \Theta_0)|$ as follows

$$
\left|\Phi_h(x, \widehat{\Theta})^\top \Sigma^{-1}(\widehat{\Theta}) \Phi_h(x, \widehat{\Theta})\right.
$$

$$
\begin{aligned}
&- \Phi_h(x,\Theta_0)^\top \Sigma^{-1}(\Theta_0)\Phi_h(x,\Theta_0)\big| \\
&= \big|[\Phi_h(x,\widehat{\Theta}) - \Phi_h(x,\Theta_0)]^\top \Sigma^{-1}(\widehat{\Theta})\Phi_h(x,\widehat{\Theta})\big| \\
&\quad + \Big|\Phi_h(x,\Theta_0)^\top(\Sigma^{-1}(\widehat{\Theta}) - \Sigma^{-1}(\Theta_0)) \\
&\quad \cdot \Phi_h(x,\widehat{\Theta})\Big| \\
&\quad + \big|\Phi_h(x,\Theta_0)^\top \Sigma^{-1}(\Theta_0) \\
&\quad \cdot (\Phi_h(x,\widehat{\Theta}) - \Phi_h(x,\Theta_0))\big| \\
&\le [\Phi_h(x,\widehat{\Theta}) - \Phi_h(x,\Theta_0)]^\top \Sigma^{-1}(\widehat{\Theta})\Phi_h(x,\widehat{\Theta}) \\
&\quad + \Big|\Phi_h(x,\Theta_0)^\top(\Sigma^{-1}(\widehat{\Theta}) - \Sigma^{-1}(\Theta_0))\Phi_h(x,\widehat{\Theta})\Big| \\
&\quad + \big|\Phi_h(x,\Theta_0)^\top \Sigma^{-1}(\Theta_0) \\
&\quad \cdot (\Phi_h(x,\widehat{\Theta}) - \Phi_h(x,\Theta_0))\big| \\
&\le \Big\|\Phi_h(x,\widehat{\Theta}) - \Phi_h(x,\Theta_0)\Big\|_2 \\
&\quad \cdot \Big\|\Sigma^{-1}(\widehat{\Theta})\Big\|_2 \Big\|\Phi_h(x,\widehat{\Theta})\Big\|_2 \\
&\quad + \|\Phi_h(x,\Theta_0)\|_2 \\
&\quad \cdot \Big\|\Sigma^{-1}(\widehat{\Theta}) - \Sigma^{-1}(\Theta_0)\Big\|_2 \Big\|\Phi_h(x,\widehat{\Theta})\Big\|_2 \\
&\quad + \|\Phi_h(x,\Theta_0)\|_2 \big\|\Sigma^{-1}(\Theta_0)\big\|_2 \\
&\quad \cdot \Big\|\Phi_h(x,\widehat{\Theta}) - \Phi_h(x,\Theta_0)\Big\|_2 \\
&= \Big\|\phi(x,\widehat{\theta}_h) - \phi(x,\theta_0)\Big\|_2 \Big\|\Sigma^{-1}(\widehat{\Theta})\Big\|_2 \Big\|\phi(x,\widehat{\theta}_h)\Big\|_2 \\
&\quad + \|\phi(x,\theta_0)\|_2 \Big\|\Sigma^{-1}(\widehat{\Theta})(\Sigma(\widehat{\Theta}) - \Sigma(\Theta_0))\Sigma^{-1}(\Theta_0)\Big\|_2 \\
&\quad \cdot \Big\|\phi(x,\widehat{\theta}_h)\Big\|_2 \\
&\quad + \|\phi(x,\theta_0)\|_2 \big\|\Sigma^{-1}(\Theta_0)\big\|_2 \Big\|\phi(x,\widehat{\theta}_h) - \phi(x,\theta_0)\Big\|_2 \\
&\le \Big\|\phi(x,\widehat{\theta}_h) - \phi(x,\theta_0)\Big\|_2 \Big\|\Sigma^{-1}(\widehat{\Theta})\Big\|_2 \Big\|\phi(x,\widehat{\theta}_h)\Big\|_2 \\
&\quad + \|\phi(x,\theta_0)\|_2 \Big\|\Sigma^{-1}(\widehat{\Theta})\Big\|_2 \Big\|\Sigma(\widehat{\Theta}) - \Sigma(\Theta_0)\Big\|_2 \\
&\quad \cdot \big\|\Sigma^{-1}(\Theta_0)\big\|_2 \Big\|\phi(x,\widehat{\theta}_h)\Big\|_2 \\
&\quad + \|\phi(x,\theta_0)\|_2 \big\|\Sigma^{-1}(\Theta_0)\big\|_2 \Big\|\phi(x,\widehat{\theta}_h) - \phi(x,\theta_0)\Big\|_2 \\
&\le \frac{1}{\lambda_1}\Big\|\phi(x,\widehat{\theta}_h) - \phi(x,\theta_0)\Big\|_2 \Big\|\phi(x,\widehat{\theta}_h)\Big\|_2 \\
&\quad + \frac{1}{\lambda_1^2}\|\phi(x,\theta_0)\|_2 \Big\|\Sigma(\widehat{\Theta}) - \Sigma(\Theta_0)\Big\|_2 \Big\|\phi(x,\widehat{\theta}_h)\Big\|_2 \\
&\quad + \frac{1}{\lambda_1}\|\phi(x,\theta_0)\|_2 \Big\|\phi(x,\widehat{\theta}_h) - \phi(x,\theta_0)\Big\|_2, \qquad (98)
\end{aligned}
$$

where the last inequality follows from the fact that $\|\Sigma(\Theta)\|_2 \ge \lambda_1$ for any $\Theta \in \mathbb{R}^{2mdH}$. For $\Sigma(\widehat{\Theta}) - \Sigma(\Theta_0)$, we have

$$
\begin{aligned}
&\Sigma(\widehat{\Theta}) - \Sigma(\Theta_0) \\
&= \sum_{\tau\in\mathcal{D}}\Big[\Phi(\tau,\widehat{\Theta})\Phi(\tau,\widehat{\Theta})^\top - \Phi(\tau,\Theta_0)\Phi(\tau,\Theta_0)^\top\Big] \\
&= \sum_{\tau\in\mathcal{D}}\big[\Phi(\tau,\widehat{\Theta})(\Phi(\tau,\widehat{\Theta}) - \Phi(\tau,\Theta_0))^\top \\
&\quad + (\Phi(\tau,\widehat{\Theta}) - \Phi(\tau,\Theta_0))\Phi(\tau,\Theta_0)^\top\big],
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\Big\|\Sigma(\widehat{\Theta}) - \Sigma(\Theta_0)\Big\|_2 \\
&\le \sum_{\tau\in\mathcal{D}}\big[\Big\|\Phi(\tau,\widehat{\Theta})\Big\|_2 \Big\|\Phi(\tau,\widehat{\Theta}) - \Phi(\tau,\Theta_0)\Big\|_2 \\
&\quad + \Big\|\Phi(\tau,\widehat{\Theta}) - \Phi(\tau,\Theta_0)\Big\|_2 \|\Phi(\tau,\Theta_0)\|_2\big]. \qquad (99)
\end{aligned}
$$

By definition of $\Phi(\tau,\Theta)$, we have the followings hold for any $\Theta, \widetilde{\Theta} \in \mathbb{R}^{2mdH}$

$$
\begin{aligned}
&\Big\|\Phi(\tau,\widehat{\Theta})\Big\|_2 = \sqrt{\sum_{h\in[H]}\Big\|\phi(x_h^\tau,\widehat{\theta}_h)\Big\|_2^2}, \qquad (100) \\
&\Big\|\Phi(\tau,\widehat{\Theta}) - \Phi(\tau,\Theta_0)\Big\|_2 \\
&= \sqrt{\sum_{h\in[H]}\Big\|\phi(x_h^\tau,\theta_h) - \phi(x_h^\tau,\tilde{\theta}_h)\Big\|_2^2}. \qquad (101)
\end{aligned}
$$

Applying Lemma 5 to eq. (100) and eq. (101), we have the followings hold with probability at least $1 - N^{-2}H^{-4}$

$$
\begin{aligned}
&\Big\|\Phi(\tau,\widehat{\Theta})\Big\|_2 \le C_\phi\sqrt{H}, \\
&\Big\|\Phi(\tau,\widehat{\Theta}) - \Phi(\tau,\Theta_0)\Big\|_2 \\
&\le \mathcal{O}\left(\frac{C_\phi H^{5/6}N^{1/6}\sqrt{\log m}}{m^{1/6}\lambda_1^{1/6}}\right).
\end{aligned}
$$

Substituting the above two inequalities into eq. (99) yields

$$
\begin{aligned}
&\Big\|\Sigma(\widehat{\Theta}) - \Sigma(\Theta_0)\Big\|_2 \\
&\le \mathcal{O}\left(\frac{C_\phi^2 H^{4/3}N^{1/6}\sqrt{\log m}}{m^{1/6}\lambda_1^{1/6}}\right). \qquad (102)
\end{aligned}
$$

Finally, combining eq. (102) and eq. (214) and eq. (215) in Lemma 5, we can bound the right hand side of eq. (98) as

$$
\begin{aligned}
&\big|\Phi_h(x,\widehat{\Theta})^\top \Sigma^{-1}(\widehat{\Theta})\Phi_h(x,\widehat{\Theta}) \\
&\quad - \Phi_h(x,\overline{\Theta})^\top \Sigma^{-1}(\overline{\Theta})\Phi_h(x,\overline{\Theta})\big| \\
&\le \frac{1}{\lambda_1}\Big\|\phi(x,\widehat{\theta}_h) - \phi(x,\theta_0)\Big\|_2 \Big\|\phi(x,\widehat{\theta}_h)\Big\|_2 \\
&\quad + \frac{1}{\lambda_1^2}\|\phi(x,\theta_0)\|_2 \Big\|\Sigma(\widehat{\Theta}) - \Sigma(\Theta_0)\Big\|_2 \\
&\quad \cdot \Big\|\phi(x,\widehat{\theta}_h)\Big\|_2 \\
&\quad + \frac{1}{\lambda_1}\|\phi(x,\theta_0)\|_2 \Big\|\phi(x,\widehat{\theta}_h) - \phi(x,\theta_0)\Big\|_2 \\
&\le \mathcal{O}\left(\frac{C_\phi^2 H^{1/3}N^{1/6}\sqrt{\log m}}{m^{1/6}\lambda_1^{7/6}}\right) \\
&\quad + \mathcal{O}\left(\frac{C_\phi^4 H^{4/3}N^{1/6}\sqrt{\log m}}{m^{1/6}\lambda_1^{13/6}}\right) \\
&= \mathcal{O}\left(\frac{C_\phi^4 H^{4/3}N^{1/6}\sqrt{\log m}}{m^{1/6}\lambda_1^{13/6}}\right). \qquad (103)
\end{aligned}
$$

Substituting eq. (103) into eq. (97), we have the following holds with probability at least $1 - N^{-2}H^{-4}$

$$\begin{aligned}&\left|b_{r,h}(x,\widehat{\Theta}) - b_{r,h}(x,\Theta_0)\right| \\ &\leq \mathcal{O}\left(\frac{C_\phi^2 H^{2/3} N^{1/12} (\log m)^{1/4}}{m^{1/12}\lambda_1^{13/12}}\right). \end{aligned} \tag{104}$$

Step IV: In Steps I and II, we show that the mean of the real reward $R_h(\cdot)$ can be well approximated by a linear function $\widetilde{R}_h(\cdot)$ with feature $\phi(\cdot,\theta_0)$ and our learned reward $\widehat{R}_h(\cdot)$ can be well approximated by a linear function $\overline{R}_h(\cdot)$ with feature $\phi(\cdot,\theta_0)$. In this step, we want to show that the reward estimation error $\left|\widehat{R}_h(\cdot) - R_h(\cdot)\right|$ is approximately $\beta_1 \cdot b_{r.h}(x,\Theta_0)$ with an approximately chosen β_1.

Recall that $\widetilde{R}_h(\cdot) = \langle \phi(\cdot,\theta_0), \tilde{\theta}_h - \theta_0\rangle$ and $\overline{R}_h(\cdot) = \langle \phi(\cdot,\theta_0), \bar{\theta}_h - \theta_0\rangle$. Considering the difference between $\overline{R}_h(\cdot)$ and $\widetilde{R}_h(\cdot)$, we have

$$\begin{aligned}&\overline{R}_h(x) - \widetilde{R}_h(x) \\ &= \langle \phi(x,\theta_0), \bar{\theta}_h - \tilde{\theta}_h\rangle \\ &= \langle \Phi_h(x,\Theta_0), \overline{\Theta} - \widetilde{\Theta}\rangle, \end{aligned} \tag{105}$$

where the last equality follows from the definition of $\Phi_h(\cdot,\Theta)$. By eq. (85), we have

$$\overline{\Theta} - \Theta_0 = \Sigma(\Theta_0)^{-1} \sum_{\tau\in\mathcal{D}} r(\tau)\Phi(\tau,\Theta_0). \tag{106}$$

By the definition of $\Sigma(\Theta)$, we have

$$\begin{aligned}&\widetilde{\Theta} - \Theta_0 \\ &= \Sigma(\Theta_0)^{-1}\Bigg[\lambda_1\left(\widetilde{\Theta} - \Theta_0\right) \\ &\quad + \left(\sum_{\tau\in\mathcal{D}} \Phi(\tau,\Theta_0)\Phi(\tau,\Theta_0)^\top\right)\left(\widetilde{\Theta} - \Theta_0\right)\Bigg]. \end{aligned} \tag{107}$$

Subtracting eq. (107) from eq. (106), we have

$$\begin{aligned}&\overline{\Theta} - \widetilde{\Theta} \\ &= -\lambda_1\Sigma(\Theta_0)^{-1}\left(\widetilde{\Theta} - \Theta_0\right) \\ &\quad + \Sigma(\Theta_0)^{-1}\sum_{\tau\in\mathcal{D}} \Phi(\tau,\Theta_0) \\ &\quad \cdot \left[r(\tau) - \langle \Phi(\tau,\Theta_0), \widetilde{\Theta} - \Theta_0\rangle\right]. \end{aligned} \tag{108}$$

Taking inter product of both sides of eq. (108) with vector $\Phi_h(x,\Theta_0)$ and using the fact that $R(\tau) = \sum_{h\in[H]} R_h(x_h^\tau)$ and $\langle \Phi(\tau,\Theta_0), \widetilde{\Theta} - \Theta_0\rangle = \sum_{h\in[H]}\langle \phi(x_h^\tau,\theta_0), \tilde{\theta}_h - \theta_0\rangle$, we have

$$\begin{aligned}&\langle \Phi_h(x,\Theta_0), \overline{\Theta} - \widetilde{\Theta}\rangle \\ &= -\lambda_1 \Phi_h(x,\Theta_0)^\top \Sigma(\Theta_0)^{-1/2}\Sigma(\Theta_0)^{-1/2} \\ &\quad \cdot \left(\widetilde{\Theta} - \Theta_0\right) \\ &\quad + \Phi_h(x,\Theta_0)^\top \Sigma(\Theta_0)^{-1/2}\Sigma(\Theta_0)^{-1/2} \\ &\quad \cdot \left(\sum_{\tau\in\mathcal{D}} \Phi(\tau,\Theta_0)\left(r(\tau) - R(\tau)\right)\right) \\ &\quad + \Phi_h(x,\Theta_0)^\top \Sigma(\Theta_0)^{-1/2}\Sigma(\Theta_0)^{-1/2} \\ &\quad \cdot \left(\sum_{\tau\in\mathcal{D}} \Phi(\tau,\Theta_0)\right. \\ &\quad \left.\cdot \left[\sum_{h\in[H]}\left(R_h(x_h^\tau) - \langle \phi(x_h^\tau,\theta_0), \tilde{\theta}_h - \theta_0\rangle\right)\right]\right). \end{aligned} \tag{109}$$

Recall that $\widetilde{R}_h(x_h^\tau) = \langle \phi(x_h^\tau,\theta_0), \tilde{\theta}_h - \theta_0\rangle$, and eq. (109) implies that

$$\begin{aligned}&\left|\langle \Phi_h(x,\Theta_0), \overline{\Theta} - \widetilde{\Theta}\rangle\right| \\ &\leq \sqrt{\lambda_1}\left\|\Phi_h(x,\Theta_0)^\top\Sigma(\Theta_0)^{-1/2}\right\|_2 \left\|\widetilde{\Theta} - \Theta_0\right\|_2 \\ &\quad + \left\|\Phi_h(x,\Theta_0)^\top\Sigma(\Theta_0)^{-1/2}\right\|_2 \\ &\quad \cdot \left\|\sum_{\tau\in\mathcal{D}} \Phi(\tau,\Theta_0)\varepsilon(\tau)\right\|_{\Sigma(\Theta_0)^{-1}} \\ &\quad + \frac{1}{\sqrt{\lambda_1}}\left\|\Phi_h(x,\Theta_0)^\top\Sigma(\Theta_0)^{-1/2}\right\|_2 \\ &\quad \cdot \left(\sum_{\tau\in\mathcal{D}} \|\Phi(\tau,\Theta_0)\|_2\right. \\ &\quad \left.\cdot \sum_{h\in[H]}\left|R_h(x_h^\tau) - \widetilde{R}_h(x_h^\tau)\right|\right), \end{aligned} \tag{110}$$

where we denote $\epsilon(\tau) = r(\tau) - R(\tau)$ and use the fact that $\left\|\Sigma(\Theta)^{-1/2}\right\|_2 \leq 1/\sqrt{\lambda_1}$ for any $\Theta \in \mathbb{R}^{2mdH}$. By the definition of $\widetilde{\Theta}$ in Step I, we have

$$\begin{aligned}\left\|\widetilde{\Theta} - \Theta\right\|_2 &= \sqrt{\sum_{h\in[H],r\in[m]}\left\|\tilde{\theta}_{h,r} - \theta_{h,r}^0\right\|_2^2} \\ &= \sqrt{\sum_{h\in[H],r\in[m]}\|\ell_r\|_2^2} \\ &\leq r_2\sqrt{H/d}. \end{aligned} \tag{111}$$

By Lemma 5 and eq. (78), we have the followings hold with probability at least $1 - N^{-2}H^{-4}$

$$\|\Phi(\tau,\Theta_0)\|_2 \leq C_\phi\sqrt{H}, \tag{112}$$

$$\begin{aligned}&\left|R_h(x_h^\tau) - \widetilde{R}_h(x_h^\tau)\right| \\ &\leq \frac{2(L_\sigma a_2 + C_\sigma^2 a_2^2)\sqrt{\log N^2H^5}}{\sqrt{m}}. \end{aligned} \tag{113}$$

Substituting eq. (111), eq. (112) and eq. (113) into eq. (110) and using the fact that $b_{r,h}(x,\Theta_0) = \left\|\Phi_h(x,\Theta_0)^\top\Sigma(\Theta_0)^{-1/2}\right\|_2$, we have

$$\begin{aligned}&\left|\langle \Phi_h(x,\Theta_0), \overline{\Theta} - \widetilde{\Theta}\rangle\right| \\ &\leq \left(a_2\sqrt{\frac{\lambda_1 H}{d}}\right. \\ &\quad + \frac{2(L_\sigma a_2 + C_\sigma^2 a_2^2)C_\phi N H^{3/2}\sqrt{\log HN}}{\sqrt{\lambda_1 m}} \\ &\quad \left. + \left\|\sum_{\tau\in\mathcal{D}} \Phi(\tau,\Theta_0)\varepsilon(\tau)\right\|_{\Sigma(\Theta_0)^{-1}}\right) b_{r,h}(x,\Theta_0). \end{aligned} \tag{114}$$

Given that the events in eq. (112) and eq. (113) occur, applying eq. (235) in Lemma 8, we have the following holds with probability at least $1 - N^{-2}H^{-4}$

$$\begin{aligned}&\left\|\sum_{\tau\in\mathcal{D}}\Phi(\tau,\Theta_0)\varepsilon(\tau)\right\|^2_{\Sigma(\Theta_0)^{-1}}\\&\le H^2\log\det(I+K_N^r/\lambda_1)\\&\quad+H^2N(\lambda_1-1)+4H^2\log(NH^2),\end{aligned}\tag{115}$$

where $K_N^r\in\mathbb{R}^{N\times N}$ is the Gram matrix defined as

$$K_N^r=[K_H(\tau_i,\tau_j)]_{i,j\in[N]}\in\mathbb{R}^{N\times N}.$$

Combining eq. (114) and eq. (115) and letting $\lambda_1=1+N^{-1}$ and m be sufficiently large such that

$$\begin{aligned}&\frac{2(L_\sigma a_2+C_\sigma^2a_2^2)C_\phi NH^{3/2}\sqrt{\log HN}}{\sqrt{\lambda_1 m}}\\&\le a_2\sqrt{\frac{\lambda_1 H}{d}},\end{aligned}$$

we have the following holds with probability at least $1-N^{-2}H^{-2}$

$$\begin{aligned}&\left|\langle\Phi_h(x,\Theta_0),\overline{\Theta}-\widetilde{\Theta}\rangle\right|\\&\le\Bigg(2a_2\sqrt{\frac{\lambda_1 H}{d}}\\&\quad+\sqrt{H^2\log\det\left(I+\frac{K_N^r}{\lambda_1}\right)}\\&\quad+\sqrt{H^2+4H^2\log(NH^2)}\Bigg)b_{r,h}(x,\Theta_0)\\&\le H\Bigg(\frac{4a_2^2\lambda_1}{d}+2\log\det\left(I+\frac{K_N^r}{\lambda_1}\right)\\&\quad+10\log(NH^2)\Bigg)^{1/2}b_{r,h}(x,\Theta_0),\end{aligned}\tag{116}$$

where in the last inequality we use the fact that $a+b\le\sqrt{2(a^2+b^2)}$. Substituting eq. (116) into eq. (105), we have the following holds with probability at least $1-N^{-2}H^{-4}$

$$\left|\overline{R}_h(x)-\widetilde{R}_h(x)\right|\le\beta_1\cdot b_{r,h}(x,\Theta_0),\tag{117}$$

where

$$\begin{aligned}\beta_1=H\Bigg(&\frac{4a_2^2\lambda_1}{d}+2\log\det\left(I+\frac{K_N^r}{\lambda_1}\right)\\&+10\log(NH^2)\Bigg)^{1/2}.\end{aligned}$$

Next, we proceed to bound the reward estimation error $\left|R_h(x)-\widehat{R}_h(x)\right|$. By the triangle inequality, we have

$$\begin{aligned}&\left|R_h(x)-\widehat{R}_h(x)\right|\\&=|R_h(x)-\widetilde{R}_h(x)+\widetilde{R}_h(x)-\overline{R}_h(x)\\&\quad+\overline{R}_h(x)-\widehat{R}_h(x)|\\&\le\left|R_h(x)-\widetilde{R}_h(x)\right|+\left|\widetilde{R}_h(x)-\overline{R}_h(x)\right|\\&\quad+\left|\overline{R}_h(x)-\widehat{R}_h(x)\right|\\&\overset{(i)}{\le}\frac{2(L_\sigma a_2+C_\sigma^2a_2^2)\sqrt{\log(HN)}}{\sqrt{m}}\\&\quad+\mathcal{O}\left(\frac{C_\phi^3H^{17/6}N^{5/3}\sqrt{\log(m)}}{m^{1/6}\lambda_1^{5/3}}\right)\\&\quad+\beta_1\cdot b_{r,h}(x,\Theta_0)\\&\overset{(ii)}{\le}\mathcal{O}\left(\frac{H^{17/6}N^{5/3}\sqrt{\log(m)}}{m^{1/6}}\right)\\&\quad+\beta_1\cdot b_{r,h}(x,\Theta_0).\end{aligned}\tag{118}$$

where (i) follows from eq. (78) and eq. (96) and (ii) follows from the fact that $\lambda_1=1+1/N$ and $L_\sigma,C_\sigma,a_2,C_\phi=\mathcal{O}(1)$.

B. Uncertainty of Estimated Transition Value Function $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$

In this subsection, we aim to bound the estimation error of the transition value function $\left|(\widehat{\mathbb{P}}_h\widehat{V})(\cdot)-(\mathbb{P}_h\widehat{V})(\cdot)\right|$. For each $h\in[H]$, since $\widehat{w}_h$ is the global minimizer of the loss function $L_v^h(w_h)$ defined in eq. (24), we have

$$\begin{aligned}&L_v^h(\widehat{w}_h)\\&=\sum_{\tau\in\mathcal{D}}\left(\widehat{V}_{h+1}(s_{h+1}^\tau)-f(x_h^\tau,\widehat{w}_h)\right)^2\\&\quad+\lambda_2\cdot\|\widehat{w}_h-w_0\|_2^2\\&\le L_v^h(\widehat{w}_0)\\&=\sum_{\tau\in\mathcal{D}}\left(\widehat{V}_{h+1}(s_{h+1}^\tau)-f(x_h^\tau,w_0)\right)^2\\&\overset{(i)}{=}\sum_{\tau\in\mathcal{D}}\left(\widehat{V}_{h+1}(s_{h+1}^\tau)\right)^2\\&\overset{(ii)}{\le}NH^2,\end{aligned}\tag{119}$$

where (i) follows from the fact that $f(x,w_0)=0$ for all $x\in\mathcal{X}$ and (ii) follows from the fact that $\widehat{V}_h(s)\le H$ for any $h\in[H]$, $s\in\mathcal{S}$, and $|\mathcal{D}|=N$. Note that eq. (119) implies

$$\|\widehat{w}_h-w_0\|_2^2\le NH^2/\lambda_2,\quad\forall h\in[H].\tag{120}$$

Hence, each $\widehat{w}_h$ belongs to the Euclidean ball $\mathcal{B}_w=\{w\in\mathbb{R}^{2md}:\|w-w_0\|_2\le H\sqrt{N/\lambda_2}\}$, where λ_2 does not depend on the network width m. Since the radius of $\mathcal{B}_w$ does not depend on m, when m is sufficient large, it can be shown that $f(\cdot,w)$ is close to its linearization at w_0, i.e.,

$$f(\cdot,w)\approx\langle\phi(\cdot,w_0),w-w_0\rangle,\quad\forall w\in\mathcal{B}_\theta,$$

where $\phi(\cdot,w)=\nabla_w f(\cdot,w)$. Furthermore, according to Assumbtion 3, there exists a function $\ell_{A_1,A_2}:\mathbb{R}^d\to\mathbb{R}^d$ such that $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ satisfies

$$(\mathbb{P}_h\widehat{V}_{h+1})(x)=\int_{\mathbb{R}^d}\sigma'(\theta^\top x)\cdot x^\top\ell_v(w)dp(w),\tag{121}$$

where $\sup_w\|\ell_v(w)\|_2\le A_1$, $\sup_w(\|\ell_v(w)\|_2/p(w))\le A_2$ and p is the density of $N(0,I_d/d)$. We then proceed to bound the difference between $(\widehat{\mathbb{P}}_h\widehat{V})(\cdot)$ and $(\mathbb{P}_h\widehat{V})(\cdot)$.

Step I: In the first step, we show that the transition value function $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ can be well-approximated by a linear function with the feature vector $\phi(\cdot,\theta_0)$. Lemma 4 in Section I implies that with probability at least $1-N^{-2}H^{-4}$ over the randomness of initialization w_0, for all $h\in[H]$, there exists a function $(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot):\mathcal{X}\to\mathbb{R}$ satisfying

$$\sup_{x\in\mathcal{X}}\left|(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(x)-(\mathbb{P}_h\widehat{V}_{h+1})(x)\right| \le \frac{2(L_\sigma A_2+C_\sigma^2A_2^2)\sqrt{\log(N^2H^5)}}{\sqrt{m}}, \tag{122}$$

where $(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ is a finite-width neural network which can be written as

$$(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(x)=\frac{1}{\sqrt{m}}\sum_{r=1}^{m}\sigma'(w_{0,r}^\top x)\cdot x^\top\ell_r^v,$$

where $\|\ell_r^v\|_2\le A_2/\sqrt{dm}$ for all $r\in[m]$ and $w_0=[w_{0,1},\cdots,w_{0,m}]$ is generated via the symmetric initialization scheme. Following steps similar to those in eq. (79), we can show that there exists a vector $\tilde{w}_h\in\mathbb{R}^{2md}$ such that

$$(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)=\langle\phi(\cdot,w_0),\tilde{w}_h-w_0\rangle,$$

where $\tilde{w}_h=[\tilde{w}_{h,1}^\top,\cdots.\tilde{w}_{h,2m}^\top]^\top$, in which $\tilde{w}_{h,r}^\top=w_{0,r}+b_{0,r}\cdot\ell_r^v/\sqrt{2}$ for all $r\in\{1,\cdots,m\}$ and $\tilde{w}_{h,r}^\top=w_{0,r}+b_{0,r}\cdot\ell_{r-m}^v/\sqrt{2}$ for all $r\in\{m+1,\cdots,2m\}$. Moreover, since $\tilde{w}_{h,r}-w_{0,r}=b_{0,r}\cdot\ell_r^v/\sqrt{2}$ or $b_{0,r}\cdot\ell_{r-m}^v/\sqrt{2}$, we have

$$\|\tilde{w}_h-w_0\|_2\le A_2\sqrt{2dm}.$$

Step II: In the second step, we show that with high probability, the estimation of the transition value function $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ in Algorithm 2 can be well-approximated by its counterpart learned with a linear function with the feature $\phi(\cdot,\theta_0)$.

Consider the following least-square loss function

$$\begin{aligned}&\bar{L}_v^h(w_h)\\&=\sum_{\tau\in\mathcal{D}}\left(\widehat{V}_{h+1}(s_{h+1}^\tau)-\langle\phi(x_h^\tau,w_0),w_h-w_0\rangle\right)^2\\&\quad+\lambda_2\cdot\|w_h-w_0\|_2^2.\end{aligned}\tag{123}$$

The global minimizer of $\bar{L}_v^h(w_h)$ is defined as

$$\overline{w}_h=\operatorname*{argmin}_{w\in\mathbb{R}^{2dm}}\bar{L}_v^h(w).\tag{124}$$

We define $(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)=\langle\phi(\cdot,w_0),\overline{w}_h-w_0\rangle$ for all $h\in[H]$. Then, in a manner similar to the construction of $\widehat{Q}_h(\cdot)$ in Algorithm 2, we combine $\overline{R}_h(\cdot)$ in eq. (80), $b_{r,h}(\cdot,\Theta_0)$, $\overline{\mathbb{P}}_h\widehat{V}_{h+1}(\cdot)$ and $b_{v,h}(\cdot,w_0)$ to construct $\overline{Q}_h(\cdot):\mathcal{X}\to\mathbb{R}$ as

$$\begin{aligned}&\overline{Q}_h(\cdot)\\&=\min\{\overline{R}_h(\cdot)+(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)\\&\quad-\beta_1\cdot b_{r,h}(\cdot,\Theta_0)-\beta_2\cdot b_{v,h}(\cdot,w_0),H\}^+.\end{aligned}\tag{125}$$

Moreover, we define the estimated optimal state value function as

$$\overline{V}_h(\cdot)=\max_{a\in\mathcal{A}}\overline{Q}_h(\cdot,a).\tag{126}$$

We then proceed to bound the estimation error $\left|(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(x)-\overline{\mathbb{P}}_h\widehat{V}_{h+1})(x)\right|$ as follows

$$\begin{aligned}&\left|(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(x)-\overline{\mathbb{P}}_h\widehat{V}_{h+1})(x)\right|\\&=\left|f(x,\widehat{w}_h)-\langle\phi(x,w_0),\overline{w}_h-w_0\rangle\right|\\&=\left|(x,\widehat{w}_h)-\langle\phi(x,w_0),\widehat{w}_h-w_0\rangle\right.\\&\quad\left.-\langle\phi(\cdot,w_0),\widehat{w}_h-\overline{w}_h\rangle\right|\\&\le|f(x,\widehat{w}_h)-\langle\phi(x,w_0),\widehat{w}_h-w_0\rangle|\\&\quad+|\langle\phi(\cdot,w_0),\widehat{w}_h-\overline{w}_h\rangle|\\&\le\underbrace{|f(x,\widehat{w}_h)-\langle\phi(\cdot,w_0),\widehat{w}_h-w_0\rangle|}_{(i)}\\&\quad+\underbrace{\|\phi(x,w_0)\|_2\|\widehat{w}_h-\overline{w}_h\|_2}_{(ii)}.\end{aligned}$$

We then bound the term (i) and term (ii) in the above inequality. According to Lemma 5 and the fact that $\|\widehat{w}_h-w_0\|_2\le H\sqrt{N/\lambda_2}$, we have the followings hold with probability at least $1-N^{-2}H^{-4}$

$$(i)\le\mathcal{O}\left(C_\phi\left(\frac{N^2H^4}{\lambda_2^2\sqrt{m}}\right)^{1/3}\sqrt{\log m}\right),\tag{127}$$

$$(ii)\le C_\phi\|\widehat{w}_h-\overline{w}_h\|_2.\tag{128}$$

We then proceed to bound $\|\widehat{w}_h-\overline{w}_h\|_2$. Consider the minimization problem defined in eq. (31) and eq. (80). By the first order optimality condition, we have

$$\begin{aligned}&\lambda_2(\widehat{w}_h-w_0)\\&=\sum_{\tau\in\mathcal{D}}\left(\widehat{V}_{h+1}(s_{h+1}^\tau)-f(x_h^\tau,\widehat{w}_h)\right)\phi(x_h^\tau,\widehat{w}_h)\end{aligned}\tag{129}$$

$$\begin{aligned}&\lambda_2(\overline{w}_h-w_0)\\&=\sum_{\tau\in\mathcal{D}}\left(\widehat{V}_{h+1}(s_{h+1}^\tau)-\langle\phi(x_h^\tau,w_0),\overline{w}_h-w_0\rangle\right)\\&\quad\cdot\phi(x_h^\tau,w_0).\end{aligned}\tag{130}$$

Note that eq. (130) implies

$$\Lambda_h(w_0)(\overline{w}_h-w_0)=\sum_{\tau\in\mathcal{D}}\widehat{V}_{h+1}(s_{h+1}^\tau)\phi(x_h^\tau,w_0).\tag{131}$$

Adding the term $\sum_{\tau\in\mathcal{D}}\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle\phi(x_h^\tau,w_0)$ on both sides of eq. (129) yields

$$\begin{aligned}&\Lambda_h(w_0)(\widehat{w}_h-w_0)\\&=\sum_{\tau\in\mathcal{D}}\widehat{V}_{h+1}(s_{h+1}^\tau)\phi(x_h^\tau,\widehat{w}_h)\\&\quad+\sum_{\tau\in\mathcal{D}}\left[\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle\phi(x_h^\tau,w_0)\right.\\&\quad\left.-f(x_h^\tau,\widehat{w}_h)\phi(x_h^\tau,\widehat{w}_h)\right].\end{aligned}\tag{132}$$

Then, by subtracting eq. (131) from eq. (132), we have

$$\begin{aligned}&\Lambda_h(w_0)(\widehat{w}_h-\overline{w}_h)\\&=\sum_{\tau\in\mathcal{D}}\widehat{V}_{h+1}(s_{h+1}^\tau)\left(\phi(x_h^\tau,\widehat{w}_h)-\phi(x_h^\tau,w_0)\right)\\&\quad+\sum_{\tau\in\mathcal{D}}\left[\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle\phi(x_h^\tau,w_0)\right.\\&\quad\left.-f(x_h^\tau,\widehat{w}_h)\phi(x_h^\tau,\widehat{w}_h)\right],\end{aligned}\tag{133}$$

which implies

$$\begin{aligned}&\|\Lambda_h(w_0)(\widehat{w}_h-\overline{w}_h)\|_2\\&=\sum_{\tau\in\mathcal{D}}\widehat{V}_{h+1}(s_{h+1}^\tau)\,\|\phi(x_h^\tau,\widehat{w}_h)-\phi(x_h^\tau,w_0)\|_2\\&\quad+\sum_{\tau\in\mathcal{D}}\big\|\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle\phi(x_h^\tau,w_0)\\&\quad-f(x_h^\tau,\widehat{w}_h)\phi(x_h^\tau,\widehat{w}_h)\big\|.\end{aligned}\tag{134}$$

To bound the term $\|\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle\phi(x_h^\tau,w_0)-f(x_h^\tau,\widehat{w}_h)\phi(x_h^\tau,\widehat{w}_h)\|$, we proceed as follows

$$\begin{aligned}&\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle\phi(x_h^\tau,w_0)-f(x_h^\tau,\widehat{w}_h)\phi(x_h^\tau,\widehat{w}_h)\\&=\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle(\phi(x_h^\tau,w_0)-\phi(x_h^\tau,\widehat{w}_h))\\&\quad-(\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle-f(x_h^\tau,\widehat{w}_h))\phi(x_h^\tau,\widehat{w}_h)\end{aligned}$$

which implies

$$\begin{aligned}&\|\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle\phi(x_h^\tau,w_0)-f(x_h^\tau,\widehat{w}_h)\phi(x_h^\tau,\widehat{w}_h)\|_2\\&\le\|\phi(x_h^\tau,w_0)\|_2\,\|\widehat{w}_h-w_0\|_2\\&\quad\cdot\|\phi(x_h^\tau,w_0)-\phi(x_h^\tau,\widehat{w}_h)\|_2\\&\quad+|\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle-f(x_h^\tau,\widehat{w}_h)|\\&\quad\cdot\|\phi(x_h^\tau,\widehat{w}_h)\|_2\,.\end{aligned}\tag{135}$$

According to Lemma 5 and the fact that $\|\widehat{w}_h-w_0\|_2\le H\sqrt{N/\lambda_2}$, we have the followings hold for all $h\in[H]$ and $\tau\in\mathcal{D}$ with probability at least $1-N^{-2}H^{-4}$

$$\|\phi(x_h^\tau,w_0)\|_2\le C_\phi,\tag{136}$$

$$\|\phi(x_h^\tau,\widehat{w}_h)\|_2\le C_\phi,\tag{137}$$

$$\begin{aligned}&\|\phi(x_h^\tau,w_0)-\phi(x_h^\tau,\widehat{w}_h)\|_2\\&\le\mathcal{O}\left(C_\phi\left(\frac{H\sqrt{N/\lambda_2}}{\sqrt{m}}\right)^{1/3}\sqrt{\log m}\right),\end{aligned}\tag{138}$$

$$\begin{aligned}&|\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle-f(x_h^\tau,\widehat{w}_h)|\\&\le\mathcal{O}\left(C_\phi\left(\frac{H^4N^2/\lambda_2^2}{\sqrt{m}}\right)^{1/3}\sqrt{\log m}\right).\end{aligned}\tag{139}$$

Substituting eq. (137), eq. (138) and eq. (139) into eq. (135), we can obtain

$$\begin{aligned}&\|\langle\phi(x_h^\tau,w_0),\widehat{w}_h-w_0\rangle\phi(x_h^\tau,w_0)\\&\quad-f(x_h^\tau,\widehat{w}_h)\phi(x_h^\tau,\widehat{w}_h)\|\\&\le(H\sqrt{N/\lambda_2})\\&\quad\cdot\mathcal{O}\left(C_\phi^2\left(\frac{H\sqrt{N/\lambda_2}}{\sqrt{m}}\right)^{1/3}\sqrt{\log m}\right)\\&\quad+\mathcal{O}\left(C_\phi^2\left(\frac{H^4N^2/\lambda_2^2}{\sqrt{m}}\right)^{1/3}\sqrt{\log m}\right)\\&\le\mathcal{O}\left(\frac{C_\phi^2H^{4/3}N^{2/3}\sqrt{\log(m)}}{m^{1/6}\lambda_2^{2/3}}\right).\end{aligned}\tag{140}$$

Substituting eq. (140) into eq. (134), we have the following holds with probability at least $1-N^{-2}H^{-4}$

$$\begin{aligned}&\|\Lambda_h(w_0)(\widehat{w}_h-\overline{w}_h)\|_2\\&\le NH\cdot\mathcal{O}\left(C_\phi\left(\frac{H\sqrt{N/\lambda_2}}{\sqrt{m}}\right)^{1/3}\sqrt{\log m}\right)\\&\quad+N\cdot\mathcal{O}\left(\frac{C_\phi^2H^{4/3}N^{2/3}\sqrt{\log(m)}}{m^{1/6}\lambda_2^{2/3}}\right)\\&\le\mathcal{O}\left(\frac{C_\phi^2H^{4/3}N^{5/3}\sqrt{\log(m)}}{m^{1/6}\lambda_2^{2/3}}\right)\end{aligned}\tag{141}$$

where we use the fact that $\widehat{V}_{h+1}(s)\le H$ for any $s\in\mathcal{S}$. We then proceed to bound $\|\widehat{w}_h-\overline{w}_h\|_2$ as follows

$$\begin{aligned}&\|\widehat{w}_h-\overline{w}_h\|_2\\&=\big\|\Lambda^{-1}(w_0)\Lambda(w_0)(\widehat{w}_h-\overline{w}_h)\big\|_2\\&\le\big\|\Lambda^{-1}(w_0)\big\|_2\,\|\Lambda(w_0)(\widehat{w}_h-\overline{w}_h)\|_2\\&\le\frac{1}{\lambda_2}\cdot\mathcal{O}\left(\frac{C_\phi^2H^{4/3}N^{5/3}\sqrt{\log(m)}}{m^{1/6}\lambda_2^{2/3}}\right)\\&\le\mathcal{O}\left(\frac{C_\phi^2H^{4/3}N^{5/3}\sqrt{\log(m)}}{m^{1/6}\lambda_2^{5/3}}\right).\end{aligned}\tag{142}$$

Substituting eq. (142) into eq. (128) yields

$$(ii)\le\mathcal{O}\left(\frac{C_\phi^3H^{4/3}N^{5/3}\sqrt{\log(m)}}{m^{1/6}\lambda_2^{5/3}}\right).\tag{143}$$

Taking summation of the upper bounds of (i) in eq. (127) and (ii) in eq. (143), respectively, we have the following holds for all $x\in\mathcal{X}$ with probability at least $1-N^{-2}H^{-4}$

$$\begin{aligned}&\left|(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(x)-\overline{\mathbb{P}}_h\widehat{V}_{h+1})(x)\right|\\&\le(i)+(ii)\\&\le\mathcal{O}\left(C_\phi\left(\frac{N^2H^4}{\lambda_2^2\sqrt{m}}\right)^{1/3}\sqrt{\log m}\right)\\&\quad+\mathcal{O}\left(\frac{C_\phi^3H^{4/3}N^{5/3}\sqrt{\log(m)}}{m^{1/6}\lambda_2^{5/3}}\right)\\&\le\mathcal{O}\left(\frac{C_\phi^3H^{4/3}N^{5/3}\sqrt{\log(m)}}{m^{1/6}\lambda_2^{5/3}}\right).\end{aligned}\tag{144}$$

Step III: In this step, we show that the bonus term $b_{v,h}(\cdot,\widehat{w}_h)$ in Algorithm 2 can be well approximated by $b_{v,h}(\cdot,w_0)$. By the definition of $b_{v,h}(\cdot,w)$, we have

$$\begin{aligned}&|b_{v,h}(x,\widehat{w}_h)-b_{v,h}(x,w_0)|\\&=\big|\left[\phi_h(x,\widehat{w}_h)^\top\Lambda^{-1}(\widehat{w}_h)\phi_h(x,\widehat{w}_h)\right]^{1/2}\\&\quad-\left[\phi_h(x,w_0)^\top\Lambda^{-1}(w_0)\phi_h(x,w_0)\right]^{1/2}\big|\\&\le\big|\phi_h(x,\widehat{w}_h)^\top\Lambda^{-1}(\widehat{w}_h)\phi_h(x,\widehat{w}_h)\\&\quad-\phi_h(x,w_0)^\top\Lambda^{-1}(w_0)\phi_h(x,w_0)\big|^{1/2},\end{aligned}\tag{145}$$

where the last inequality follows from the fact that $\big|\sqrt{x}-\sqrt{y}\big|\le\sqrt{|x-y|}$. Following steps similar to those in eq. (98), we can obtain

$$\begin{aligned}&|\phi_h(x,\widehat{w}_h)^\top\Lambda^{-1}(\widehat{w}_h)\phi_h(x,\widehat{w}_h)\\&\quad-\phi_h(x,w_0)^\top\Lambda^{-1}(w_0)\phi_h(x,w_0)|\\&\le\|\phi(x,\widehat{w}_h)-\phi(x,w_0)\|_2\,\big\|\Lambda^{-1}(\widehat{w}_h)\big\|_2\end{aligned}$$

$$
\begin{aligned}
&\cdot \|\phi(x,\widehat{w}_h)\|_2 \\
&+ \|\phi(x,w_0)\|_2 \left\|\Lambda^{-1}(\widehat{w}_h)\right\|_2 \|\Lambda(\widehat{w}_h)-\Lambda(w_0)\|_2 \\
&\cdot \left\|\Lambda^{-1}(w_0)\right\|_2 \|\phi(x,\widehat{w}_h)\|_2 \\
&+ \|\phi(x,w_0)\|_2 \left\|\Lambda^{-1}(w_0)\right\|_2 \\
&\cdot \|\phi(x,\widehat{w}_h)-\phi(x,w_0)\|_2 \\
\le\ & \frac{1}{\lambda_2}\|\phi(x,\widehat{w}_h)-\phi(x,w_0)\|_2 \|\phi(x,\widehat{w}_h)\|_2 \\
&+ \frac{\|\phi(x,\widehat{w}_h)\|_2}{\lambda_2^2}\|\phi(x,w_0)\|_2 \|\Lambda(\widehat{w}_h)-\Lambda(w_0)\|_2 \\
&+ \frac{1}{\lambda_2}\|\phi(x,w_0)\|_2 \|\phi(x,\widehat{w}_h)-\phi(x,w_0)\|_2 , \qquad (146)
\end{aligned}
$$

where the last inequality follows from the fact that $\|\Lambda(w)\|_2 \ge \lambda_2$ for any $w \in \mathbb{R}^{2md}$. For $\Lambda(\widehat{w})-\Lambda(w_0)$, by following steps similar to those in eq. (99), we can obtain

$$
\begin{aligned}
&\|\Lambda(\widehat{w}_h)-\Lambda(w_0)\|_2 \\
&\le \sum_{\tau\in\mathcal{D}} \big[\|\phi(x_h^\tau,\widehat{w}_h)\|_2 \|\phi(x_h^\tau,\widehat{w}_h)-\phi(x_h^\tau,w_0)\|_2 \\
&\quad + \|\phi(x_h^\tau,\widehat{w}_h)-\phi(x_h^\tau,w_0)\|_2 \|\phi(x_h^\tau,w_0)\|_2 \big]. \qquad (147)
\end{aligned}
$$

Applying Lemma 5 to eq. (147), we have the followings hold with probability at least $1-N^{-2}H^{-4}$

$$
\begin{aligned}
&\|\phi(x_h^\tau,\widehat{w}_h)\|_2 \le C_\phi, \\
&\|\phi(x_h^\tau,\widehat{w}_h)-\phi(x_h^\tau,w_0)\|_2 \\
&\le \mathcal{O}\left(\frac{C_\phi H^{1/3}N^{1/6}\sqrt{\log m}}{m^{1/6}\lambda_2^{1/6}}\right).
\end{aligned}
$$

Substituting the above two inequalities into eq. (99) yields

$$
\begin{aligned}
&\|\Lambda(\widehat{w}_h)-\Lambda(w_0)\|_2 \\
&\le \mathcal{O}\left(\frac{C_\phi^2 H^{1/3}N^{1/6}\sqrt{\log m}}{m^{1/6}\lambda_2^{1/6}}\right). \qquad (148)
\end{aligned}
$$

Finally, combining eq. (148) and eq. (214) and eq. (215) in Lemma 5, the right hand side of eq. (146) can be bounded by

$$
\begin{aligned}
&\left|\phi_h(x,\widehat{w}_h)^\top\Lambda^{-1}(\widehat{w}_h)\phi_h(x,\widehat{w}_h)\right. \\
&\quad \left. -\phi_h(x,w_0)^\top\Lambda^{-1}(w_0)\phi_h(x,w_0)\right| \\
&\le \frac{1}{\lambda_2}\|\phi(x,\widehat{w}_h)-\phi(x,w_0)\|_2 \|\phi(x,\widehat{w}_h)\|_2 \\
&\quad + \frac{1}{\lambda_2^2}\|\phi(x,w_0)\|_2 \|\Lambda(\widehat{w}_h)-\Lambda(w_0)\|_2 \|\phi(x,\widehat{w}_h)\|_2 \\
&\quad + \frac{1}{\lambda_2}\|\phi(x,w_0)\|_2 \|\phi(x,\widehat{w}_h)-\phi(x,w_0)\|_2 \\
&\le \mathcal{O}\left(\frac{C_\phi^2 H^{1/3}N^{1/6}\sqrt{\log m}}{m^{1/6}\lambda_2^{7/6}}\right) \\
&\quad + \mathcal{O}\left(\frac{C_\phi^4 H^{1/3}N^{1/6}\sqrt{\log m}}{m^{1/6}\lambda_2^{13/6}}\right).
\end{aligned}
$$

By eq. (145), we have the following holds with probability at least $1-N^{-2}H^{-4}$

$$
\begin{aligned}
&|b_{v,h}(x,\widehat{w}_h)-b_{v,h}(x,w_0)| \\
&\le \left|\phi_h(x,\widehat{w}_h)^\top\Lambda^{-1}(\widehat{w}_h)\phi_h(x,\widehat{w}_h)\right. \\
&\quad \left. -\phi_h(x,w_0)^\top\Lambda^{-1}(w_0)\phi_h(x,w_0)\right|^{1/2} \\
&\le \mathcal{O}\left(\frac{C_\phi^2 H^{1/6}N^{1/12}(\log m)^{1/4}}{m^{1/12}\lambda_2^{13/12}}\right). \qquad (149)
\end{aligned}
$$

Step IV: In Steps I and II, we show that $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ can be well approximated by a linear function $(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ with the feature $\phi(\cdot,\theta_0)$, and $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ can be well approximated by a linear function $(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ with the feature $\phi(\cdot,\theta_0)$. In this step, we want to show that the difference between $(\mathbb{P}_h\widehat{V}_{h+1})(\cdot)$ and $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ is approximately $\beta_2\cdot b_{v.h}(x,\Theta_0)$ with an approximately chosen β_2.

Recall that $(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot) = \langle\phi(\cdot,w_0),\tilde{w}_h-w_0\rangle$ and $(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot) = \langle\phi(\cdot,w_0),\bar{w}_h-w_0\rangle$. Consider the difference between $(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ and $(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$. We have

$$
\begin{aligned}
&(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(x)-(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(x) \\
&= \langle\phi(x,w_0),\bar{w}_h-\tilde{w}_h\rangle, \qquad (150)
\end{aligned}
$$

By eq. (130), we have

$$
\begin{aligned}
&\overline{w}-w_0 \\
&= \Lambda(w_0)^{-1}\sum_{\tau\in\mathcal{D}}\widehat{V}_{h+1}(s_{h+1}^\tau)\phi(x_h^\tau,w_0). \qquad (151)
\end{aligned}
$$

By the definition of $\Lambda(w)$, we have

$$
\begin{aligned}
&\widetilde{w}-w_0 \qquad (152)\\
&= \Lambda(w_0)^{-1}\Bigg[\lambda_2(\widetilde{w}-w_0) \\
&\quad + \left(\sum_{\tau\in\mathcal{D}}\phi(x_h^\tau,w_0)\phi(x_h^\tau,w_0)^\top\right)(\widetilde{w}-w_0)\Bigg].
\end{aligned}
$$

Subtracting eq. (152) from eq. (151), we have

$$
\begin{aligned}
&\overline{w}-\widetilde{w} \qquad (153)\\
&= -\lambda_2\Lambda(w_0)^{-1}(\widetilde{w}-w_0) \\
&\quad + \Lambda(w_0)^{-1}\sum_{\tau\in\mathcal{D}}\phi(x_h^\tau,w_0) \\
&\quad \cdot\left[\widehat{V}_{h+1}(s_{h+1}^\tau)-\langle\phi(x_h^\tau,w_0),\widetilde{w}-w_0\rangle\right].
\end{aligned}
$$

Taking inter product of both sides of eq. (153) with vector $\phi(x_h^\tau,w_0)$ and using the fact that $(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(s_{h+1}^\tau) = \langle\phi(x_h^\tau,w_0),\tilde{w}_h-w_0\rangle$, we have

$$
\begin{aligned}
&\langle\phi_h(x_h^\tau,w_0),\overline{w}-\widetilde{w}\rangle \\
&= -\lambda_2\phi_h(x_h^\tau,w_0)^\top\Lambda(w_0)^{-1/2}\Lambda(w_0)^{-1/2}(\widetilde{w}-w_0) \\
&\quad + \phi_h(x_h^\tau,w_0)^\top\Lambda(w_0)^{-1/2}\Lambda(w_0)^{-1/2} \\
&\quad \left(\sum_{\tau\in\mathcal{D}}\phi(x_h^\tau,w_0)\right. \\
&\quad \left.\left(\widehat{V}_{h+1}(s_{h+1}^\tau)-(\mathbb{P}_h\widehat{V}_{h+1})(x_h^\tau)\right)\right) \\
&\quad + \phi_h(x,w_0)^\top\Lambda(w_0)^{-1/2}\Lambda(w_0)^{-1/2} \\
&\quad \left(\sum_{\tau\in\mathcal{D}}\phi(x_h^\tau,w_0)\right. \\
&\quad \left.\left((\mathbb{P}_h\widehat{V}_{h+1})(s_{h+1}^\tau)-\langle\phi(x_h^\tau,w_0),\tilde{w}_h-w_0\rangle\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= -\lambda_2 \phi_h(x_h^\tau, w_0)^\top \Lambda(w_0)^{-1/2} \Lambda(w_0)^{-1/2} \\
&\quad (\widetilde{w} - w_0) \\
&\quad + \phi_h(x_h^\tau, w_0)^\top \Lambda(w_0)^{-1/2} \Lambda(w_0)^{-1/2} \\
&\quad \left(\sum_{\tau \in \mathcal{D}} \phi(x_h^\tau, w_0) \left(\overline{V}_{h+1}(s_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1})(x_h^\tau) \right) \right) \\
&\quad + \phi_h(x_h^\tau, w_0)^\top \Lambda(w_0)^{-1/2} \Lambda(w_0)^{-1/2} \\
&\quad \left(\sum_{\tau \in \mathcal{D}} \phi(x_h^\tau, w_0) \left(\Delta V_{h+1}(s_{h+1}^\tau) - (\mathbb{P}_h \Delta V_{h+1})(x_h^\tau) \right) \right) \\
&\quad + \phi_h(x, w_0)^\top \Lambda(w_0)^{-1/2} \Lambda(w_0)^{-1/2} \\
&\quad \left(\sum_{\tau \in \mathcal{D}} \phi(x_h^\tau, w_0) \left((\mathbb{P}_h \widehat{V}_{h+1})(x_h^\tau) - (\widetilde{\mathbb{P}}_h \widehat{V}_{h+1})(x_h^\tau) \right) \right),
\end{aligned} \tag{154}
$$

where in the last equality we denote $\Delta V_h(s) \coloneqq \widehat{V}_h(s) - \overline{V}_h(s)$. By the definition of $\widehat{V}_h(\cdot)$ in Algorithm 2 and $\overline{V}_h(\cdot)$ in eq. (126), we have

$$
\begin{aligned}
&\left| \widehat{V}_h(x) - \overline{V}_h(x) \right| \\
&\le \sup_{x \in \mathcal{X}} \left| \widehat{Q}_h(x) - \overline{Q}_h(x) \right| \\
&\le \left| f(x, \widehat{\theta}_h) - \langle \phi(x, \theta_0), \bar{\theta}_h - \theta_0 \rangle \right| \\
&\quad + \left| f(x, \widehat{w}_h) - \langle \phi(x, w_0), \overline{w}_h - w_0 \rangle \right| \\
&\quad + \beta_1 \left| b_{r,h}(x, \widehat{\Theta}) - b_{r,h}(x, \Theta_0) \right| \\
&\quad + \beta_2 \left| b_{v,h}(x, \widehat{w}) - b_{v,h}(x, w_0) \right| \\
&\overset{(i)}{\le} \mathcal{O}\left(C_\phi \left(\frac{H^4 N^2 / \lambda_1^2}{\sqrt{m}} \right)^{1/3} \sqrt{\log m} \right) \\
&\quad + \mathcal{O}\left(C_\phi \left(\frac{H^4 N^2 / \lambda_2^2}{\sqrt{m}} \right)^{1/3} \sqrt{\log m} \right) \\
&\quad + \beta_1 \cdot \mathcal{O}\left(\frac{C_\phi^2 N^{1/12} (\log m)^{1/4}}{m^{1/12} \lambda_1^{13/12}} \right) \\
&\quad + \beta_2 \cdot \mathcal{O}\left(\frac{C_\phi^2 H^{1/6} N^{1/12} (\log m)^{1/4}}{m^{1/12} \lambda_2^{13/12}} \right) \\
&\overset{(ii)}{\le} \mathcal{O}\left(C_\phi \left(\frac{H^4 N^2}{\sqrt{m}} \right)^{1/3} \sqrt{\log m} \right) \\
&\quad + \max\{H^{2/3} \beta_1, H^{1/6} \beta_2\} \\
&\quad \cdot \mathcal{O}\left(\frac{C_\phi^2 N^{1/12} (\log m)^{1/4}}{m^{1/12}} \right),
\end{aligned} \tag{155}
$$

where (i) follows from eq. (216) in Lemma 5, eq. (104) and eq. (149), and (ii) follows from the fact that $\lambda_1, \lambda_2 > 1$. Denoting

$$
\begin{aligned}
&\varepsilon_v \\
&= \mathcal{O}\left(C_\phi \left(\frac{H^4 N^2}{\sqrt{m}} \right)^{1/3} \sqrt{\log m} \right) \\
&\quad + \max\{H^{2/3} \beta_1, H^{1/6} \beta_2\} \\
&\quad \cdot \mathcal{O}\left(\frac{C_\phi^2 N^{1/12} (\log m)^{1/4}}{m^{1/12}} \right),
\end{aligned}
$$

we then have the following holds for all $h \in [H]$ and $s \in \mathcal{S}$

$$
|\Delta V_h(s)| \le \varepsilon_v.
$$

eq. (154) together with eq. (155) imply

$$
\begin{aligned}
&|\langle \phi_h(x_h^\tau, w_0), \overline{w} - \widetilde{w} \rangle| \\
&\le \sqrt{\lambda_2} \left\| \phi_h(x_h^\tau, w_0)^\top \Lambda(w_0)^{-1/2} \right\|_2 \|\widetilde{w} - w_0\|_2 \\
&\quad + \left\| \phi_h(x_h^\tau, w_0)^\top \Lambda(w_0)^{-1/2} \right\|_2 \\
&\quad \left\| \sum_{\tau \in \mathcal{D}} \phi(x_h^\tau, w_0) \varepsilon_v(x_h^\tau) \right\|_{\Lambda(w_0)^{-1}} \\
&\quad + \frac{2\varepsilon_v}{\sqrt{\lambda_2}} \left\| \phi_h(x_h^\tau, w_0)^\top \Lambda(w_0)^{-1/2} \right\|_2 \\
&\quad \sum_{\tau \in \mathcal{D}} \|\phi(x_h^\tau, w_0)\|_2 \\
&\quad + \frac{1}{\sqrt{\lambda_2}} \left\| \phi_h(x_h^\tau, w_0)^\top \Lambda(w_0)^{-1/2} \right\|_2 \\
&\quad \left(\sum_{\tau \in \mathcal{D}} \|\phi(x_h^\tau, w_0)\|_2 \right. \\
&\quad \left. \left| (\mathbb{P}_h \widehat{V}_{h+1})(x_h^\tau) - (\widetilde{\mathbb{P}}_h \widehat{V}_{h+1})(x_h^\tau) \right| \right),
\end{aligned} \tag{156}
$$

where we denote $\varepsilon_v(x_h^\tau) \coloneqq \overline{V}_{h+1}(s_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1})(x_h^\tau)$ and use the fact that $\left\| \Lambda(w)^{-1/2} \right\|_2 \le 1/\sqrt{\lambda_2}$ for any $w \in \mathbb{R}^{2md}$. By the definition of $\widetilde{w}$ in Step I, we have

$$
\|\widetilde{w} - w_0\|_2 = \|\ell_v\|_2 \le A_2 \sqrt{H/d}. \tag{157}
$$

By Lemma 5 and eq. (78), we have the followings hold with probability at least $1 - N^{-2} H^{-4}$ over the randomness of initialization w_0

$$
\|\phi(x_h^\tau, w_0)\|_2 \le C_\phi, \tag{158}
$$

$$
\begin{aligned}
&\left| (\mathbb{P}_h \widehat{V}_{h+1})(x_h^\tau) - (\widetilde{\mathbb{P}}_h \widehat{V}_{h+1})(x_h^\tau) \right| \\
&\le \frac{2(L_\sigma A_2 + C_\sigma^2 A_2^2) \sqrt{\log N^2 H^5}}{\sqrt{m}}.
\end{aligned} \tag{159}
$$

Substituting eq. (157), eq. (158) and eq. (159) into eq. (156) and using the fact that $b_{v,h}(x, w_0) = \left\| \phi_h(x, w_0)^\top \Lambda(w_0)^{-1/2} \right\|_2$, we have

$$
\begin{aligned}
&|\langle \phi_h(x_h^\tau, w_0), \overline{w} - \widetilde{w} \rangle| \\
&\le b_{v,h}(x_h^\tau, w_0) \left(A_2 \sqrt{\frac{\lambda_2 H}{d}} \right. \\
&\quad + \frac{2(L_\sigma A_2 + C_\sigma^2 A_2^2) C_\phi N H^{3/2} \sqrt{\log HN}}{\sqrt{\lambda_2 m}} \\
&\quad \left. + \left\| \sum_{\tau \in \mathcal{D}} \phi(x_h^\tau, w_0) \varepsilon_v(x_h^\tau) \right\|_{\Lambda(w_0)^{-1}} \right).
\end{aligned} \tag{160}
$$

Given that the events in eq. (158) and eq. (159) occur, applying eq. (234) in Lemma 8, we have the following holds with

probability at least $1-N^{-2}H^{-4}$

$$\begin{aligned}&\left\|\sum_{\tau\in\mathcal{D}}\phi(x_h^\tau,w_0)\varepsilon_v(x_h^\tau)\right\|^2_{\Lambda(w_0)^{-1}}\\&\le 2H^2\log\det(I+K^v_{N,h}/\lambda_2)+2H^2N(\lambda_2-1)\\&\quad+4H^2\log(\mathcal{N}^v_{\epsilon,h}/\delta)+8N^2C_\phi^2\epsilon^2/\lambda_2,\end{aligned}\tag{161}$$

where $K^v_{N,h}\in\mathbb{R}^{N\times N}$ is the Gram matrix defined as

$$K^v_{N,h}=[K(x_h^{\tau_i},x_h^{\tau_j})]_{i,j\in[N]}\in\mathbb{R}^{N\times N},$$

and $\mathcal{N}^v_{\epsilon,h}$ is the cardinality of the following function class

$$\begin{aligned}&\mathcal{V}_h(x,R_\theta,R_w,R_{\beta_1},R_{\beta_2},\lambda_1,\lambda_2)\\&\quad=\{\max_{a\in\mathcal{A}}\{\overline{Q}_h(s,a)\}:\mathcal{S}\to[0,H]\\&\|\theta\|_2\le R_\theta,\|w\|_2\le R_w,\beta_1\in[0,R_{\beta_1}],\beta_2\in[0,R_{\beta_2}],\\&\|\Sigma\|_2\ge\lambda_1,\|\Lambda\|_2\ge\lambda_2\},\end{aligned}$$

where $R_\theta=H\sqrt{N/\lambda_1}$, $R_w=H\sqrt{N/\lambda_2}$ and

$$\begin{aligned}&\overline{Q}_h(x)\\&=\min\{\langle\phi(x,\theta_0),\theta-\theta_0\rangle+\langle\phi(x,w_0),w-w_0\rangle\\&-\beta_1\cdot\sqrt{\Phi_h(x,\theta_0)^\top\Sigma^{-1}\Phi_h(x,\theta_0)}\\&\quad-\beta_2\cdot\sqrt{\phi(x,w_0)^\top\Lambda^{-1}\phi(x,w_0)},H\}^+.\end{aligned}$$

Combining eq. (160) and eq. (161), defining $\mathcal{N}^v_\epsilon=\max_{h\in[H]}\{\mathcal{N}^v_{\epsilon,h}\}$ and letting

$$\epsilon=\sqrt{\lambda_2C_\epsilon}H/(2NC_\phi),\tag{162}$$

$$C_\epsilon\ge1,\tag{163}$$

$$\lambda_2=1+N^{-1},\tag{164}$$

and m be sufficiently large such that

$$\begin{aligned}&\frac{2(L_\sigma A_2+C_\sigma^2A_2^2)C_\phi NH^{3/2}\sqrt{\log HN}}{\sqrt{\lambda_2 m}}\\&\le A_2\sqrt{\frac{\lambda_2H}{d}},\end{aligned}$$

we have the following holds with probability at least $1-N^{-2}H^{-4}$

$$\begin{aligned}&|\langle\phi_h(x_h^\tau,w_0),\overline{w}-\widetilde{w}\rangle|\\&\le\Bigg(2A_2\sqrt{\frac{\lambda_2H}{d}}\\&\quad+\sqrt{2H^2\log\det\left(I+\frac{K^v_{N,h}}{\lambda_2}\right)+3C_\epsilon H^2}\\&\quad+\sqrt{8H^2\log(NH^2\mathcal{N}^v_\epsilon)}\Bigg)b_{v,h}(x,w_0)\\&\le H\Bigg(\frac{8A_2^2\lambda_2}{d}+4\max_{h\in[H]}\left\{\log\det\left(I+\frac{K^v_{N,h}}{\lambda_2}\right)\right\}\\&\quad+6C_\epsilon+16\log(NH^2\mathcal{N}^v_\epsilon)\Bigg)^{1/2}b_{v,h}(x,w_0),\end{aligned}\tag{165}$$

where in the last inequality we use the fact that $a+b\le\sqrt{2(a^2+b^2)}$. Substituting eq. (165) into eq. (150), we conclude that the following holds with probability at least $1-N^{-2}H^{-4}$

$$\left|(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(x)-(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(x)\right|\le\beta_2\cdot b_{v,h}(x,w_0),\tag{166}$$

where

$$\begin{aligned}&\beta_2\\&=H\Bigg(\frac{8A_2^2\lambda_2}{d}+4\max_{h\in[H]}\left\{\log\det\left(I+\frac{K^v_{N,h}}{\lambda_2}\right)\right\}\\&\quad+22\log(NH^2\mathcal{N}^v_\epsilon)\Bigg)^{1/2}.\end{aligned}$$

Next, we proceed to bound the term $\left|(\mathbb{P}_h\widehat{V}_{h+1})(x)-(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(x)\right|$. By the triangle inequality, we have

$$\begin{aligned}&\left|(\mathbb{P}_h\widehat{V}_{h+1})(x)-(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(x)\right|\\&=\big|(\mathbb{P}_h\widehat{V}_{h+1})(x)-(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(x)\\&\quad+(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(x)-(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(x)\\&\quad+(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(x)-(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(x)\big|\\&\le\left|(\mathbb{P}_h\widehat{V}_{h+1})(x)-(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(x)\right|\\&\quad+\left|(\widetilde{\mathbb{P}}_h\widehat{V}_{h+1})(x)-(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(x)\right|\\&\quad+\left|(\overline{\mathbb{P}}_h\widehat{V}_{h+1})(x)-(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(x)\right|\\&\overset{(i)}{\le}\frac{2(L_\sigma A_2+C_\sigma^2A_2^2)\sqrt{\log(N^2H^5)}}{\sqrt{m}}\\&\quad+\beta_2\cdot b_{v,h}(x,w_0)\\&\quad+\mathcal{O}\left(\frac{C_\phi^3H^{4/3}N^{5/3}\sqrt{\log(m)}}{m^{1/6}\lambda_2^{5/3}}\right)\\&\overset{(ii)}{\le}\mathcal{O}\left(\frac{H^{4/3}N^{5/3}\sqrt{\log(N^2H^5m)}}{m^{1/6}}\right)\\&\quad+\beta_2\cdot b_{v,h}(x,w_0),\end{aligned}\tag{167}$$

where (i) follows from eq. (122), eq. (144) and eq. (166) and (ii) follows from the fact that $\lambda_2=1+1/N$ and $L_\sigma,C_\sigma,A_2,C_\phi=\mathcal{O}(1)$.

C. Upper and Lower Bounds on Evaluate Error $\delta_h(\cdot)$

By definition, we have the following holds with probability $1-2N^{-2}H^{-4}$

$$\begin{aligned}&\left|(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)-(\mathbb{B}_h\widehat{V}_{h+1})(x)\right|\\&=\left|\widehat{R}_h(x)+(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(x)-R_h(x)-(\mathbb{P}_h\widehat{V}_{h+1})(x)\right|\\&\le\left|\widehat{R}_h(x)-R_h(x)\right|+\left|(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(x)-(\mathbb{P}_h\widehat{V}_{h+1})(x)\right|\\&\overset{(i)}{\le}\beta_1\cdot b_{r,h}(x,\Theta_0)+\beta_2\cdot b_{v,h}(x,w_0)\\&\quad+\mathcal{O}\left(\frac{H^{4/3}N^{5/3}\sqrt{\log(N^2H^5m)}}{m^{1/6}}\right)\end{aligned}$$

$$
\begin{aligned}
&+\mathcal{O}\left(\frac{H^{17/6}N^{5/3}\sqrt{\log(m)}}{m^{1/6}}\right)\\
&\leq \beta_1\cdot b_{r,h}(x,\Theta_0)+\beta_2\cdot b_{v,h}(x,w_0)\\
&\quad+\mathcal{O}\left(\frac{H^{17/6}N^{5/3}\sqrt{\log(N^2H^5m)}}{m^{1/6}}\right),
\end{aligned}\tag{168}
$$

where (i) follows from eq. (118) and eq. (168). Moreover, by the triangle inequality, eq. (104) and eq. (149), we have the following holds with probability $1-2N^{-2}H^{-4}$

$$
\begin{aligned}
&\beta_1\cdot b_{r,h}(x,\Theta_0)+\beta_2\cdot b_{v,h}(x,w_0)\\
&\leq \beta_1\cdot b_{r,h}(x,\widehat{\Theta})+\beta_2\cdot b_{v,h}(x,\widehat{w})\\
&\quad+\beta_1\cdot\left|b_{r,h}(x,\widehat{\Theta})-b_{r,h}(x,\Theta_0)\right|\\
&\quad+\beta_2\cdot|b_{r,h}(x,\widehat{w})-b_{r,h}(x,w_0)|\\
&\leq \beta_1\cdot b_{r,h}(x,\widehat{\Theta})+\beta_2\cdot b_{v,h}(x,\widehat{w}_h)\\
&\quad+\beta_1\cdot\mathcal{O}\left(\frac{C_\phi^2H^{2/3}N^{1/12}(\log m)^{1/4}}{m^{1/12}\lambda_1^{13/12}}\right)\\
&\quad+\beta_2\cdot\mathcal{O}\left(\frac{C_\phi^2H^{1/6}N^{1/12}(\log m)^{1/4}}{m^{1/12}\lambda_2^{13/12}}\right)\\
&\overset{(i)}{\leq} \beta_1\cdot b_{r,h}(x,\widehat{\Theta})+\beta_2\cdot b_{v,h}(x,\widehat{w}_h)\\
&\quad+\max\{\beta_1H^{2/3},\beta_2H^{1/6}\}\\
&\quad\mathcal{O}\left(\frac{N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right),
\end{aligned}\tag{169}
$$

where (i) follows from the fact that $\lambda_1=\lambda_2=1+1/N$ and $C_\phi=\mathcal{O}(1)$. Substituting eq. (169) into eq. (168), we can obtain

$$
\begin{aligned}
&\left|(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)-(\mathbb{B}_h\widehat{V}_{h+1})(x)\right|\\
&\leq \beta_1\cdot b_{r,h}(x,\widehat{\Theta})+\beta_2\cdot b_{v,h}(x,\widehat{w}_h)\\
&\quad+\max\{\beta_1H^{2/3},\beta_2H^{1/6}\}\\
&\quad\mathcal{O}\left(\frac{N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right)\\
&\quad+\mathcal{O}\left(\frac{H^{17/6}N^{5/3}\sqrt{\log(N^2H^5m)}}{m^{1/6}}\right).
\end{aligned}
$$

Denoting

$$
\begin{aligned}
\varepsilon_b=&\max\{\beta_1H^{2/3},\beta_2H^{1/6}\}\mathcal{O}\left(\frac{N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right)\\
&+\mathcal{O}\left(\frac{H^{17/6}N^{5/3}\sqrt{\log(N^2H^5m)}}{m^{1/6}}\right),
\end{aligned}
$$

we have

$$
\begin{aligned}
&\left|(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)-(\mathbb{B}_h\widehat{V}_{h+1})(x)\right|\\
&\quad\leq \beta_1\cdot b_{r,h}(x,\widehat{\Theta})+\beta_2\cdot b_{v,h}(x,\widehat{w}_h)+\varepsilon_b.
\end{aligned}\tag{170}
$$

Up to this point, we characterize the uncertainty of $(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\cdot)$. Next, we proceed to bound the suboptimality of Algorithm 2. Recalling the construction of $\widehat{Q}_h(x)$ in Algorithm 2, we have

$$
\begin{aligned}
&\widehat{Q}_h(\cdot)\\
&=\min\{(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\cdot)-\beta_1\cdot b_{r,h}(\cdot,\widehat{\Theta})\\
&\quad-\beta_2\cdot b_{v,h}(\cdot,\widehat{w}_h),H\}^+.
\end{aligned}
$$

If $(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)<\beta_1\cdot b_{r,h}(x,\widehat{\Theta})+\beta_2\cdot b_{v,h}(x,\widehat{w}_h)$, we have

$$
\widehat{Q}_h(\cdot)=0.
$$

Note that $\widehat{V}_{h+1}(\cdot)$ is nonnegative. Recalling the definition of $\delta_h(x)$ in eq. (38), we have

$$
\delta_h(x)=(\mathbb{B}_h\widehat{V}_{h+1})(x)-\widehat{Q}_h(x)=(\mathbb{B}_h\widehat{V}_{h+1})(x)>0.
$$

Otherwise, if $(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)>\beta_1\cdot b_{r,h}(x,\widehat{\Theta})+\beta_2\cdot b_{v,h}(x,\widehat{w}_h)$, we have

$$
\begin{aligned}
&\widehat{Q}_h(x)\\
&=\min\{(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)-\beta_1\cdot b_{r,h}(x,\widehat{\Theta})\\
&-\beta_2\cdot b_{v,h}(x,\widehat{w}_h),H\}^+\\
&\leq(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)-\beta_1\cdot b_{r,h}(x,\widehat{\Theta})-\beta_2\cdot b_{v,h}(x,\widehat{w}_h),
\end{aligned}
$$

which implies that

$$
\begin{aligned}
&\delta_h(x)\\
&\geq(\mathbb{B}_h\widehat{V}_{h+1})(x)-\Big[(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)-\beta_1\cdot b_{r,h}(x,\widehat{\Theta})\\
&\quad-\beta_2\cdot b_{v,h}(x,\widehat{w}_h)\Big]\\
&=\Big[(\mathbb{B}_h\widehat{V}_{h+1})(x)-(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)\Big]\\
&\quad+\beta_1\cdot b_{r,h}(x,\widehat{\Theta})+\beta_2\cdot b_{v,h}(x,\widehat{w}_h).
\end{aligned}
$$

Note that eq. (170) implies the followings hold with probability $1-2N^{-2}H^{-4}$

$$
\begin{aligned}
&(\mathbb{B}_h\widehat{V}_{h+1})(x)-(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)\\
&\geq-\beta_1\cdot b_{r,h}(x,\widehat{\Theta})-\beta_2\cdot b_{v,h}(x,\widehat{w}_h)-\varepsilon_b.
\end{aligned}\tag{171}
$$

$$
\begin{aligned}
&(\mathbb{B}_h\widehat{V}_{h+1})(x)-(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(x)\\
&\leq\beta_1\cdot b_{r,h}(x,\widehat{\Theta})+\beta_2\cdot b_{v,h}(x,\widehat{w}_h)+\varepsilon_b.
\end{aligned}\tag{172}
$$

As a result, we have the following holds with probability $1-2N^{-2}H^{-4}$

$$
\delta_h(x)\geq-\varepsilon_b.\tag{173}
$$

It remains to establish the upper bound of $\delta_h(x)$. Considering the event in eq. (172) occurs, we have

$$
\begin{aligned}
&(\widehat{\mathbb{B}}_h\widehat{V}_{h+1})(\cdot)-\beta_1\cdot b_{r,h}(\cdot,\widehat{\Theta})-\beta_2\cdot b_{v,h}(\cdot,\widehat{w}_h)\\
&\leq\Big[(\mathbb{B}_h\widehat{V}_{h+1})(x)+\beta_1\cdot b_{r,h}(x,\widehat{\Theta})\\
&\quad+\beta_2\cdot b_{v,h}(x,\widehat{w}_h)+\varepsilon_b\Big]-\beta_1\cdot b_{r,h}(\cdot,\widehat{\Theta})\\
&\quad-\beta_2\cdot b_{v,h}(\cdot,\widehat{w}_h)\\
&=(\mathbb{B}_h\widehat{V}_{h+1})(x)+\varepsilon_b\leq H+\varepsilon_b,
\end{aligned}
$$

where the last inequality follows from the fact that $R_h(x) \leq 1$ and $\widehat{V}_{h+1}(s) \leq H$ for all $x \in \mathcal{X}$ and $s \in \mathcal{S}$. Hence, we have

$$\begin{aligned}
&\widehat{Q}_h(x) \\
&= \min\{(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x) - \beta_1 \cdot b_{r,h}(x, \widehat{\Theta}) \\
&\quad - \beta_2 \cdot b_{v,h}(x, \widehat{w}_h), H\}^+ \\
&\geq \min\{(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x) - \beta_1 \cdot b_{r,h}(x, \widehat{\Theta}) \\
&\quad - \beta_2 \cdot b_{v,h}(x, \widehat{w}_h) - \varepsilon_b, H\}^+ \\
&= \max\{(\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x) - \beta_1 \cdot b_{r,h}(x, \widehat{\Theta}) \\
&\quad - \beta_2 \cdot b_{v,h}(x, \widehat{w}_h) - \varepsilon_b, 0\} \\
&\geq (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x) - \beta_1 \cdot b_{r,h}(x, \widehat{\Theta}) \\
&\quad - \beta_2 \cdot b_{v,h}(x, \widehat{w}_h) - \varepsilon_b, \qquad (174)
\end{aligned}$$

which by definition of $\delta_h(x)$ implies

$$\begin{aligned}
&\delta_h(x) \\
&= (\mathbb{B}_h \widehat{V}_{h+1})(x) - \widehat{Q}_h(x) \\
&\leq (\mathbb{B}_h \widehat{V}_{h+1})(x) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(x) + \beta_1 \cdot b_{r,h}(x, \widehat{\Theta}) \\
&\quad + \beta_2 \cdot b_{v,h}(x, \widehat{w}_h) + \varepsilon_b \\
&\leq 2\left[\beta_1 \cdot b_{r,h}(x, \widehat{\Theta}) + \beta_2 \cdot b_{v,h}(x, \widehat{w}_h) + \varepsilon_b\right], \qquad (175)
\end{aligned}$$

where the last inequality follows from eq. (172). Combining eq. (173) and eq. (175), with probability $1 - 2N^{-2}H^{-4}$, we have

$$\begin{aligned}
&-\varepsilon_b \\
&\leq \delta_h(x) \\
&\leq 2\left[\beta_1 \cdot b_{r,h}(x, \widehat{\Theta}) + \beta_2 \cdot b_{v,h}(x, \widehat{w}_h) + \varepsilon_b\right],
\end{aligned}$$

which completes the proof.

Appendix G
Proof of Lemma 2

For $\sum_{h=1}^H b_{r,h}(x, \widehat{\Theta})$, we have the following holds with probability $1 - N^{-2}H^{-4}$

$$\begin{aligned}
&\sum_{h=1}^H b_{r,h}(x, \widehat{\Theta}) \\
&\leq \sum_{h=1}^H b_{r,h}(x, \Theta_0) \\
&\quad + \sum_{h=1}^H \left| b_{r,h}(x, \widehat{\Theta}) - b_{r,h}(x, \Theta_0) \right| \\
&\overset{(i)}{\leq} \sum_{h=1}^H b_{r,h}(x, \Theta_0) \\
&\quad + \mathcal{O}\left(\frac{H^{5/3} N^{1/12} (\log m)^{1/4}}{m^{1/12}}\right), \qquad (176)
\end{aligned}$$

where (i) follows from eq. (104). We next proceed to bound the term $\sum_{h=1}^H b_{r,h}(x, \Theta_0)$. Recall that in Assumbtion 4 we define $\overline{M}(\Theta_0) = \mathbb{E}_\mu\left[\Phi(\tau, \Theta_0)\Phi(\tau, \Theta_0)^\top\right]$. For all $\tau \in \mathcal{D}$, we define the following random matrix $\widehat{M}(\Theta_0)$

$$\widehat{M}(\Theta_0) = \sum_{\tau \in \mathcal{D}} A_\tau(\Theta_0), \qquad (177)$$

$$A_\tau(\Theta_0) = \Phi(\tau, \Theta_0)\Phi(\tau, \Theta_0)^\top - \overline{M}(\Theta_0). \qquad (178)$$

Note that eq. (90) implies $\|\Phi(\tau, \Theta_0)\|_2 \leq C_\phi \sqrt{H}$. By Jensen's inequality, we have

$$\begin{aligned}
&\left\|\overline{M}(\Theta_0)\right\|_2 \\
&\leq \mathbb{E}_\mu\left[\left\|\Phi(\tau, \Theta_0)\Phi(\tau, \Theta_0)^\top\right\|_2\right] \\
&\leq C_\phi^2 H. \qquad (179)
\end{aligned}$$

For any vector $v \in \mathbb{R}^{2mdH}$ with $\|v\|_2 = 1$, we have

$$\begin{aligned}
&\|A_\tau(\Theta_0)v\|_2 \\
&\leq \left\|\Phi(\tau, \Theta_0)\Phi(\tau, \Theta_0)^\top v\right\|_2 + \left\|\overline{M}(\Theta_0)v\right\|_2 \\
&\leq \left\|\Phi(\tau, \Theta_0)\Phi(\tau, \Theta_0)^\top\right\|_2 \|v\|_2 \\
&\quad + \left\|\overline{M}(\Theta_0)\right\|_2 \|v\|_2 \\
&\leq 2C_\phi^2 H \|v\|_2 \\
&= 2C_\phi^2 H,
\end{aligned}$$

which implies

$$\begin{aligned}
&\|A_\tau(\Theta_0)\|_2 \leq 2C_\phi^2 H, \qquad (180) \\
&\left\|A_\tau(\Theta_0)A_\tau(\Theta_0)^\top\right\|_2 \\
&\quad \leq \|A_\tau(\Theta_0)\|_2 \left\|A_\tau(\Theta_0)^\top\right\|_2 \\
&\quad \leq 4C_\phi^4 H^2. \qquad (181)
\end{aligned}$$

Since $\{A_\tau(\Theta_0)\}_{\tau \in \mathcal{D}}$ are i.i.d. and $\mathbb{E}[A_\tau(\Theta_0)] = 0$ for all τ, we have

$$\begin{aligned}
&\left\|E_\mu[\widehat{M}(\Theta_0)\widehat{M}(\Theta_0)^\top]\right\|_2 \\
&= \left\|\sum_{\tau \in \mathcal{D}} E_\mu\left[A_\tau(\Theta_0)A_\tau(\Theta_0)^\top\right]\right\|_2 \\
&= N \cdot \left\|E_\mu\left[A_{\tau_1}(\Theta_0)A_{\tau_1}(\Theta_0)^\top\right]\right\|_2 \\
&\overset{(i)}{\leq} N \cdot E_\mu\left[\left\|A_{\tau_1}(\Theta_0)A_{\tau_1}(\Theta_0)^\top\right\|_2\right] \\
&\leq 4C_\phi^4 H^2 \ N,
\end{aligned}$$

where (i) follows from Jensen's inequality. Similarly, we can also obtain

$$\left\|E_\mu[\widehat{M}(\Theta_0)^\top \widehat{M}(\Theta_0)]\right\|_2 \leq 4C_\phi^4 H^2 \ N.$$

Applying Lemma 10 to $\widehat{M}(\Theta_0)$, for any fixed $h \in [H]$ and any $\xi_1 > 0$, we have

$$\begin{aligned}
&\mathrm{P}\left(\left\|\widehat{M}(\Theta_0)\right\|_2 \geq \xi_1\right) \\
&\leq 4mdH \cdot \exp\left(-\frac{\xi_1^2/2}{4C_\phi^4 H^2 N + 2C_\phi^2 H/3 \cdot \xi_1}\right).
\end{aligned}$$

For any $\delta_1 \in (0, 1)$, let

$$\begin{aligned}
\xi_1 &= C_\phi^2 H \sqrt{10N \log\left(\frac{4mdH}{\delta_1}\right)}, \\
N &\geq \frac{40}{9} \log\left(\frac{4mdH}{\delta_1}\right).
\end{aligned}$$

Then, we have

$$\mathrm{P}\left(\left\|\widehat{M}(\Theta_0)\right\|_2 \geq \xi_1\right)$$

$$\leq 4mdH \cdot \exp\left(-\frac{\xi_1^2/2}{4C_\phi^4 H^2 N + 2C_\phi^2 H/3 \cdot \xi_1}\right)$$
$$\leq 4mdH \cdot \exp\left(-\frac{\xi_1^2}{10C_\phi^4 H^2 N}\right) = \delta_1,$$

which implies that the following holds with probability at least $1-\delta_1$ taken with respect to the randomness of $\mathcal{D}$

$$\left\|\widehat{M}(\Theta_0)/N\right\|_2 = \left\|\frac{1}{N}\sum_{\tau\in\mathcal{D}}\Phi(\tau,\Theta_0)\Phi(\tau,\Theta_0)^\top - \overline{M}(\Theta_0)\right\|_2 \leq C_\phi^2 H\sqrt{\frac{10}{N}\log\left(\frac{4mdH}{\delta_1}\right)}. \tag{182}$$

By the definition of $\Sigma(\Theta_0)$, we have

$$\widehat{M}(\Theta_0) = (\Sigma(\Theta_0) - \lambda_1 \cdot I_{2mdH}) - N \cdot \overline{M}(\Theta_0). \tag{183}$$

By Assumbtion 4, there exists an absolute constant $C_\sigma > 0$ such that $\lambda_{\min}(\overline{M}(\Theta_0)) \geq C_\sigma$, which implies that $\left\|\overline{M}(\Theta_0)^{-1}\right\|_2 \leq 1/C_\sigma$. Letting N be sufficiently large such that

$$N \geq \max\left\{\frac{40C_\phi^4 H^2}{C_\sigma^2}, \frac{40}{9}\right\}\log\left(\frac{4mdH}{\delta_1}\right)$$

and combining eq. (182) and eq. (183), we have

$$\begin{aligned}
&\lambda_{\min}(\Sigma(\Theta_0)/N)\\
&= \lambda_{\min}(\overline{M}(\Theta_0) + \widehat{M}(\Theta_0)/N + \lambda_1/N \cdot I_{2mdH})\\
&\geq \lambda_{\min}(\overline{M}(\Theta_0)) - \left\|\widehat{M}(\Theta_0)/N\right\|_2\\
&\geq C_\sigma - C_\phi^2 H\sqrt{\frac{10}{N}\log\left(\frac{4mdH}{\delta}\right)}\\
&\geq C_\sigma/2.
\end{aligned}$$

Hence, the following holds with probability $1-\delta_1$ with respect to randomness of $\mathcal{D}$

$$\left\|\Sigma(\Theta_0)^{-1}\right\|_2 \leq (N\cdot\lambda_{\min}(\Sigma(\Theta_0)/N))^{-1} \leq \frac{2}{NC_\sigma},$$

which implies the following holds for all $x\in\mathcal{X}$ and $h\in[H]$

$$\begin{aligned}
&b_{r,h}(x,\Theta_0)\\
&= \sqrt{\Phi_h(x,\Theta_0)^\top\Sigma^{-1}(\Theta_0)\Phi_h(x,\Theta_0)}\\
&\leq \|\Phi_h(x,\Theta_0)\|_2 \cdot \left\|\Sigma^{-1}(\Theta_0)\right\|_2^{1/2}\\
&\leq \frac{\sqrt{2}C_\phi}{\sqrt{C_\sigma}\sqrt{N}},
\end{aligned} \tag{184}$$

where we use the fact that $\|\Phi_h(x,\Theta_0)\|_2 = \|\phi(x_h^\tau,\theta_0)\|_2 \leq C_\phi$. Substituting eq. (184) into eq. (176), we have

$$\sum_{h=1}^H b_{r,h}(x,\widehat{\Theta}) \leq \frac{\sqrt{2}HC_\phi}{\sqrt{C_\sigma}\sqrt{N}} + \mathcal{O}\left(\frac{H^{5/3}N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right). \tag{185}$$

Next, we proceed to bound the term $\sum_{h=1}^H b_{v,h}(x,\widehat{w}_h)$. According to eq. (149), we have the following holds with probability at least $1-N^{-2}H^{-4}$

$$\begin{aligned}
&\sum_{h=1}^H b_{v,h}(x,\widehat{w}_h)\\
&\leq \sum_{h=1}^H b_{v,h}(x,w_0) + \sum_{h=1}^H |b_{v,h}(x,\widehat{w}_h) - b_{v,h}(x,w_0)|\\
&\overset{(i)}{\leq} \sum_{h=1}^H b_{v,h}(x,w_0) + \mathcal{O}\left(\frac{H^{7/6}N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right).
\end{aligned} \tag{186}$$

We then proceed to bound the summation of the penalty terms $\sum_{h=1}^H b_{v,h}(x,w_0)$. Recall that in Assumbtion 4 we define $\overline{m}_h(w_0) = \mathbb{E}_\mu\left[\phi(x_h^\tau,w_0)\phi(x_h^\tau,w_0)^\top\right]$. For all $h\in[H]$ and $\tau\in\mathcal{D}$, we define the following random matrix $\widehat{m}(w_0)$

$$\widehat{m}_h(w_0) = \sum_{\tau\in\mathcal{D}} B_h^\tau(w_0), \tag{187}$$
$$B_h^\tau(w_0) = \phi(x_h^\tau,w_0)\phi(x_h^\tau,w_0)^\top - \overline{m}_h(w_0). \tag{188}$$

Note that eq. (90) implies $\|\phi(x_h^\tau,w_0)\|_2 \leq C_\phi$. By Jensen's inequality, we have

$$\begin{aligned}
&\|\overline{m}_h(w_0)\|_2\\
&\leq \mathbb{E}_\mu\left[\left\|\phi(x_h^\tau,w_0)\phi(x_h^\tau,w_0)^\top\right\|_2\right]\\
&\leq C_\phi^2.
\end{aligned} \tag{189}$$

For any vector $v\in\mathbb{R}^{2md}$ with $\|v\|_2 = 1$, we have

$$\begin{aligned}
&\|B_h^\tau(w_0)v\|_2\\
&\leq \left\|\phi(x_h^\tau,w_0)\phi(x_h^\tau,w_0)^\top v\right\|_2 + \|\overline{m}_h(w_0)v\|_2\\
&\leq \left\|\phi(x_h^\tau,w_0)\phi(x_h^\tau,w_0)^\top\right\|_2\|v\|_2 + \|\overline{m}_h(w_0)\|_2\|v\|_2\\
&\leq 2C_\phi^2\|v\|_2 = 2C_\phi^2,
\end{aligned}$$

which implies

$$\|B_h^\tau(w_0)\|_2 \leq 2C_\phi^2, \tag{190}$$
$$\begin{aligned}
&\left\|B_h^\tau(w_0)B_h^\tau(w_0)^\top\right\|_2\\
&\leq \|B_h^\tau(w_0)\|_2\left\|B_h^\tau(w_0)^\top\right\|_2 \leq 4C_\phi^4.
\end{aligned} \tag{191}$$

Since $\{B_h^\tau(w_0)\}_{\tau\in\mathcal{D}}$ are i.i.d. and $\mathbb{E}[B_h^\tau(w_0)] = 0$ for all τ, we have

$$\begin{aligned}
&\left\|E_\mu[\overline{m}_h(w_0)\overline{m}_h(w_0)^\top]\right\|_2\\
&= \left\|\sum_{\tau\in\mathcal{D}} E_\mu\left[B_h^\tau(w_0)B_h^\tau(w_0)^\top\right]\right\|_2\\
&= N\cdot\left\|E_\mu\left[B_h^{\tau_1}(w_0)B_h^{\tau_1}(w_0)^\top\right]\right\|_2\\
&\overset{(i)}{\leq} N\cdot E_\mu\left[\left\|B_h^{\tau_1}(w_0)B_h^{\tau_1}(w_0)^\top\right\|_2\right]\\
&\leq 4C_\phi^4 N,
\end{aligned}$$

where (i) follows from Jensen's inequality. Similarly, we can also obtain

$$\left\|E_\mu[\overline{m}_h(w_0)^\top \overline{m}_h(w_0)]\right\|_2 \le 4C_\phi^4 N.$$

Applying Lemma 10 to $\widehat{m}_h(w_0)$, for any fixed $h \in [H]$ and any $\xi_2 > 0$, we have

$$\begin{aligned}&\mathrm{P}\left(\|\widehat{m}_h(w_0)\|_2 \ge \xi_2\right)\\ &\le 4md \cdot \exp\left(-\frac{\xi_2^2/2}{4C_\phi^4 N + 2C_\phi^2/3 \cdot \xi_2}\right).\end{aligned}$$

For any $\delta_2 \in (0,1)$, let

$$\begin{aligned}\xi_2 &= C_\phi^2\sqrt{10N\log\left(\frac{4mdH}{\delta_2}\right)},\\ N &\ge \frac{40}{9}\log\left(\frac{4mdH}{\delta_2}\right).\end{aligned}$$

Then, we have

$$\begin{aligned}&\mathrm{P}\left(\|\widehat{m}_h(w_0)\|_2 \ge \xi_2\right)\\ &\le 4md \cdot \exp\left(-\frac{\xi_2^2/2}{4C_\phi^4 N + 2C_\phi^2/3 \cdot \xi_2}\right)\\ &\le 4md \cdot \exp\left(-\frac{\xi_2^2}{10C_\phi^4 N}\right) = \frac{\delta_2}{H},\end{aligned}$$

which implies that we have the following holds with probability at least $1-\delta_2/H$ taken with respect to the randomness of $\mathcal{D}$

$$\begin{aligned}&\|\widehat{m}_h(w_0)/N\|_2\\ &= \left\|\frac{1}{N}\sum_{\tau\in\mathcal{D}}\phi(x_h^\tau, w_0)\phi(x_h^\tau, w_0)^\top - \overline{m}_h(w_0)\right\|_2\\ &\le C_\phi^2\sqrt{\frac{10}{N}\log\left(\frac{4mdH}{\delta_2}\right)}.\end{aligned} \tag{192}$$

By the definition of $\Lambda_h(w_0)$, we have

$$\widehat{m}_h(w_0) = (\Lambda_h(w_0) - \lambda_2 \cdot I_{2md}) - N \cdot \overline{m}_h(w_0). \tag{193}$$

By Assumbtion 4, there exists an absolute constant $C_\varsigma > 0$ such that $\lambda_{\min}(\overline{m}_h(\Theta_0)) \ge C_\varsigma$, which implies that $\left\|\overline{m}(w_0)^{-1}\right\|_2 \le 1/C_\varsigma$. Letting N be sufficiently large such that

$$N \ge \max\left\{\frac{40C_\phi^4}{C_\varsigma^2}, \frac{40}{9}\right\}\log\left(\frac{4mdH}{\delta_2}\right)$$

and combining eq. (192) and eq. (193), we have

$$\begin{aligned}&\lambda_{\min}(\Lambda_h(w_0)/N)\\ &= \lambda_{\min}(\overline{m}(w_0) + \widehat{m}(w_0)/N + \lambda_1/N \cdot I_{2md})\\ &\ge \lambda_{\min}(\overline{m}(w_0)) - \|\widehat{m}(w_0)/N\|_2\\ &\ge C_\varsigma - C_\phi^2 H\sqrt{\frac{10}{N}\log\left(\frac{4mdH}{\delta_2}\right)}\\ &\ge C_\varsigma/2.\end{aligned}$$

Hence, the following holds with probability $1-\delta_2/H$ with respect to randomness of $\mathcal{D}$

$$\begin{aligned}\left\|\Lambda_h(w_0)^{-1}\right\|_2 &\le \left(N \cdot \lambda_{\min}(\Lambda_h(w_0)/N)\right)^{-1}\\ &\le \frac{2}{NC_\varsigma}.\end{aligned} \tag{194}$$

Taking union bound of eq. (194) over $[H]$, we have the following holds for all $x \in \mathcal{X}$ and $h \in [H]$ with probability $1-\delta_2$

$$\begin{aligned}b_{v,h}(x, w_0) &= \sqrt{\phi_h(x, w_0)^\top \Lambda_h^{-1}(w_0)\phi_h(x, w_0)}\\ &\le \|\phi_h(x, w_0)\|_2 \cdot \left\|\Lambda_h(w_0)^{-1}\right\|_2^{1/2}\\ &\le \frac{\sqrt{2}C_\phi}{\sqrt{C_\varsigma}\sqrt{N}},\end{aligned} \tag{195}$$

where we use the fact that $\|\phi(x_h^\tau, \theta_0)\|_2 \le C_\phi$. Substituting eq. (195) into eq. (186), we have

$$\begin{aligned}&\sum_{h=1}^H b_{v,h}(x, \widehat{w})\\ &\le \frac{\sqrt{2}HC_\phi}{\sqrt{C_\varsigma}\sqrt{N}} + \mathcal{O}\left(\frac{H^{7/6}N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right).\end{aligned} \tag{196}$$

Finally, letting $\delta_1 = N^{-2}H^{-4}/2$ and $\delta_2 = N^{-2}H^{-4}/2$ and combining eq. (185) and eq. (196), we have the following holds with probability $1 - N^{-2}H^{-4}$

$$\begin{aligned}&\beta_1 \cdot \sum_{h=1}^H b_{r,h}(x, \widehat{\Theta}) + \beta_2 \cdot \sum_{h=1}^H b_{v,h}(x, \widehat{w})\\ &\le \left(\frac{\beta_1}{\sqrt{C_\sigma}} + \frac{\beta_2}{\sqrt{C_\varsigma}}\right)\frac{\sqrt{2}HC_\phi}{\sqrt{N}}\\ &\quad + \max\{\beta_1 H^{5/3}, \beta_2 H^{7/6}\} \cdot \mathcal{O}\left(\frac{N^{1/12}(\log m)^{1/4}}{m^{1/12}}\right),\end{aligned}$$

which completes the proof.

APPENDIX H
PROOF OF LEMMA 3

Similarly to the proof of Lemma 1, we first bound the uncertainty of the estimated reward $\widehat{R}_h(\cdot)$ in eq. (67) and then bound the uncertainty of the estimated transition value function $(\widehat{\mathbb{P}}_h\widehat{V}_{h+1})(\cdot)$ in eq. (69).

A. Uncertainty of Estimated Reward $\widehat{R}_h(\cdot)$

Following steps similar to those in the proof of Lemma B.1 in [26], we can obtain

$$\|\Theta^*\|_2 \le H\sqrt{dH} \quad \text{and} \quad \left\|\widehat{\Theta}\right\|_2 \le H\sqrt{dHN/\lambda_1}. \tag{197}$$

For simplicity, we denote $r(\tau) = \sum_{h\in[H]} r(x_h^\tau)$, $R(\tau) = \sum_{h\in[H]} R(x_h^\tau)$ and $\varepsilon(\tau) = R(\tau) - r(\tau)$. Consider the estimation error $R_h(\cdot) - \widehat{R}_h(\cdot)$. We have

$$\begin{aligned}&R_h(x) - \widehat{R}_h(x)\\ &= \langle\phi(x), \theta_h^* - \widehat{\theta}_h\rangle\\ &= \langle\Phi_h(x), \Theta^* - \widehat{\Theta}\rangle\end{aligned}$$

$$
\begin{aligned}
&= \langle \Phi_h(x), \Theta^* \rangle - \Phi_h(x)^\top \Sigma^{-1} \left(\sum_{\tau \in \mathcal{D}} \Phi(\tau) r(\tau) \right) \\
&= \langle \Phi_h(x), \Theta^* \rangle \\
&\quad - \Phi_h(x)^\top \Sigma^{-1} \left(\sum_{\tau \in \mathcal{D}} \Phi(\tau) \Phi(\tau)^\top \Theta^* \right) \\
&\quad + \Phi_h(x)^\top \Sigma^{-1} \left(\sum_{\tau \in \mathcal{D}} \Phi(\tau) \varepsilon(\tau) \right) \\
&= \langle \Phi_h(x), \Theta^* \rangle \\
&\quad - \Phi_h(x)^\top \Sigma^{-1} \left(\Sigma - \lambda_1 \cdot I_{dH} \right) \Theta^* \\
&\quad + \Phi_h(x)^\top \Sigma^{-1} \left(\sum_{\tau \in \mathcal{D}} \Phi(\tau) \varepsilon(\tau) \right) \\
&= -\lambda_1 \cdot \Phi_h(x)^\top \Sigma^{-1} \Theta^* \\
&\quad + \Phi_h(x)^\top \Sigma^{-1} \left(\sum_{\tau \in \mathcal{D}} \Phi(\tau) \varepsilon(\tau) \right). \qquad (198)
\end{aligned}
$$

Applying the triangle inequality to eq. (198), we have

$$
\begin{aligned}
&\left| R_h(x) - \widehat{R}_h(x) \right| \\
&\le \underbrace{\lambda_1 \cdot \left| \Phi_h(x)^\top \Sigma^{-1} \Theta^* \right|}_{(i)} \\
&\quad + \underbrace{\left| \Phi_h(x)^\top \Sigma^{-1} \left(\sum_{\tau \in \mathcal{D}} \Phi(\tau) \varepsilon(\tau) \right) \right|}_{(ii)}. \qquad (199)
\end{aligned}
$$

We then proceed to bound (i) and (ii) separately. For (i), we have

$$
\begin{aligned}
(i) &= \lambda_1 \cdot \left| \Phi_h(x)^\top \Sigma^{-1/2} \Sigma^{-1/2} \Theta^* \right| \\
&\le \lambda_1 \left\| \Phi_h(x) \right\|_{\Sigma^{-1}} \left\| \Theta^* \right\|_{\Sigma^{-1}} \\
&\overset{(i.1)}{\le} H \sqrt{dH\lambda_1} \left\| \Phi_h(x) \right\|_{\Sigma^{-1}}, \qquad (200)
\end{aligned}
$$

where $(i.1)$ follows from eq. (197) and the following inequality

$$
\begin{aligned}
&\left\| \Theta^* \right\|_{\Sigma^{-1}} \\
&= \sqrt{\Theta^{*\top} \Sigma^{-1} \Theta^*} \\
&\le \left\| \Sigma^{-1} \right\|_2^{1/2} \left\| \Theta^* \right\|_2 \\
&\le H \sqrt{dH/\lambda_1}.
\end{aligned}
$$

For (ii), we have

$$
\begin{aligned}
(ii) &= \left| \Phi_h(x)^\top \Sigma^{-1/2} \Sigma^{-1/2} \left(\sum_{\tau \in \mathcal{D}} \Phi(\tau) \varepsilon(\tau) \right) \right| \\
&\le \underbrace{\left\| \sum_{\tau \in \mathcal{D}} \Phi(\tau) \varepsilon(\tau) \right\|_{\Sigma^{-1}}}_{(iii)} \cdot \left\| \Phi_h(x) \right\|_{\Sigma^{-1}}. \qquad (201)
\end{aligned}
$$

Following steps similar to those in eq. (116) and Lemma B.2 in [26], we have the following holds with probability at least $1 - \delta$

$$
(iii) \le H \cdot \sqrt{2 \log(1/\delta) + dH \cdot \log(1 + N/\lambda_1)},
$$

which implies

$$
\begin{aligned}
(ii) &\le H \sqrt{2 \log(1/\delta) + dH \cdot \log(1 + N/\lambda_1)} \\
&\quad \cdot \left\| \Phi_h(x) \right\|_{\Sigma^{-1}}. \qquad (202)
\end{aligned}
$$

Recalling that $b_{r,h}(x) = \left\| \Phi_h(x) \right\|_{\Sigma^{-1}}$ and substituting eq. (202) and eq. (200) into eq. (199), we can obtain

$$
\left| R_h(x) - \widehat{R}_h(x) \right| \le R_{\beta_1} \cdot b_{r,h}(x), \qquad (203)
$$

where R_{β_1} is an absolute constant satisfying

$$
\begin{aligned}
&R_{\beta_1} \\
&\ge H \big(\sqrt{dH\lambda_1} + \\
&\quad \sqrt{2 \log(1/\delta) + dH \cdot \log(1 + N/\lambda_1)} \big).
\end{aligned}
$$

Letting $\lambda_1 = 1$ and $C_{\beta_1} > 0$ be a sufficiently large constant, we can verify that $R_{\beta_1} = C_{\beta_1} H \sqrt{dH \log(N/\delta)}$ satisfies the above inequality.

B. Uncertainty of Estimated Transition Value Function $(\widehat{\mathbb{P}}_h \widehat{V}_{h+1})(\cdot)$

Following steps similar to those in the proof of Lemma B.1 in [26], we can obtain

$$
\left\| w^* \right\|_2 \le H \sqrt{d} \quad \text{and} \quad \left\| \widehat{w} \right\|_2 \le H \sqrt{dN/\lambda_2}. \qquad (204)
$$

Consider the estimation error $(\mathbb{P}_h \widehat{V}_{h+1})(\cdot) - (\widehat{\mathbb{P}}_h \widehat{V}_{h+1})(\cdot)$. For simplicity, we define $\varepsilon_v(x) = (\mathbb{P}_h \widehat{V}_{h+1})(x) - (\widehat{\mathbb{P}}_h \widehat{V}_{h+1})(x)$ for all $x \in \mathcal{X}$. Following steps similar to those in eq. (198), we can obtain

$$
\begin{aligned}
&(\mathbb{P}_h \widehat{V}_{h+1})(x) - (\widehat{\mathbb{P}}_h \widehat{V}_{h+1})(x) \\
&\le -\lambda_2 \cdot \phi(x)^\top \Lambda_h^{-1} w_h^* \\
&\quad + \phi(x)^\top \Lambda_h^{-1} \left(\sum_{\tau \in \mathcal{D}} \phi(x_h^\tau) \varepsilon_v(x_h^\tau) \right). \qquad (205)
\end{aligned}
$$

Applying the triangle inequality to eq. (205), we have

$$
\begin{aligned}
&\left| (\mathbb{P}_h \widehat{V}_{h+1})(x) - (\widehat{\mathbb{P}}_h \widehat{V}_{h+1})(x) \right| \\
&\le \underbrace{\lambda_2 \cdot \left| \phi(x)^\top \Lambda_h^{-1} w_h^* \right|}_{(i)} \\
&\quad + \underbrace{\left| \phi(x)^\top \Lambda_h^{-1} \left(\sum_{\tau \in \mathcal{D}} \phi(x_h^\tau) \varepsilon_v(x_h^\tau) \right) \right|}_{(ii)}. \qquad (206)
\end{aligned}
$$

Following steps similar to those in eq. (199), we can obtain

$$
(i) \le H \sqrt{d\lambda_2} \left\| \phi(x) \right\|_{\Lambda_h^{-1}}. \qquad (207)
$$

For (ii), we have

$$
\begin{aligned}
(ii) &= \left| \phi(x)^\top \Lambda_h^{-1/2} \Lambda_h^{-1/2} \left(\sum_{\tau \in \mathcal{D}} \phi(x_h^\tau) \varepsilon_v(x_h^\tau) \right) \right| \\
&\le \underbrace{\left\| \sum_{\tau \in \mathcal{D}} \phi(x_h^\tau) \varepsilon_v(x_h^\tau) \right\|_{\Lambda_h^{-1}}}_{(iii)} \left\| \phi(x) \right\|_{\Lambda_h^{-1}}. \qquad (208)
\end{aligned}
$$

We then proceed to upper bound the term (iii). Following steps similar to those in eq. (161) and Lemma B.2 in [26], we have the following holds with probability at least $1-\delta$

$$(iii) \leq R_{\beta_2} \left\|\phi(x)\right\|_{\Lambda_h^{-1}} \tag{209}$$

where R_{β_2} is an absolute constant satisfying

$$R_{\beta_2} \geq 2H \cdot \sqrt{\log(H \cdot \mathcal{N}^v_{\epsilon,h}/\delta) + d \cdot \log(1+N/\lambda_2) + 8\epsilon^2 N^2/\lambda_2}, \tag{210}$$

and $\mathcal{N}^v_{\epsilon,h}$ is the cardinality of the following function class

$$\begin{aligned}&\mathcal{V}_h(x, R_\theta, R_w, R_{\beta_1}, R_{\beta_2}, \lambda_1, \lambda_2)\\ &= \{\max_{a\in\mathcal{A}}\{\overline{Q}_h(s,a)\} : \mathcal{S} \to [0,H]\\ &\text{with } \left\|\Theta\right\|_2 \leq R_\theta, \left\|w\right\|_2 \leq R_w, \beta_1 \in [0, R_{\beta_1}], \beta_2 \in [0, R_{\beta_2}],\\ &\left\|\Sigma\right\|_2 \geq \lambda_1, \left\|\Lambda\right\|_2 \geq \lambda_2\},\end{aligned}$$

where $R_\theta = H\sqrt{dHN/\lambda_1}$, $R_w = H\sqrt{dN/\lambda_2}$, and

$$\begin{aligned}\overline{Q}_h(x) = \min\{&\langle \Phi_h(x), \Theta\rangle + \langle \phi(x), w\rangle\\ &- \beta_1 \cdot \sqrt{\Phi_h(x)^\top \Sigma^{-1} \Phi_h(x)}\\ &- \beta_2 \cdot \sqrt{\phi(x)^\top \Lambda^{-1} \phi(x)}, H-h+1\}^+.\end{aligned}$$

Then, following steps similar to those in Section D, we have

$$\begin{aligned}&\left|\max_{a\in\mathcal{A}}\{\overline{Q}_h(s,a,\theta,w,\beta_1,\beta_2,\Sigma,\Lambda)\}\right.\\ &\quad\left.- \max_{a\in\mathcal{A}}\{\overline{Q}_h(s,a,\theta',w',\beta_1',\beta_2',\Sigma',\Lambda')\}\right|\\ &\leq \max_{a\in\mathcal{A}} |\langle \Phi_h(x), \Theta-\Theta'\rangle|\\ &\quad + \max_{a\in\mathcal{A}} |\langle \phi(x), w-w'\rangle| + \frac{1}{\sqrt{\lambda_1}}|\beta_1-\beta_1'|\\ &\quad + \frac{1}{\sqrt{\lambda_2}}|\beta_2-\beta_2'|\\ &\quad + R_{\beta_1} \max_{a\in\mathcal{A}} |\left\|\Phi_h(x)\right\|_{\Sigma^{-1}} - \left\|\Phi_h(x)\right\|_{\Sigma'^{-1}}|\\ &\quad + R_{\beta_2} \max_{a\in\mathcal{A}} |\left\|\phi(x)\right\|_{\Lambda^{-1}} - \left\|\phi(x)\right\|_{\Lambda'^{-1}}|\\ &\overset{(i)}{\leq} \left\|\Theta-\Theta'\right\|_2 + \left\|w-w'\right\|_2 + |\beta_1-\beta_1'|\\ &\quad + |\beta_2-\beta_2'|\\ &\quad + R_{\beta_1}\sqrt{\left\|\Sigma^{-1}-\Sigma'^{-1}\right\|_F}\\ &\quad + R_{\beta_2}\sqrt{\left\|\Lambda^{-1}-\Lambda'^{-1}\right\|_F},\end{aligned} \tag{211}$$

where (i) follows from the fact that $\left\|\phi(x)\right\|_2 \leq 1$ and $\lambda_1, \lambda_2 \geq 1$. Following arguments similar to those used to obtain eq. (60) and applying Lemma 8.6 in [55], we have

$$\begin{aligned}&\log \mathcal{N}^v_{\epsilon,h}\\ &\overset{(i)}{\leq} \mathcal{N}(\epsilon/6, \mathbb{R}^{dH}, R_\theta) + \mathcal{N}(\epsilon/6, \mathbb{R}^d, R_w)\\ &\quad + \mathcal{N}(\epsilon/6, R_{\beta_1}) + \mathcal{N}(\epsilon/6, R_{\beta_2})\\ &\quad + \mathcal{N}(\epsilon^2/(36R_{\beta_1}^2), \mathcal{F}, \sqrt{dH}/\lambda_1)\\ &\quad + \mathcal{N}(\epsilon^2/(36R_{\beta_2}^2), \mathcal{F}, \sqrt{d}/\lambda_2)\\ &\overset{(ii)}{\leq} dH\log(1+12R_\theta/\epsilon) + d\log(1+12R_w/\epsilon)\\ &\quad + \log(1+12R_{\beta_1}/\epsilon) + \log(1+12R_{\beta_2}/\epsilon)\\ &\quad + d^2H^2\log(1+36R_{\beta_1}^2\sqrt{dH}/\epsilon^2)\\ &\quad + d^2\log(1+36R_{\beta_2}^2\sqrt{d}/\epsilon^2)\\ &\overset{(iii)}{\leq} dH\log(1+12H\sqrt{dHN}/\epsilon)\\ &\quad + d\log(1+12H\sqrt{dN}/\epsilon)\\ &\quad + \log(1+12C_{\beta_1}H\sqrt{dH\log(N/\delta)}/\epsilon)\\ &\quad + \log(1+12R_{\beta_2}/\epsilon)\\ &\quad + d^2H^2\log(1+36C_{\beta_1}^2dH^3\sqrt{dH}\log(N/\delta)/\epsilon^2)\\ &\quad + d^2\log(1+36R_{\beta_2}^2\sqrt{d}/\epsilon^2)\\ &\overset{(iv)}{\lesssim} C_1d^2H^2\log(d^{3/2}H^{7/2}N^{1/2}/\epsilon^2)\\ &\quad + C_2d^2\log(R_{\beta_2}^2\sqrt{d}/\epsilon^2),\end{aligned} \tag{212}$$

where in (i) we use $\mathcal{N}(\epsilon, \mathbb{R}^d, B)$ to denote the ϵ-covering of ball with radius B in the space $\mathbb{R}^d$, $\mathcal{N}(\epsilon, B)$ to denote the ϵ-covering of interval $[0,B]$, and $\mathcal{N}(\epsilon, \mathcal{F}, B)$ to denote the ϵ-covering of the function class $\mathcal{F} = \{M : \left\|M\right\|_F \leq B\}$, (ii) follows from Lemma. 8.6 in [55], (iii) follows from the definition of R_θ, R_w and R_{β_1}, and in (iv) we let C_1 and C_2 be sufficiently large and waive the $\log(\log(\cdot))$ term.

Substituting eq. (212) into eq. (209), we can obtain

$$\begin{aligned}&2H \cdot \sqrt{\log(H \cdot \mathcal{N}^v_{\epsilon,h}/\delta) + d \cdot \log(1+N/\lambda_2) + 8\epsilon^2 N/\lambda_2}\\ &\leq 2H \cdot \left(\sqrt{\log(H/\delta)} + \sqrt{\log \mathcal{N}^v_{\epsilon,h}}\right.\\ &\quad \left. + \sqrt{d\cdot\log(1+N)} + \sqrt{8\epsilon^2N^2}\right)\\ &\leq 2H \cdot \left(\sqrt{\log(H/\delta)}\right.\\ &\quad + \sqrt{C_1d^2H^2\log(d^{3/2}H^{7/2}N^{1/2}/\epsilon^2)}\\ &\quad + \sqrt{C_2d^2\log(R_{\beta_2}^2\sqrt{d}/\epsilon^2)}\\ &\quad \left. + \sqrt{d\cdot\log(1+N)} + \sqrt{8\epsilon^2N^2}\right).\end{aligned} \tag{213}$$

Letting $\epsilon = (dH)^{1/4}/N$, we can see that when $R_{\beta_2} = C_{\beta_2}dH^2\sqrt{\log(dH^3N^{5/2}/\delta)}$, where C_{β_2} is a sufficiently large constant, we have

$$R_{\beta_2} \geq \text{R.H.S of eq. (213)},$$

which satisfies the inequality in eq. (210).

C. Upper and Lower Bounds on Evaluation Error $\delta_h(\cdot)$

Using the properties that we obtained from Section H-A & H-B and following steps similar to those in Section F-C, we can obtain

$$0 \leq \delta_h(x) \leq 2\left[\beta_1 \cdot b_{r,h}(x) + \beta_2 \cdot b_{v,h}(x)\right],$$

where $\beta_1 = R_{\beta_1} = C_{\beta_1}H\sqrt{dH\log(N/\delta)}$ and $R_{\beta_2} = C_{\beta_2}dH^2\sqrt{\log(dH^3N^{5/2}/\delta)}$.

APPENDIX I
SUPPORTING LEMMAS FOR OVERPARAMETERIZED NEURAL NETWORKS

The following lemma shows that an infinite-width neural network can be well-approximated by a finite-width neural network.

Lemma 4 (Approximation by Finite Sum): Let $g(x) = \int_{\mathbb{R}^d} \sigma'(w^\top x)x^\top \ell(w)dp(w) \in \mathcal{F}_{g_1,g_2}$. Then for any $\epsilon > 0$, with probability at least $1-\epsilon$ over $w_1, \cdots, w_m$ drawn i.i.d. from $N(0, I_d/d)$, there exist $\ell_1, \cdots, \ell_m$ where $\ell_i \in \mathbb{R}^d$ and $\|\ell_i\|_2 \le g_2/\sqrt{dm}$ for all $i \in [m]$ such that the function $\widehat{g}(x) = (1/\sqrt{m})\sum_{i=1}^m \sigma'(w_i^\top x)x^\top \ell_i$ satisfies

$$\sup_x |g(x) - \widehat{g}(x)| \le \frac{2L_\sigma g_2}{\sqrt{m}} + \frac{\sqrt{2}C_\sigma^2 g_2^2}{\sqrt{m}}\sqrt{\log\left(\frac{1}{\delta}\right)}$$

with probability at least $1-\delta$.

Proof: The proof of Lemma 4 follows from the proof of Proposition C.1 in [68] with some modifications. In Lemma 4 we consider a different distribution of w_i and upper bound on $\|\ell_i\|_2$ from those in [68]. First, we define the following random variable

$$a(w_1, \cdots, w_m) = \sup_x |g(x) - \widehat{g}(x)|.$$

Then, we proceed to show that $a(\cdot)$ is robust to the perturbation of one of its arguments. Let $\ell_i = \ell(w_i)/(\sqrt{dm}p(w_i))$. For $w_1, \cdots, w_m$ and $\tilde{w}_i$ $(1 \le i \le m)$, we have

$$\begin{aligned}
&|a(w_1, \cdots, w_m) - a(w_1, \cdots, \tilde{w}_i, \cdots, w_m)| \\
&= \frac{1}{\sqrt{dm}} \left|\sigma'(w_i^\top x)x^\top \ell_i - \sigma'(\tilde{w}_i^\top x)x^\top \ell_i\right| \\
&= \frac{1}{\sqrt{dm}} \left|\frac{\sigma'(w_i^\top x)x^\top \ell(w_i)}{p(w_i)} - \frac{\sigma'(\tilde{w}_i^\top x)x^\top \ell(\tilde{w}_i)}{p(\tilde{w}_i)}\right| \\
&\le \frac{1}{\sqrt{dm}} \sup_{x\in\mathcal{X}} \left|\frac{\sigma'(w_i^\top x)x^\top \ell(w_i)}{p(w_i)} - \frac{\sigma'(\tilde{w}_i^\top x)x^\top \ell(\tilde{w}_i)}{p(\tilde{w}_i)}\right| \\
&\le \frac{1}{\sqrt{dm}} \sup_{x\in\mathcal{X}} \left(\left|\frac{\sigma'(w_i^\top x)x^\top \ell(w_i)}{p(w_i)}\right| + \left|\frac{\sigma'(\tilde{w}_i^\top x)x^\top \ell(\tilde{w}_i)}{p(\tilde{w}_i)}\right|\right) \\
&\le \frac{1}{\sqrt{dm}} \sup_{x\in\mathcal{X}} \left(\left\|\sigma'(w_i^\top x)x\right\| \left\|\frac{\ell(w_i)}{p(w_i)}\right\|_2 \right. \\
&\quad \left. + \left\|\sigma'(\tilde{w}_i^\top x)x\right\|_2 \left\|\frac{\ell(\tilde{w}_i)}{p(\tilde{w}_i)}\right\|_2\right) \\
&\le \frac{2C_\sigma g_2}{\sqrt{dm}} = \zeta,
\end{aligned}$$

where the last inequality follows from the facts that $\|x\|_2 = 1$, $|\sigma'(\cdot)| \le C_\sigma$ and $\sup_x \|\ell(w)/p(w)\|_2 \le g_2$. Then, we proceed to bound the expectation of $a(\cdot)$. Note that our choice of ℓ_i ensures that $\sqrt{d}\cdot\mathbb{E}_{w_1,\cdots,w_m}\widehat{g}_h(\cdot) = g(\cdot)$. By symmetrization, we have

$$\begin{aligned}
\mathbb{E}a &= \sqrt{d}\cdot\mathbb{E}\sup_{x\in\mathcal{X}} |\widehat{g}(x) - \mathbb{E}\widehat{g}(x)| \\
&\le \frac{2\sqrt{d}}{\sqrt{m}}\cdot\mathbb{E}_{w,\varepsilon}\sup_{x\in\mathcal{X}}\left|\sum_{i=1}^m \varepsilon_i\sigma'(w_i^\top x)x^\top\ell_i\right|,
\end{aligned}$$

where $\{\varepsilon_i\}_{i\in[m]}$ are a sequence of Rademacher random variables. Since $|x^\top\ell_i| \le \|\ell_i\|_2 \le g_2/\sqrt{m}$ and $\sigma'(\cdot)$ is L_σ-Lipschitz, we have that the function $b(\cdot) = \sigma'(\cdot)x^\top\ell_i$ is $(L_\sigma g_2/\sqrt{m})$-Lipschitz. We then proceed as follows

$$\begin{aligned}
\mathbb{E}a &\le \frac{2\sqrt{d}}{\sqrt{m}}\cdot\mathbb{E}_{w,\varepsilon}\sup_{x\in\mathcal{X}}\left|\sum_{i=1}^m \varepsilon_i\sigma'(w_i^\top x)x^\top\ell_i\right| \\
&\overset{(i)}{\le} \frac{2\sqrt{d}L_\sigma g_2}{m}\cdot\mathbb{E}_{w,\varepsilon}\sup_{x\in\mathcal{X}}\left|\left(\sum_{i=1}^m \varepsilon_i w_i\right)^\top x\right| \\
&\overset{(ii)}{\le} \frac{2\sqrt{d}L_\sigma g_2}{m}\cdot\mathbb{E}_w\left\|\sum_{i=1}^m \varepsilon_i w_i\right\|_2 \\
&\overset{(iii)}{\le} \frac{2\sqrt{d}L_\sigma g_2}{\sqrt{m}}\cdot\sqrt{\mathbb{E}_{w\sim N(0,I_d/d)}\|w\|_2^2} \\
&= \frac{2L_\sigma g_2}{\sqrt{m}},
\end{aligned}$$

where (i) follows from Talagrand's Lemma (Lemma 5.7) in [78], (ii) follows from the fact that $\|x\|_2 = 1$ for all $x \in \mathcal{X}$ and Cauchy-Schwartz inequality and (iii) follows from Jensen's inequality. Then, applying McDiarmid's inequality, we can obtain

$$\begin{aligned}
&\mathrm{P}\left(a \ge \frac{2L_\sigma g_2}{\sqrt{m}} + \epsilon\right) \\
&\le \mathrm{P}(a \ge \mathbb{E}a + \epsilon) \\
&\le \exp\left(-\frac{2\epsilon^2}{m\zeta^2}\right) = \exp\left(-\frac{m\epsilon^2}{2C_\sigma^2 g_2^2}\right).
\end{aligned}$$

Letting $\epsilon = \frac{\sqrt{2}C_\sigma^2 g_2^2}{\sqrt{m}}\sqrt{\log\left(\frac{1}{\delta}\right)}$, we have

$$\mathrm{P}\left(a \ge \frac{2L_\sigma g_2}{\sqrt{m}} + \frac{\sqrt{2}C_\sigma^2 g_2^2}{\sqrt{m}}\sqrt{\log\left(\frac{1}{\delta}\right)}\right) \le \delta,$$

which completes the proof. □

The following lemma bounds the perturbed gradient and value of local linearization of overparameterized neural networks around the initialization, which is provided as Lemma C.2 in [71].

Lemma 5: Consider the overparameterized neural network. Consider any fixed input $x \in \mathcal{X}$. Let $R \le c\sqrt{m}/(\log m)^3$ for some sufficiently small constant c. Then, with probability at least $1-m^{-2}$ over the random initialization, we have for any $w \in \mathcal{B}(w_0, R)$, where $\mathcal{B}(w_0, R)$ denotes the Euclidean ball centred at w_0 with radius R, the followings hold

$$\|\phi(x,w)\|_2 \le C_\phi, \tag{214}$$

$$\|\phi(x,w) - \phi(x,w_0)\|_2 \le \mathcal{O}\left(C_\phi\left(\frac{R}{\sqrt{m}}\right)^{1/3}\sqrt{\log m}\right), \tag{215}$$

$$\left|f(x,w)-\langle\phi(x,w_0)^\top(w-w_0)\rangle\right| \le \mathcal{O}\left(C_\phi\left(\frac{R^4}{\sqrt{m}}\right)^{1/3}\sqrt{\log m}\right), \tag{216}$$

where $C_\phi=\mathcal{O}(1)$ is a constant independent from m and d.

Proof: Please see Lemma C.2 in [71] for a detailed proof, which is based on Lemma F.1, F.2 in [72], Lemma A.5, A.6 in [68] and Theorem 1 in [79]. □

APPENDIX J
SUPPORTING LEMMAS FOR RKHS

In this section, we provide some useful lemmas for general RKHS. Consider a variable space $\mathcal{X}$. Given a mapping $\phi(\cdot):\mathcal{X}\to\mathbb{R}^d$, we can assign a feature vector $\phi(x)\in\mathbb{R}^d$ for each $x\in\mathcal{X}$. We further define a kernel function $K(\cdot,\cdot):\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ as $K(x,x')=\phi(x)^\top\phi(x')$ for any $x,x'\in\mathcal{X}$. Let $\mathcal{H}$ be a RKHS defined on $\mathcal{X}$ with the kernel function $K(\cdot,\cdot)$. Let $\langle\cdot,\cdot\rangle_{\mathcal{H}}:\mathcal{H}\times\mathcal{H}\to\mathbb{R}$ and $\|\cdot\|_{\mathcal{H}}:\mathcal{H}\to\mathbb{R}$ denote the inner product and RKHS norm on $\mathcal{H}$, respectively. Since $\mathcal{H}$ is a RKHS, there exists a feature mapping $\psi(\cdot):\mathcal{X}\to\mathcal{H}$, such that $f(x)=\langle f(\cdot),\psi(x)\rangle_{\mathcal{H}}$ for all $f\in\mathcal{H}$ and all $x\in\mathcal{X}$. Moreover, for any $x,x'\in\mathcal{X}$ we have $K(x,x')=\langle\psi(x),\psi(x')\rangle_{\mathcal{H}}$. Without loss of generality, we further assume $\|\phi(x)\|_2\le C_\phi$ and $\|\psi(x)\|_{\mathcal{H}}\le C_\psi$ for all $x\in\mathcal{X}$.

Let $\mathcal{L}^2(\mathcal{X})$ be the space of square-integrable functions on $\mathcal{X}$ with respect to the Lebesgue measure and let $\langle\cdot,\cdot\rangle_{\mathcal{L}^2}$ be the inner product on $\mathcal{L}^2(\mathcal{X})$. The kernel function $K(\cdot,\cdot)$ induces an integral operator $T_K:\mathcal{L}^2(\mathcal{X})\to\mathcal{L}^2(\mathcal{X})$ defined as

$$T_Kf(z)=\int_{\mathcal{X}}K(x,x')\cdot f(x')dx',\quad \forall f\in\mathcal{L}^2(\mathcal{X}). \tag{217}$$

Consider the kernel function $K(\cdot,\cdot)$ of the RHKS $\mathcal{H}$. Let $\{x_i\}_{i=1}^\infty\subset\mathcal{X}$ be a discrete time stochastic process that is adapted to a filtration $\{\mathcal{F}_t\}_{i=0}^\infty$, i.e., x_i is $\mathcal{F}_{i-1}$ measurable for all $i\ge1$. We define the Gram matrix $K_N\in\mathbb{R}^{N\times N}$ and function $k_N(\cdot):\mathcal{X}\to\mathbb{R}^N$ as

$$K_N=[K(x_i,x_j)]_{i,j\in[N]}\in\mathbb{R}^{N\times N}, \tag{218}$$

$$k_N(x)=[K(x_1,x),\cdots,K(x_N,x)]^\top\in\mathbb{R}^N. \tag{219}$$

Note that K_N and $k_N(x)$ can also be expressed as

$$K_N=\Phi\Phi^\top=\Psi\Psi^\top\in\mathbb{R}^{N\times N},$$
$$k_N(x)=\Phi\phi(x)=\Psi\psi(x)\in\mathbb{R}^{N\times1},$$

where $\Phi=[\phi(x_1),\cdots,\phi(x_N)]^\top\in\mathbb{R}^{N\times d}$ and $\Psi=[\psi(x_1),\cdots,\psi(x_N)]^\top\in\mathbb{R}^{N\times\infty}$. Given a regularization parameter $\lambda>1$, we define the matrix Ω_N based on Φ and an operator Υ_N in RKHS $\mathcal{H}$ based on Ψ as

$$\Omega_N=\Phi^\top\Phi+\lambda\cdot I_d, \tag{220}$$

$$\Upsilon_N=\Psi^\top\Psi+\lambda\cdot I_{\mathcal{H}}. \tag{221}$$

We next provide some fundamental properties for the RKHS $\mathcal{H}$.

Lemma 6: For any $x\in\mathcal{X}$, considering K_N, $k_N(\cdot)$, Ω_N and Υ_N defined in eq. (219) and eq. (221), we have the followings hold

$$\Phi^\top(K_N+I_N)^{-1}=\Omega_N^{-1}\Phi^\top, \tag{222}$$

$$\Psi^\top(K_N+I_N)^{-1}=\Upsilon_N^{-1}\Psi^\top, \tag{223}$$

$$\begin{aligned}&\phi(x)^\top\Omega_N^{-1}\phi(x)\\ &\overset{(i)}{=}\frac{1}{\lambda}\Big[K(x,x)\\ &\quad-k_N(x)^\top(K_N+\lambda\cdot I_N)^{-1}k_N(x)\Big]\\ &\overset{(ii)}{=}\psi(x)^\top\Upsilon_N^{-1}\psi(x).\end{aligned} \tag{224}$$

Proof: The result in Lemma 6 can be obtained from steps spread out in [71]. We provide a detailed proof here for completeness.

We first proceed to prove eq. (222) and (i) in eq. (224). According to the definition of Σ_N, we have

$$\begin{aligned}&\Omega_N\Phi^\top\\ &=\Phi^\top\Phi\Phi^\top+\lambda\Phi^\top\\ &=\Phi^\top(\Phi\Phi^\top+\lambda I_N)\\ &=\Phi^\top(K_N+I_N).\end{aligned}$$

Multiplying Ω_N^{-1} on both sides of the above equality yields

$$\Phi^\top=\Omega_N^{-1}\Phi^\top(K_N+I_N),$$

which implies eq. (222) as follows

$$\Phi^\top(K_N+I_N)^{-1}=\Omega_N^{-1}\Phi^\top. \tag{225}$$

We next proceed as follows

$$\begin{aligned}\phi(x)&=\Omega_N^{-1}\Omega_N\phi(x)\\ &=\Omega_N^{-1}(\Phi^\top\Phi+\lambda\cdot I_d)\phi(x)\\ &=(\Omega_N^{-1}\Phi^\top)\Phi\phi(x)+\lambda\Omega_N^{-1}\phi(x)\\ &\overset{(i)}{=}\Phi^\top(K_N+I_N)^{-1}\Phi\phi(x)\\ &\quad+\lambda\Omega_N^{-1}\phi(x),\end{aligned} \tag{226}$$

where (i) follows from eq. (225). Taking inter product with $\phi(x)$ on both sides of eq. (226) yields

$$\begin{aligned}&K(x,x)\\ &=\phi(x)^\top\phi(x)\\ &=\phi(x)^\top\Phi^\top(K_N+I_N)^{-1}\Phi\phi(x)+\lambda\phi(x)^\top\Omega_N^{-1}\phi(x)\\ &=k_N(x)^\top(K_N+I_N)^{-1}k_N(x)+\lambda\phi(x)^\top\Omega_N^{-1}\phi(x),\end{aligned}$$

which implies

$$\begin{aligned}\phi(x)^\top\Omega_N^{-1}\phi(x)=\frac{1}{\lambda}\Big[&K(x,x)\\ &-k_N(x)^\top(K_N+I_N)^{-1}k_N(x)\Big].\end{aligned} \tag{227}$$

We next proceed to prove eq. (223) and (ii) in eq. (224). According to the definition of Υ_N, we have

$$\begin{aligned}&\Upsilon_N\Psi^\top\\ &=\Psi^\top\Psi\Psi^\top+\lambda\Psi^\top\\ &=\Psi^\top(\Psi\Psi^\top+I_N)\\ &=\Psi^\top(K_N+I_N).\end{aligned} \tag{228}$$

Multiplying Υ_N^{-1} on both sides of the above equality yields

$$\Psi^\top = \Upsilon_N^{-1}\Psi^\top(K_N + I_N),$$

which further implies eq. (223) as follows

$$\Psi^\top(K_N + I_N)^{-1} = \Upsilon_N^{-1}\Psi^\top. \tag{229}$$

We next proceed as follows

$$\begin{aligned}\psi(x) &= \Upsilon_N^{-1}\Upsilon_N\psi(x)\\ &= \Upsilon_N^{-1}(\Psi^\top\Psi + \lambda\cdot I_{\mathcal{H}})\psi(x)\\ &= (\Upsilon_N^{-1}\Psi^\top)\Psi\psi(x) + \lambda\Upsilon_N^{-1}\psi(x)\\ &\overset{(i)}{=} \Psi^\top(K_N + I_N)^{-1}\Psi\psi(x) + \lambda\Upsilon_N^{-1}\psi(x),\end{aligned} \tag{230}$$

where (i) follows from eq. (229). Taking inter product with $\psi(x)$ on both sides of eq. (230) yields

$$\begin{aligned}&K(x,x)\\ &= \langle\psi(x),\psi(x)\rangle_{\mathcal{H}}\\ &= \psi(x)^\top\Psi^\top(K_N + I_N)^{-1}\Psi\psi(x)\\ &\quad + \lambda\psi(x)^\top\Upsilon_N^{-1}\psi(x)\\ &= k_N(x)^\top(K_N + I_N)^{-1}k_N(x)\\ &\quad + \lambda\psi(x)^\top\Upsilon_N^{-1}\psi(x),\end{aligned}$$

which implies

$$\begin{aligned}&\psi(x)^\top\Upsilon_N^{-1}\psi(x)\\ &= \frac{1}{\lambda}\Big[K(x,x)\\ &\quad - k_N(x)^\top(K_N + I_N)^{-1}k_N(x)\Big].\end{aligned} \tag{231}$$

Combining eq. (228) and eq. (231) completes the proof. □

The following two lemmas characterize the concentration property of self-normalized processes.

Lemma 7 (Concentration of Self-Normalized Process in RKHS [80]):

Let $\{\varepsilon_i\}_{i=1}^\infty$ be a real-valued stochastic process such that (i) $\epsilon_i \in \mathcal{F}_t$ and (ii) ϵ_i is zero-mean and σ-sub-Gaussian conditioned on $\mathcal{F}_{i-1}$ satisfying $\forall\kappa\in\mathbb{R}$

$$\mathbb{E}\left[\varepsilon_i|\mathcal{F}_{i-1}\right] = 0, \tag{232}$$

$$\mathbb{E}\left[e^{\kappa\varepsilon_i} \le e^{\kappa^2\sigma^2/2}|\mathcal{F}_{i-1}\right]. \tag{233}$$

Moreover, for any $t \ge 2$, let $E_N = [\varepsilon_1,\cdots,\varepsilon_{N-1}]^\top \in \mathbb{R}^{N-1}$.

For any $\eta > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, we have the following holds simultaneously for all $N \ge 1$:

$$\begin{aligned}&E_N^\top\left[(K_N + \eta\cdot I_{N-1})^{-1} + I_{N-1}\right]^{-1}E_N\\ &\le \sigma^2\cdot\log det[(1+\eta)\cdot I_{N+1} + K_N]\\ &\quad + 2\sigma^2\cdot\log(1/\delta).\end{aligned}$$

Moreover, if K_N is positive definite for all $N \ge 2$ with probability one, then the above inequality also holds with $\eta = 0$.

Lemma 8: Let $\mathcal{G} \subset \{G : \mathcal{X} \to [0, C_g]\}$ be a class of bounded functions on $\mathcal{X}$. Let $\mathcal{G}_\epsilon \subset \mathcal{G}$ be the minimal ϵ-cover of $\mathcal{G}$ such that $\mathcal{N}_\epsilon = |\mathcal{G}_\epsilon|$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}&\sup_{G\in\mathcal{G}}\left\|\sum_{i=1}^N\phi(x_i)\left(G(x_i) - \mathbb{E}\left[G(x_i)|\mathcal{F}_{i-1}\right]\right)\right\|^2_{\Omega_N^{-1}}\\ &\le 2C_g^2\log\det(I + K_N/\lambda)\\ &\quad + 2C_g^2N(\lambda-1) + 4C_g^2\log(\mathcal{N}_\epsilon/\delta)\\ &\quad + 8N^2C_\phi^2\epsilon^2/\lambda.\end{aligned} \tag{234}$$

Moreover, if $G(\cdot)$ does not depend on $\{x_i\}_{i\in[N]}$, we have

$$\begin{aligned}&\left\|\sum_{i=1}^N\phi(x_i)\left(G(x_i) - \mathbb{E}\left[G(x_i)|\mathcal{F}_{i-1}\right]\right)\right\|^2_{\Omega_N^{-1}}\\ &\le C_g^2\log\det(I + K_N/\lambda)\\ &\quad + C_g^2N(\lambda-1) + 2C_g^2\log(1/\delta).\end{aligned} \tag{235}$$

Proof: The proof is adapted but different from the proof of Lemma E.2 in [71]. We first proceed to prove eq. (234) and will show that eq. (235) can be obtained as a by-product of proving eq. (234). For any $G \in \mathcal{G}$, there exists a function G' in $\mathcal{G}_\epsilon$ such that $\sup_{x\in\mathcal{X}}|G(x) - G'(x)| \le \epsilon$. Denote $\Delta_G(x) = G(x) - G'(x)$. We have the following holds

$$\begin{aligned}&\left\|\sum_{i=1}^N\phi(x_i)\left(G(x_i) - \mathbb{E}\left[G(x_i)|\mathcal{F}_{i-1}\right]\right)\right\|^2_{\Omega_N^{-1}}\\ &\le 2\left\|\sum_{i=1}^N\phi(x_i)\left(G'(x_i) - \mathbb{E}\left[G'(x_i)|\mathcal{F}_{i-1}\right]\right)\right\|^2_{\Omega_N^{-1}}\\ &\quad + 2\left\|\sum_{i=1}^N\phi(x_i)(\Delta_G(x_i)\right.\\ &\quad \left.- \mathbb{E}\left[\Delta_G(x_i)|\mathcal{F}_{i-1}\right])\right\|^2.\end{aligned} \tag{236}$$

For the second term on the right hand side of eq. (236), we have

$$\begin{aligned}&\left\|\sum_{i=1}^N\phi(x_i)\left(\Delta_G(x_i) - \mathbb{E}\left[\Delta_G(x_i)|\mathcal{F}_{i-1}\right]\right)\right\|^2_{\Omega_N^{-1}}\\ &\le N^2C_\phi^2\cdot(2\epsilon)^2/\lambda = 4N^2C_\phi^2\epsilon^2/\lambda.\end{aligned} \tag{237}$$

To bound the first term on the right hand side of eq. (236), we apply Lemma 5 to $G'(x_i) - \mathbb{E}\left[G'(x_i)|\mathcal{F}_{i-1}\right]$. We fix $G' \in \mathcal{G}$ and let $\varepsilon_i = G'(x_i) - \mathbb{E}\left[G'(x_i)|\mathcal{F}_{i-1}\right]$ and $E_N = [\varepsilon_1,\cdots,\varepsilon_{N-1}]^\top \in \mathbb{R}^{N-1}$. Using this notation, we have

$$\begin{aligned}&\left\|\sum_{i=1}^N\phi(x_i)\left(G'(x_i) - \mathbb{E}\left[G'(x_i)|\mathcal{F}_{i-1}\right]\right)\right\|^2_{\Omega_N^{-1}}\\ &= \left\|\sum_{i=1}^N\phi(x_i)\varepsilon_i\right\|^2_{\Omega_N^{-1}}\\ &= \left\|\Phi^\top E_N\right\|^2_{\Omega_N^{-1}}\\ &= E_N^\top\Phi\Omega_N^{-1}\Phi^\top E_N\\ &\overset{(i)}{=} E_N^\top\Phi\Phi^\top(K_N + \lambda I_N)^{-1}E_N =\\ &\quad E_N^\top K_N(K_N + \lambda I_N)^{-1}E_N\end{aligned}$$

$$\overset{(ii)}{\leq} E_N^\top (K_N + (\lambda-1)I_N)(K_N+\lambda I_N)^{-1} E_N$$
$$= E_N^\top (K_N + (\lambda-1)I_N)$$
$$[I_N + (K_N + (\lambda-1)I_N)]^{-1} E_N$$
$$= E_N^\top [(K_N + (\lambda-1)I_N)^{-1} + I_N] E_N, \tag{238}$$

where (i) follows from eq. (222) in Lemma 6 and (ii) follows from the fact that $\lambda > 1$ and $K_N + \lambda I_N$ is positive definite. Note that each entry of E_N is bounded by C_g in absolute value. Applying Lemma 7 to eq. (238) and taking a union bound over $\mathcal{G}_\epsilon$, for any $0 < \delta < 1$, we have the following holds with probability at least $1-\delta$

$$\sup_{G' \in \mathcal{G}_\epsilon} \left\| \sum_{i=1}^N \phi(x_i) \left(G'(x_i) - \mathbb{E}\left[G'(x_i)|\mathcal{F}_{i-1}\right]\right) \right\|^2_{\Omega_N^{-1}}$$
$$\leq C_g^2 \log\det[(1+\eta)I + K_N]$$
$$+ 2C_g^2 \log(\mathcal{N}_\epsilon/\delta). \tag{239}$$

Moreover, note that $(1+\eta)I + K_N = [I + (1+\eta)^{-1}K_N][(1+\eta)I]$, which implies

$$\log\det[(1+\eta)I + K_N]$$
$$= \log\det[I + (1+\eta)^{-1}K_N] + N\log(1+\eta)$$
$$\leq \log\det[I + (1+\eta)^{-1}K_N] + N\eta. \tag{240}$$

Combining eq. (236), eq. (237), eq. (238), eq. (239) and eq. (240) and letting $\eta = \lambda - 1$, we have the following holds with probability $1-\delta$

$$\left\| \sum_{i=1}^N \phi(x_i) \left(G(x_i) - \mathbb{E}\left[G(x_i)|\mathcal{F}_{i-1}\right]\right) \right\|^2_{\Omega_N^{-1}}$$
$$\leq 2C_g^2 \log\det(I + K_N/\lambda) + 2C_g^2 N(\lambda-1)$$
$$+ 4C_g^2 \log(\mathcal{N}_\epsilon/\delta) + 8N^2 C_\phi^2 \epsilon^2/\lambda,$$

which completes the proof of eq. (234). To prove eq. (235) we do not need to go through the "ϵ-cover" argument since $G(\cdot)$ is independent from $\{x_i\}_{i\in N}$. We can directly apply Lemma 7 and then follow steps similar to those in eq. (240) to obtain eq. (235). □

For any integer N and $\lambda > 0$, we define the maximal information gain associated with the RKHS $\mathcal{H}$ as

$$\Gamma_K(N,\lambda) = \sup_{\mathcal{D}\subset\mathcal{X}} \{1/2 \cdot \log\det(I_d + \lambda^{-1}\cdot K_N)\},$$

where the supremum is taken over all discrete subset $\mathcal{D}$ of $\mathcal{X}$ with the cardinality no more than N.

Lemma 9 (Finite Spectrum/Effective Dimension Property): Let $\{\sigma_j\}_{j\geq 1}$ be the eigenvalues of T_K defined in eq. (217) in the descending order. Let $\lambda \in [c_1, c_2]$ with c_1 and c_2 being absolute constants. If $\sigma_j = 0$ for all $j \geq D+1$, where D is a positive integer. Then, we have $\Gamma_K(N,\lambda) = C_K \cdot D \cdot \log N$, where C_K is an absolute constant that depends on C_1, C_2, c_1, c_2 and C_ϕ.

Proof: See the proof of Lemma D.5 in [71] for a detailed proof. □

Appendix K
Other Useful Lemmas

Lemma 10 (Matrix Bernstein Inequality [81]): Suppose that $\{A_i\}_{i=1}^N$ are independent and centered random matrices in $\mathbb{R}^{d_1\times d_2}$, that is, $\mathbb{E}[A_i] = 0$ for all $i \in [N]$. Also, suppose $\|A_i\|_2 \leq C_A$ for all $i \in [n]$. Let $Z = \sum_{i=1}^N A_i$ and

$$v(Z) = \max\left\{\left\|\mathbb{E}\left[ZZ^\top\right]\right\|_2, \left\|\mathbb{E}\left[Z^\top Z\right]\right\|_2\right\}.$$

For all $\xi \geq 0$, we have

$$\mathrm{P}(\|Z\|_2 \geq \xi)$$
$$\leq (d_1+d_2)\cdot \exp\left(-\frac{\xi^2/2}{v(Z) + C_A/3\cdot\xi}\right).$$

Proof: See Theorem 1.6.2 in [81] for a detailed proof. □

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[2] V. Mnih, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[3] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2016.

[4] D. Silver et al., "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[5] A. W. Senior et al., "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 7792.

[6] H. R. Maei, "Gradient temporal-difference learning algorithms," Ph.D. thesis, Dept. Comput. Sci., Univ. Alberta, Edmonton, AB, Canada, 2011.

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

[8] N. Chatterji, A. Pacchiano, P. Bartlett, and M. Jordan, "On the theory of reinforcement learning with once-per-episode feedback," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 3401–3412.

[9] Y. Gong, M. Abdel-Aty, Q. Cai, and M. S. Rahman, "Decentralized network level adaptive signal control by multi-agent deep reinforcement learning," *Transp. Res. Interdiscipl. Perspect.*, vol. 1, Jun. 2019, Art. no. 100020.

[10] M. Olivecrona, T. Blaschke, O. Engkvist, and H. Chen, "Molecular de-novo design through deep reinforcement learning," *J. Cheminformatics*, vol. 9, no. 1, pp. 1–14, Dec. 2017.

[11] K. Lin, R. Zhao, Z. Xu, and J. Zhou, "Efficient large-scale fleet management via multi-agent deep reinforcement learning," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1774–1783.

[12] D. Hein et al., "A benchmark environment motivated by industrial control problems," in *Proc. IEEE Symp. Ser. Computat. Intell. (SSCI)*, Nov. 2017, pp. 1–8.

[13] H. Rahmandad, N. Repenning, and J. Sterman, "Effects of feedback delay on learning," *Syst. Dyn. Rev.*, vol. 25, no. 4, pp. 309–338, Oct. 2009.

[14] J. A. Arjona-Medina, M. Gillhofer, M. Widrich, T. Unterthiner, J. Brandstetter, and S. Hochreiter, "Rudder: Return decomposition for delayed rewards," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 32, 2019, pp. 1–12.

[15] A. Pacchiano, A. Saha, and J. Lee, "Dueling RL: Reinforcement learning with trajectory preferences," 2021, *arXiv:2111.04850*.

[16] Y. Liu, Y. Luo, Y. Zhong, X. Chen, Q. Liu, and J. Peng, "Sequence modeling of temporal credit assignment for episodic reinforcement learning," 2019, *arXiv:1905.13420*.

[17] T. Gangwani, Y. Zhou, and J. Peng, "Learning guidance rewards with trajectory-space smoothing," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 33, 2020, pp. 822–832.

[18] Z. Ren, R. Guo, Y. Zhou, and J. Peng, "Learning long-term reward redistribution via randomized return decomposition," 2021, *arXiv:2111.13485*.

[19] Y. Efroni, N. Merlis, and S. Mannor, "Reinforcement learning with trajectory feedback," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 7288–7295.

[20] O. Gottesman et al., "Guidelines for reinforcement learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 16–18, Jan. 2019.

[21] R. Wang, D. P. Foster, and S. M. Kakade, "What are the statistical limits of offline RL with linear function approximation?" 2020, *arXiv:2010.11895*.

[22] B. Han, Z. Ren, Z. Wu, Y. Zhou, and J. Peng, "Off-policy reinforcement learning with delayed rewards," 2021, *arXiv:2106.11854*.

[23] Z. Zheng, J. Oh, and S. Singh, "On learning intrinsic rewards for policy gradient methods," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 31, 2018, pp. 1–11.

[24] M. Klissarov and D. Precup, "Reward propagation using graph convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 33, 2020, pp. 12895–12908.

[25] J. Oh, Y. Guo, S. Singh, and H. Lee, "Self-imitation learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3878–3887.

[26] Y. Jin, Z. Yang, and Z. Wang, "Is pessimism provably efficient for offline RL?" in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 5084–5096.

[27] M. Yin, Y. Bai, and Y.-X. Wang, "Near-optimal offline reinforcement learning via double variance reduction," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 7677–7688.

[28] M. Yin, Y. Bai, and Y.-X. Wang, "Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning," 2020, *arXiv:2007.03760*.

[29] M. Yin, Y. Duan, M. Wang, and Y.-X. Wang, "Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism," 2022, *arXiv:2203.05804*.

[30] M. Yin and Y.-X. Wang, "Towards instance-optimal offline reinforcement learning with pessimism," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 4065–4078.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 1–11.

[33] A. Cohen, H. Kaplan, T. Koren, and Y. Mansour, "Online Markov decision processes with aggregate bandit feedback," in *Proc. Conf. Learn. Theory (COLT)*, 2021, pp. 1301–1329.

[34] Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill, "Provably good batch reinforcement learning without great exploration," 2020, *arXiv:2007.08202*.

[35] R. Dadashi, S. Rezaeifar, N. Vieillard, L. Hussenot, O. Pietquin, and M. Geist, "Offline reinforcement learning with pseudometric learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2307–2318.

[36] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 2052–2062.

[37] S. Fujimoto, E. Conti, M. Ghavamzadeh, and J. Pineau, "Benchmarking batch deep reinforcement learning algorithms," 2019, *arXiv:1910.01708*.

[38] Z. Wang et al., "Critic regularized regression," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7768–7778.

[39] S. Fujimoto and S. S. Gu, "A minimalist approach to offline reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 20132–20145.

[40] J. Buckman, C. Gelada, and M. G. Bellemare, "The importance of pessimism in fixed-dataset policy optimization," 2020, *arXiv:2009.06799*.

[41] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1179–1191.

[42] L. Shi, G. Li, Y. Wei, Y. Chen, and Y. Chi, "Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity," 2022, *arXiv:2202.13890*.

[43] Y. Yan, G. Li, Y. Chen, and J. Fan, "The efficacy of pessimism in asynchronous Q-learning," 2022, *arXiv:2203.07368*.

[44] G. Li, L. Shi, Y. Chen, Y. Chi, and Y. Wei, "Settling the sample complexity of model-based offline reinforcement learning," 2022, *arXiv:2204.05275*.

[45] M. Yin and Y.-X. Wang, "Optimal uniform ope and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021.

[46] T. Ren, J. Li, B. Dai, S. S. Du, and S. Sanghavi, "Nearly horizon-free offline reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 15621–15634.

[47] T. Xie, N. Jiang, H. Wang, C. Xiong, and Y. Bai, "Policy finetuning: Bridging sample-efficient offline and online reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 27395–27407.

[48] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell, "Bridging offline reinforcement learning and imitation learning: A tale of pessimism," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 11702–11716.

[49] T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal, "Bellman-consistent pessimism for offline reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 6683–6694.

[50] A. Zanette, M. J. Wainwright, and E. Brunskill, "Provable benefits of actor-critic methods for offline reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 13626–13640.

[51] A. Zanette, "Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 12287–12297.

[52] D. J. Foster, A. Krishnamurthy, D. Simchi-Levi, and Y. Xu, "Offline reinforcement learning: Fundamental barriers for value function approximation," 2021, *arXiv:2111.10919*.

[53] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Proc. Conf. Learn. Theory*, 2020, pp. 2137–2143.

[54] L. F. Yang and M. Wang, "Sample-optimal parametric Q-learning using linearly additive features," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6995–7004.

[55] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement learning: Theory and algorithms," Dept. CS, UW Seattle, Seattle, WA, USA, Tech. Rep., 2019, vol. 32, p. 96.

[56] B. Scherrer, "Approximate policy iteration schemes: A comparison," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1314–1322.

[57] J. Chen and N. Jiang, "Information-theoretic considerations in batch reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1042–1051.

[58] N. Jiang, "On value functions and the agent-environment boundary," 2019, *arXiv:1905.13341*.

[59] L. Wang, Q. Cai, Z. Yang, and Z. Wang, "Neural policy gradient methods: Global optimality and rates of convergence," 2019, *arXiv:1909.01150*.

[60] P. Liao, Z. Qi, R. Wan, P. Klasnja, and S. Murphy, "Batch policy learning in average reward Markov decision processes," 2020, *arXiv:2007.11771*.

[61] B. Liu, Q. Cai, Z. Yang, and Z. Wang, "Neural trust region/proximal policy optimization attains globally optimal policy," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 1–12.

[62] Y. Duan, Z. Jia, and M. Wang, "Minimax-optimal off-policy evaluation with linear function approximation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2701–2709.

[63] T. Xie and N. Jiang, "Batch value-function approximation with only realizability," 2020, *arXiv:2008.04990*.

[64] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020, *arXiv:2005.01643*.

[65] A. M. Farahmand, R. Munos, and C. Szepesvári, "Error propagation for approximate policy and value iteration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1–9.

[66] R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims, "MOReL: Model-based offline reinforcement learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21810–21823.

[67] T. Yu et al., "MOPO: Model-based offline policy optimization," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 14129–14142.

[68] R. Gao, T. Cai, H. Li, C.-J. Hsieh, L. Wang, and J. D. Lee, "Convergence of adversarial training in overparametrized neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 32, 2019, pp. 1–12.

[69] Y. Bai and J. D. Lee, "Beyond linearization: On quadratic and higher-order approximation of wide neural networks," 2019, *arXiv:1910.01619*.

[70] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 31, 2018, pp. 1–10.

[71] Z. Yang, C. Jin, Z. Wang, M. Wang, and M. I. Jordan, "On function approximation in reinforcement learning: Optimism in the face of large state spaces," 2020, *arXiv:2011.04622*.

[72] Q. Cai, Z. Yang, J. D. Lee, and Z. Wang, "Neural temporal-difference and Q-learning provably converge to global optima," 2019, *arXiv:1905.10027*.
[73] T. Xu, Y. Liang, and G. Lan, "CRPO: A new approach for safe reinforcement learning with convergence guarantee," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 11480–11491.
[74] S. Qiu, J. Ye, Z. Wang, and Z. Yang, "On reward-free RL with kernel and neural function approximations: Single-agent MDP and Markov game," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8737–8747.
[75] D. Zhou, L. Li, and Q. Gu, "Neural contextual bandits with UCB-based exploration," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11492–11502.
[76] M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini, "Finite-time analysis of kernelised contextual bandits," 2013, *arXiv:1309.6869*.
[77] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.
[78] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2018.
[79] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 242–252.
[80] S. R. Chowdhury and A. Gopalan, "On kernelized multi-armed bandits," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 844–853.
[81] J. A. Tropp, "An introduction to matrix concentration inequalities," 2015, *arXiv:1501.01571*.

Tengyu Xu received the B.E. degree from Xi'an Jiaotong University, Xi'an, China, in 2015, and the Ph.D. degree in electrical and computer engineering from The Ohio State University, Columbus, OH, USA, in 2022. Since 2022, he has been a member of the GenAI Meta AI Team, Meta Platforms, Menlo Park. His research interests include reinforcement learning, large language model post-tuning, optimization, and statistical learning theory.

Yue Wang received the B.S. degree from Nanjing University, Nanjing, China, in 2019, and the Ph.D. degree in electrical engineering from SUNY at Buffalo, Buffalo, NY, USA, in 2023. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Central Florida. His research interests include reinforcement learning, machine learning, and generative models. He received the 2023 AAAI Distinguished Paper Award.

Shaofeng Zou (Member, IEEE) received the B.E. degree (Hons.) from Shanghai Jiao Tong University, Shanghai, China, in 2011, and the Ph.D. degree in electrical and computer engineering from Syracuse University, Syracuse, NY, USA, in 2016. From 2016 to 2018, he was a Post-Doctoral Research Associate with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign. He is currently an Assistant Professor with the Department of Electrical Engineering, University at Buffalo, The State University of New York. His research interests include reinforcement learning, machine learning, statistical signal processing, and information theory. He received the National Science Foundation CAREER Award in 2024, the National Science Foundation CRII Award in 2019, and the 2023 AAAI Distinguished Paper Award. He has been serving as an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING since 2023.

Yingbin Liang (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign in 2005. She is currently a Professor with the Department of Electrical and Computer Engineering, The Ohio State University (OSU), and a Core Faculty Member of Ohio State Translational Data Analytics Institute (TDAI). She is also the Deputy Director of the AI-EDGE Institute, OSU. She was a Faculty Member of the University of Hawai'i and Syracuse University before she joined OSU. Her research interests include machine learning, optimization, information theory, and statistical signal processing. She received the National Science Foundation CAREER Award and the State of Hawaii Governor Innovation Award in 2009. She received the EURASIP Best Paper Award in 2014.