Trajectory Data Suffices for Statistically Efficient Policy Evaluation in Fixed-Horizon Offline RL with Linear q^π-Realizability and Concentrability

Volodymyr Tkachuk VTKACHUK@UALBERTA.CA
University of Alberta

Csaba Szepesvári SZEPESVA@UALBERTA.CA
University of Alberta

Xiaoqi Tan XIAOQI.TAN@UALBERTA.CA
University of Alberta

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study fixed-horizon offline reinforcement learning (RL) with function approximation for both policy evaluation and policy optimization. Prior work established that statistically efficient learning is impossible for either of these problems when the only assumptions are that the data has good coverage (concentrability) and the state-action value function of every policy is linearly realizable (q^π-realizability) [Foster et al., 2022]. Recently, Tkachuk et al. [2024] gave a statistically efficient learner for policy optimization, if in addition the data is assumed to be given as trajectories. In this work we present a statistically efficient learner for policy evaluation under the same assumptions, with the additional requirement that the behavior policy is known. Further, we show that the sample complexity of the learner used by Tkachuk et al. [2024] for policy optimization can be improved by a tighter analysis.

Keywords: Reinforcement learning, Offline RL, Sample complexity, Linear function approximation, Policy evaluation, Policy optimization

1. Introduction and Overview of Results

In offline RL a learner is given access to a dataset and is tasked with either evaluating a policy or finding the optimal policy [Jiang and Xie, 2024]. If the dataset does not cover the state-action space well, then certain parts of the Markov decision process (MDP) will be unseen, and no learner will be able to succeed [Chen and Jiang, 2019]. Thus, some assumption needs to be made on the coverage of the dataset to get positive results. One such assumption, which will be central to this work, is when the dataset covers the states and actions reachable by **all** policies well. This assumption is known as *concentrability*, with parameter C_0 indicating the degree of coverage (see Assumption 3).

When the state space is large it is undesirable for the size of the dataset to scale with the number of states in the MDP. As such, function approximation is often used to represent particular aspects of the problem. In this work we focus on the case where a function class is used to represent state-action value functions (q-functions). Foster et al. [2022] showed that if the dataset satisfies concentrability, and the q-function of **all** policies is realizable by the function class, then achieving ϵ-optimal policy evaluation or optimization with high probability requires a dataset size that scales with the number of states, which is considered statistically inefficient.

	Trajectory Data		
Objective	Obj π Realizable	All π Lin-Realizable	All π Realizable
Policy Evaluation	x [Jia et al., 2024]	✓ **[This Work]**	x [Foster et al., 2022]
Policy Optimization	x [Liu et al., 2026]	✓ [Tkachuk et al., 2024]	x [Foster et al., 2022]

Table 1: Concentrability is assumed in all cases. *Obj π Realizable* means **only** the evaluation/optimal policy is realizable by the function class for policy evaluation/optimization. *All π (Lin-)Realizable* means the q-function of **all** memoryless policies is realizable by the (linear) function class. The ✓ means a $\mathrm{poly}(d, H, C_0, 1/\epsilon)$ sample complexity upper bound, and x means a lower bound in terms of the state space size or exponential in the horizon H.

Naturally, the question of how much these assumptions need to be strengthened to allow for statistically efficient learning remained. Tkachuk et al. [2024] showed that if the dataset has the extra structure of being full rollouts from some *behavior policy* (i.e., *trajectories*), and the q-function of every policy is realizable by a linear function class (with known features), then the sample complexity of policy optimization can be improved to scale polynomially with the feature dimension d and other problem parameters. Importantly, this is considered statistically efficient, since the feature dimension d is often chosen much smaller than the number of states. Their result does not imply a similar bound for policy evaluation, since their learner follows a value-iteration approach, which does not evaluate intermediate policies (see Section 4.4 for further discussion). Thus, whether a similar result is possible for policy evaluation remained open.

In this work we give a positive answer, by providing a statistically efficient learner for policy evaluation under the same assumptions as Tkachuk et al. [2024], except we additionally assume the behavior policy is known. Interestingly, Jia et al. [2024] showed that if the q-function of **only** the evaluation policy is realizable by the function class, then the dataset size must scale exponentially with the horizon H of the MDP. Our result complements their work by showing that if the function class is assumed to linearly realize the q-function of **all** memoryless policies, then the sample complexity can be improved to scale polynomially with the horizon H, feature dimension d, and other problem parameters. Recently, Liu et al. [2026] showed a negative result for policy optimization when only the q-function of the optimal policy is realizable, which complements the result of Jia et al. [2024] for policy evaluation. In Table 1 we summarize our above discussion. For a detailed account of the offline RL literature we refer the reader to the excellent survey by Jiang and Xie [2024].

Another contribution comes as a consequence of our analysis building on the work of Tkachuk et al. [2024]. In particular, we noticed that a bound in a key lemma used by Tkachuk et al. [2024] can be improved. Therefore, we also provide a sample complexity bound for the policy optimization objective, which improves on that of Tkachuk et al. [2024] by a factor of $C_0 d$.

The remainder of the paper is structured as follows. Section 2 defines the offline RL setting. In Section 3 we present our results. Then, in Section 4 we define the learners for which the results hold.

2. Problem Setting

Since we study the same setting as Tkachuk et al. [2024], we use similar notation and definitions.

Notation: Throughout we fix the integer $d \geq 1$. Let $\mathbf{0} \in \mathbb{R}^d$ be the d-dimensional, all zero vector. For $L > 0$, let $\mathcal{B}(L) = \{x \in \mathbb{R}^d : \|x\|_2 \leq L\}$ denote the d-dimensional Euclidean ball of radius L centered at the origin, where $\|\cdot\|_2$ denotes the Euclidean norm. The inner product $\langle x, y\rangle$ for $x, y \in \mathbb{R}^d$ is defined as the dot product $x^\top y$. The closure of a set $\mathcal{X}$ is denoted by $\mathrm{cl}(\mathcal{X})$. Let $\mathbb{I}\{B\}$ be the indicator function of a boolean-valued variable B, taking the value 1 if B is true and 0 if false. Let $\mathcal{M}_1(X)$ denote the set of probability distributions over the set X. Let $\mathbb{E}_{B\sim\mathcal{P}}[f(B)]$ denote the expectation of a random variable $f(B)$ under distribution $\mathcal{P}$. For integers i, j, let $[i] = \{1, \ldots, i\}$ and $[i : j] = \{i, \ldots, j\}$. For any two functions f, g, comparisons are always pointwise (e.g. $(f \leq g)(x) \coloneqq f(x) \leq g(x)$ for all x).

The environment is modeled by a fixed-horizon Markov decision process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, H)$. The state space $\mathcal{S}$ is finite but arbitrarily large, and organized by stages: $\mathcal{S} = \bigcup_{h\in[H+1]} \mathcal{S}_h$, starting from a designated initial state s_1 ($\mathcal{S}_1 = \{s_1\}$)[1] , and culminating in a designated terminal state $s_\top$ ($\mathcal{S}_{H+1} = \{s_\top\}$)[2]. Without loss of generality, we assume $\mathcal{S}_h$ and $\mathcal{S}_{h'}$ for $h \neq h'$ are disjoint sets. The action space $\mathcal{A}$ is finite. The transition kernel is $P : (\bigcup_{h\in[H]} \mathcal{S}_h) \times \mathcal{A} \to \mathcal{M}_1(\mathcal{S})$, with the property that transitions occur between successive stages. Specifically, for any $h \in [H]$, state $s_h \in \mathcal{S}_h$, and action $a \in \mathcal{A}$, $P(s_h, a) \in \mathcal{M}_1(\mathcal{S}_{h+1})$. The reward kernel is $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathcal{M}_1([0, 1])$. To ensure that the terminal state $s_\top$ has no influence on the learner we force the reward kernel to deterministically give zero reward for all actions $a \in \mathcal{A}$ in $s_\top$.

We will define an agent's interaction with the MDP through a *memoryless policy* $\pi : \mathcal{S} \to \mathcal{M}_1(\mathcal{A})$, which assigns a probability distribution over actions based on a state. The set of all memoryless policies is $\Pi = \{\pi : \mathcal{S} \to \mathcal{M}_1(\mathcal{A})\}$. For $\pi \in \Pi$, we write $\pi(a|s)$ to denote the probability $\pi(s)$ assigns to action a. For deterministic policies only (i.e., all the probability is concentrated on a single action) we sometimes abuse notation by writing $\pi(s)$ to denote $\arg\max_{a\in\mathcal{A}} \pi(a|s)$. An agent interacts with the MDP sequentially from a state $s \in \mathcal{S}$, by sampling an action according to its policy π, receiving a reward specified by $\mathcal{R}$, transitioning to a subsequent state according to P, and repeating this process until receiving a reward at the terminal state $s_\top$. This interaction induces a probability distribution over trajectories, denoted as $\mathbb{P}_{\pi,s}$. Formally, for any $h \in [H + 1]$ and $s \in \mathcal{S}_h$, we write $\mathrm{Traj} \sim \mathbb{P}_{\pi,s}$ to denote a trajectory $\mathrm{Traj} = (s, A_h, R_h, \ldots, S_{H+1}, A_{H+1}, R_{H+1})$ distributed according to $\mathbb{P}_{\pi,s}$, where $S_h = s$, $A_i \sim \pi(S_i)$ for $i \in [h : H + 1]$, $S_{i+1} \sim P(S_i, A_i)$ for $i \in [h : H]$, and $R_i \sim \mathcal{R}(S_i, A_i)$ for $i \in [h : H + 1]$. For any $a \in \mathcal{A}$, we also define $\mathbb{P}_{\pi,s,a}$ as a distribution over trajectories when first action a is taken in state s, and then policy π is followed. For $h \in [H + 1]$, we write $\mathbb{P}^h_{\pi,s}$ (and $\mathbb{P}^h_{\pi,s,a}$) for the marginal distribution of (S_h, A_h) (i.e., the state-action pair of stage h) based on the joint distribution of $\mathbb{P}_{\pi,s}$ (and $\mathbb{P}_{\pi,s,a}$).

For $1 \leq t \leq t' \leq H + 1$, we use the notation $R_{t:t'} = \sum_{u=t}^{t'} R_u$. At a stage $h \in [H + 1]$, the state-value and action-value functions v^π and q^π, for a policy $\pi \in \Pi$, and $s_h \in \mathcal{S}_h$, $a \in \mathcal{A}$, are:

$$v^\pi(s_h) \coloneqq \mathbb{E}_{\mathrm{Traj}\sim\mathbb{P}_{\pi,s_h}} R_{h:H+1} \quad \text{and} \quad q^\pi(s_h, a) \coloneqq \mathbb{E}_{\mathrm{Traj}\sim\mathbb{P}_{\pi,s_h,a}} R_{h:H+1}\,.$$

For any function $f \in \mathbb{R}^{\mathcal{S}}$, we write $f_h \in \mathbb{R}^{\mathcal{S}_h}$ for its restriction to stage h. Similarly, for any function $g \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, we write $g_h \in \mathbb{R}^{\mathcal{S}_h\times\mathcal{A}}$ for its restriction to stage h. Let $\pi^\star \in \Pi$ be the optimal policy, satisfying $q^{\pi^\star}(s, a) \coloneqq \sup_{\pi\in\Pi} q^\pi(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. The optimal policy is known to exist and is deterministic [Puterman, 2014].

1. A deterministic start state s_1 is added for simplicity of presentation. It is easy to show that adding an additional stage to the MDP allows for the transition dynamics to encode an arbitrary start state distribution.
2. A terminal state $s_\top$ is added purely as a technical convenience for the analysis. We will focus on the interaction of learners for stages $h \in [H]$ (not $[H + 1]$), since the terminal state will have no effect on the learner.

2.1. Assumptions and Objective

For a fixed feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathcal{B}(L_\phi)$, $L_\phi > 0$ (known to the learner), the function class at stage $h \in [H+1]$ containing linear functions is defined as follows:

$$\mathcal{F}_h \coloneqq \{f : \mathcal{S}_h \times \mathcal{A} \to [0, H] \mid f(\cdot, \cdot) = \langle \phi(\cdot, \cdot),\, \theta \rangle, \theta \in \mathcal{B}(L_\theta)\} \quad \text{for some } L_\theta > 0\,.$$

The first assumption is that $\mathcal{F}_h$ realizes the q-function of all memoryless policies exactly[3].

Assumption 1 (q^π-Realizability[4]) *$q_h^\pi \in \mathcal{F}_h$, for all $\pi \in \Pi$, and stages $h \in [H+1]$. For any $h \in [H+1]$, let $\theta_h^\star : \Pi \to \mathcal{B}(L_\theta)$ be a mapping that satisfies $q_h^\pi(\cdot, \cdot) = \langle \phi(\cdot, \cdot),\, \theta_h^\star(\pi) \rangle$.*

Now, we introduce our assumptions on the data.

Assumption 2 (Trajectory Data) *The learner is given $n \geq 1$ trajectories[5] $(\text{Traj}^j)_{j=1}^n$, where each $\text{Traj}^j = (S_h^j, A_h^j, R_h^j)_{h \in [H+1]}$ is an independent sample from $\mathbb{P}_{\pi^b, s_1}$, for **behavior policy** $\pi^b \in \Pi$.*

A trajectory is a sample from the joint distribution $\mathbb{P}_{\pi^b, s_1}$. This is in contrast to just having samples from the marginal distributions $(\mathbb{P}^h_{\pi^b, s_1})_{h \in [H]}$[6], which hide temporal dependencies between states and actions across stages that will be crucial for our learner (see Section 4.2). A sequence $\nu = (\nu_h)_{h \in [H]}$, where $\nu_h \in \mathcal{M}_1(\mathcal{S}_h \times \mathcal{A})$, is admissible if there exists a policy $\pi \in \Pi$ such that

$$\nu_h(s, a) = \mathbb{P}^h_{\pi, s_1}(s, a) \quad \text{for all } (s, a) \in \mathcal{S}_h \times \mathcal{A},\ h \in [H]\,.$$

We use $\mu \coloneqq (\mu_h)_{h \in [H]}$, where $\mu_h \coloneqq \mathbb{P}^h_{\pi^b, s_1}$, for the behavior policy's admissible distributions. The following assumption requires the behavior policy's coverage of the state-action space to be nearly as good as any memoryless policy.

Assumption 3 (Concentrability) *For all admissible distributions $\nu = (\nu_h)_{h \in [H]}$,*

$$\max_{h \in [H]} \max_{(s,a) \in \mathcal{S}_h \times \mathcal{A}} (\nu_h(s, a) / \mu_h(s, a)) \leq C_0, \quad \textit{where } C_0 \geq 1\,.$$

The policy optimization and evaluation objectives are as follows. Let $\epsilon > 0$. The policy optimization objective is to minimize the number of trajectories n, while outputting a policy $\hat{\pi}$, such that $v^{\pi^\star}(s_1) - v^{\hat{\pi}}(s_1) \leq \epsilon$. Let π^e be an arbitrary *evaluation policy*. The policy evaluation objective is to minimize the number of trajectories n, while outputting a value estimate $\hat{v} \in \mathbb{R}$, such that $|v^{\pi^e}(s_1) - \hat{v}| \leq \epsilon$.

3. We assume exact realizability for simplicity of presentation. Our results can be extended to the approximate case with minor modifications to the proofs, as was done by Tkachuk et al. [2024].
4. Since $\mathcal{F}_h$ is a linear function class, this should be called **linear** q^π-realizability; however, since we always work with linear function classes in this paper, we omit the word "linear" for brevity. The same is done for future assumptions that use $\mathcal{F}_h$ (e.g., Assumptions 4 and 5).
5. Our learner does not require explicit knowledge of the states within each trajectory; the features alone are sufficient.
6. Since samples from the marginals are just state-action pairs, we of course mean that a reward and next state are also provided for each state-action pair. Even weaker is having samples from a general distribution over $\mathcal{S} \times \mathcal{A}$, for which the negative result of Foster et al. [2022] applies.

3. Results

Our main result is the first polynomial sample complexity for policy evaluation (proof: Appendix A).

Theorem 1 (Policy Evaluation) *Let Assumptions 1 to 3 hold and the behavior policy π^b be **known**. For any evaluation policy π^e, $\delta \in (0,1)$ and $\epsilon > 0$, if the number of trajectories $n = \tilde{\Theta}(C_0^5 H^7 d^3/\epsilon^2)$, then with probability at least $1-\delta$, the value estimate $\hat{v}$ output by Algorithm 2 satisfies*

$$\left|v^{\pi^e}(s_1) - \hat{v}\right| \leq \epsilon\,. \tag{1}$$

The notations $\tilde{\Omega}, \tilde{\mathcal{O}}$ and $\tilde{\Theta}$ hide polylogarithmic factors of $(1/\epsilon, 1/\delta, H, d, C_0, L_\phi, L_\theta)$. Due to our improved analysis (in Appendix C.4), we obtain a sample complexity for policy optimization that improves the bound on n by Tkachuk et al. [2024] by a factor of $C_0 d$ (proof: Appendix B).

Theorem 2 (Policy Optimization) *Let Assumptions 1 to 3 hold and the behavior policy π^b be **unknown**. For any $\delta \in (0,1)$ and $\epsilon > 0$, if the number of trajectories $n = \tilde{\Theta}(C_0^3 H^7 d^3/\epsilon^2)$, then with probability at least $1-\delta$, the policy $\hat{\pi}$ output by Algorithm 1 satisfies*

$$v^{\pi^\star}(s_1) - v^{\hat{\pi}}(s_1) \leq \epsilon\,. \tag{2}$$

4. Learners and Background

In Section 4.4 we present our main contribution, a sample efficient learner for policy evaluation (Algorithm 2). We use the first three subsections to present **known** results, which we believe are crucial for understanding our contribution. In Section 4.1 we present the fitted Q-iteration/evaluation (FQE/FQI) algorithms for policy evaluation and optimization respectively. Since FQE/FQI works with the Bellman completeness assumptions (Assumptions 4 and 5), we show how to get Bellman completeness from q^π-realizability in Section 4.2. In Section 4.3 we show how Tkachuk et al. [2024] used these ideas along with trajectory data to develop their learner for policy optimization. We finally present our policy evaluation learner in Section 4.4.

4.1. Fitted Q-Iteration/Evaluation (FQI/FQE) with Bellman Completeness

The *Bellman policy operator* $T^\pi : \cup_{h\in[2:H+1]}\mathbb{R}^{\mathcal{S}_h\times\mathcal{A}} \to \cup_{h\in[H]}\mathbb{R}^{\mathcal{S}_h\times\mathcal{A}}$ for a policy π, is defined as follows: for all $h \in [H]$, $q_{h+1} \in \mathbb{R}^{\mathcal{S}_{h+1}\times\mathcal{A}}$,

$$T^\pi q_{h+1}(s,a) \coloneqq \mathbb{E}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}}[R_h + q_{h+1}(S_{h+1},\pi)], \quad \text{for all } (s,a) \in \mathcal{S}_h \times \mathcal{A}\,, \tag{3}$$

where $q_{h+1}(s,\pi) \coloneqq \sum_{a\in\mathcal{A}} q_{h+1}(s,a)\pi(a|s)$ is notation that will be used throughout the paper. Notice that $q_h^\pi = T^\pi q_{h+1}^\pi$. Since we do not have access to T^π, and $T^\pi q_{h+1}^\pi$ is a conditional expectation, we can approximate it from data by solving a regression problem. Let $\mu' = (\mu'_1, \ldots, \mu'_H)$, be an arbitrary sequence of state-action distributions. Only for this subsection, we assume each offline data sample $j \in [n]$ is generated as follows for all $h \in [H]$: $(S_h^j, A_h^j) \sim \mu'_h$, $R_h^j \sim \mathcal{R}(S_h^j, A_h^j)$, $\bar{S}_{h+1}^j \sim P(\cdot|S_h^j, A_h^j)$. This is more general than trajectory data, since the state-action pairs (S_h^j, A_h^j) are not necessarily generated by following the behavior policy π^b, and the next state $\bar{S}_{h+1}^j$ need not be equal to S_{h+1}^j. The empirical Bellman policy operator $\hat{T}^\pi : \cup_{h\in[2:H+1]}\mathbb{R}^{\mathcal{S}_h\times\mathcal{A}} \to \cup_{h\in[H]}\mathbb{R}^{\mathcal{S}_h\times\mathcal{A}}$ for a policy π, is defined as follows: for all $h \in [H]$, $q_{h+1} \in \mathbb{R}^{\mathcal{S}_{h+1}\times\mathcal{A}}$,

$$\hat{T}^\pi q_{h+1} \coloneqq \arg\min_{q_h\in\mathcal{F}_h} \tfrac{1}{n}\textstyle\sum_{j=1}^n (q_h(S_h^j, A_h^j) - R_h^j - q_{h+1}(\bar{S}_{h+1}^j, \pi))^2.$$

For policy evaluation, FQE starts with $\hat{q}_{H+1} = \mathbf{0}$ and calculates $\hat{q}_h = \hat{T}^{\pi^e} \hat{q}_{h+1}$ for $h = H, \dots, 1$. It then outputs $\hat{v} = \hat{q}_1(s_1, \pi^e)$. A sufficient condition (on $\mathcal{F}_h$) for FQE to work is:

Assumption 4 (Bellman Completeness (for T^π)) $T^\pi q_{h+1} \in \mathcal{F}_h$, $\forall q_{h+1} \in \mathcal{F}_{h+1}$ *and* $\forall h \in [H]$.

Roughly speaking, Bellman completeness ensures that the approximation error does not multiplicatively accumulate across stages. If $\mu' = (\mu'_h)_{h\in[H]}$ satisfies concentrability (Assumption 3) and Bellman completeness (Assumption 4) holds[7], then the policy evaluation objective can be achieved with a polynomial number of samples that is independent of the state space size $|\mathcal{S}|$ [Le et al., 2019]. Notice how trajectory data (Assumption 2) was not necessary, if we have Bellman completeness. Unfortunately, we cannot directly use FQE with q^π-realizability (Assumption 1), since it does not imply Bellman completeness [Zanette et al., 2020]. In fact, without further assumptions on the data, Foster et al. [2022] showed that a policy evaluation bound (like in Eq. (1)), cannot be achieved unless the sample complexity scales with the size of the state space $|\mathcal{S}|$, which is undesirable. Naturally, this leads to imposing an extra assumption on the data, which we choose to be trajectory data (Assumption 2), whose benefits we discuss in the next section.

Before moving on, we address policy optimization. One could try to evaluate every policy using FQE and then select the one with the largest value from the start state; however, since the optimal policy is deterministic, there are $|\mathcal{A}|^{|\mathcal{S}|}$ policies to evaluate, and thus even by taking a union bound over this set of policies we would obtain a bound that scales as $\log(|\mathcal{A}|^{|\mathcal{S}|}) = |\mathcal{S}| \log(|\mathcal{A}|)$. Fortunately, this can be improved if our q-function class $\mathcal{F}$ is small or structured, by making use of the *Bellman optimality operator*. In particular, the Bellman optimality operator $T : \cup_{h\in[2:H+1]} \mathbb{R}^{\mathcal{S}_h \times \mathcal{A}} \to \cup_{h\in[H]} \mathbb{R}^{\mathcal{S}_h \times \mathcal{A}}$ and empirical Bellman optimality operator $\hat{T} : \cup_{h\in[2:H+1]} \mathbb{R}^{\mathcal{S}_h \times \mathcal{A}} \to \cup_{h\in[H]} \mathbb{R}^{\mathcal{S}_h \times \mathcal{A}}$ are defined as follows[8] : for all $h \in [H]$, $\pi \in \Pi$, $q_{h+1} \in \mathbb{R}^{\mathcal{S}_{h+1} \times \mathcal{A}}$,

$$Tq_{h+1}(s,a) \coloneqq \mathbb{E}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}}[R_h + \max_{a'} q_{h+1}(S_{h+1}, a')], \quad \text{for all } (s,a) \in \mathcal{S}_h \times \mathcal{A}\,. \tag{4}$$

$$\hat{T}q_{h+1} \coloneqq \arg\min_{q_h \in \mathcal{F}_h} \tfrac{1}{n} \textstyle\sum_{j=1}^n (q_h(S_h^j, A_h^j) - R_h^j - \max_{a'} q_{h+1}(\bar{S}_{h+1}^j, a'))^2 \tag{5}$$

For policy optimization, FQI starts with $\hat{q}_{H+1} = \mathbf{0}$ and calculates $\hat{q}_h = \hat{T}\hat{q}_{h+1}$ for $h = H, \dots, 1$. It then outputs $\hat{\pi}(\cdot) = \arg\max_a \hat{q}(\cdot, a)$. The Bellman completeness assumption for T is:

Assumption 5 (Bellman Completeness (of T)) $Tq_{h+1} \in \mathcal{F}_h$, $\forall q_{h+1} \in \mathcal{F}_{h+1}$ *and* $\forall h \in [H]$.

If $\mu' = (\mu'_h)_{h\in[H]}$ satisfies Assumption 3, and Bellman completeness (Assumption 5) holds, it can be shown that $\hat{q}$ is a good estimate of $q^{\pi^\star}$ in expectation. Finally, greedifying with respect to a good estimate of $q^{\pi^\star}$ (see Lemma 17) achieves the policy optimization objective (Eq. (2)) with a polynomial sample complexity [Munos and Szepesvári, 2008; Chen and Jiang, 2019]. Now, let us see what happens when we consider q^π-realizability, and why trajectories are helpful.

4.2. From q^π-realizability to Bellman Completeness

What we will see in this subsection is that although an MDP that is q^π-realizable is not necessarily Bellman complete, there exists a minor modification of the MDP such that Bellman completeness does hold in the modified MDP. Then, if we knew the modified MDP we could simply use FQE/FQI,

7. The result holds even if each function class $\mathcal{F}_h$ is not necessarily linear, but still satisfies Bellman completeness.
8. We use purple to highlight important differences between equations throughout the paper.

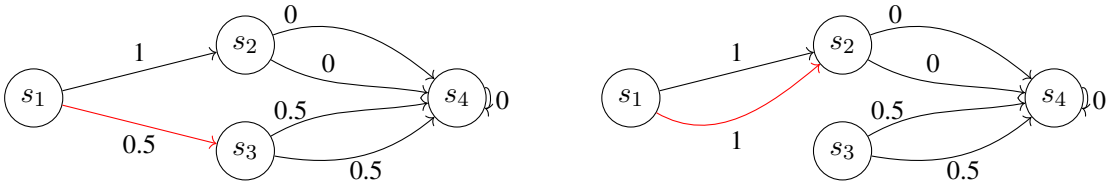

Figure 1: The features for both MDPs are $\phi(s_1, \cdot) = 1, \phi(s_3, \cdot) = 0.5, \phi(\cdot, \cdot) = 0$ otherwise. **Left:** q^π-realizable. **Right:** Linear MDP, due to skipping s_1 via the up action.

and since the modification is minor, it will imply good bounds in the original MDP. This observation was made by Weisz et al. [2023][9], and is necessary background for the following subsections.

To relate the q^π-realizability assumption to Bellman completeness, we will make use of another commonly used assumption called *linear MDP* [Jin et al., 2020]. Importantly, a linear MDP always satisfies Bellman completeness and q^π-realizability, but not vice versa [Zanette et al., 2020]. As such, if we can show that a modification to the original q^π-realizable MDP gives a linear MDP, then we will also have Bellman completeness in the modified MDP. We begin by building some intuition for why a q^π-realizable MDP can fail to be a linear MDP. Informally, a linear MDP is an MDP where the transition dynamics and expected rewards can be expressed as linear functions of the same state-action features ϕ. It is easy to construct an MDP for which there is a state that has the same q-function for all policies and actions; however, it does not have the same reward function or transition dynamics for all actions. Such an example is given on the left in Figure 1[10] , where in state s_1 the q-function is 1 for all policies and actions; however, the reward for taking the up action is 1, while the reward for taking the down action is 0.5. As such, to satisfy q^π-realizability, the features for the linear function class $\mathcal{F}$ can be defined as $\phi(s_1, \cdot) = 1$, since $\langle \phi(s_1, \cdot), 1 \rangle = 1$. But, such a feature map cannot capture the difference in rewards between the two actions, as it assigns the same value to both the up and down action. A similar issue arises for the transitions in state s_1.

To formally describe such problematic states we define the following function:

$$\text{range}(s) := \sup_{\pi \in \Pi} \max_{a,a' \in \mathcal{A}} q^\pi(s, a) - q^\pi(s, a'), \qquad \text{for all } s \in \mathcal{S}\,.$$

Then, if the range is zero for a state s, we can encounter the issue described above. More generally, we can encounter issues representing the rewards or transitions if there are states with low range.

One way to address this issue is to modify the MDP such that all the rewards and transitions are the same at any states s with range less than a threshold $\alpha \geq 0$. In particular, at each such state, select any action a and make the reward and transition of every other action equal to that of a. We will refer to modifying a state in this way as *skipping* the state via action a. An example of this is shown on the right in Figure 1, where we have skipped state s_1 via the up action. An MDP modified in this way is indeed linear (with parameter norm bound L_θ/α) and thus satisfies Bellman completeness (see Lemma 5). At the end of this subsection we will see how the error due to modifying the MDP depends on α. But first, we formalize the above intuition that the modified MDP is linear.

As we have seen, to modify the MDP we need to know the range function. But, the learner is not given the range function, so it uses a parametric representation of it, to avoid estimating

9. Weisz et al. [2023] studied the online RL setting with linear q^π-realizability. Since then Mhammedi [2025] and Tkachuk et al. [2024] have also made use of this helpful observation.
10. A reader familiar with the work of Weisz et al. [2023] may wonder why our figure is different from their Figure 1. First, this view of modifying the MDP is identical from a value function perspective, and secondly we find this view to be more aligned with the skippy Bellman operators and skippy policies we will define later.

the q-function of every policy. For any $h \in [H]$, let $\Theta_h^\star := \mathrm{cl}(\{\theta_h^\star(\pi) : \pi \in \Pi\}) \subseteq \mathcal{B}(L_\theta)$ be a (compact) set containing the parameter values of all policies. When q^π-realizability holds the range function can be expressed in terms of the parameters in these sets:

$$\mathrm{range}(s) = \sup_{\theta_h^\star \in \Theta_h^\star} \max_{a,a' \in \mathcal{A}} \langle \phi(s,a) - \phi(s,a'),\, \theta_h^\star \rangle, \qquad \text{for all } s \in \mathcal{S}_h, h \in [H]\,.$$

Now we will construct a parametric bound on the range. For all $h \in [H]$, fix a subset $G_h^\star \subseteq \Theta_h^\star$ of size $|G_h^\star| = d_0 := \lceil 4d \log\log(d) + 16 \rceil$ that is the basis of a near-optimal design for $\Theta_h^\star$ (i.e., satisfying Definition 18). The existence of such a near-optimal design follows from [Todd, 2016, Part (ii) of Lemma 3.7]. Any $G = (G_h)_{h\in[H]} \in \mathbf{G}$, where $\mathbf{G} := (\mathcal{B}(L_\theta)^{d_0})^H$, can be used to define an approximate range function that is completely specified by $\tilde{O}(Hd^2)$ parameters:

$$\mathrm{range}_G(s) := \max_{\vartheta \in G_h} \max_{a,a' \in \mathcal{A}} \langle \phi(s,a) - \phi(s,a'),\, \vartheta \rangle, \qquad \text{for all } s \in \mathcal{S}_h, h \in [H]\,.$$

Importantly, when $G^\star := (G_h^\star)_{h\in[H]}$ is used, the $\mathrm{range}_{G^\star}$ function upper bounds the true range:

Lemma 3 (Prop. 4.5 [Weisz et al., 2023]) $\mathrm{range}(s) \le \sqrt{2d} \cdot \mathrm{range}_{G^\star}(s)$, *for all* $s \in \mathcal{S}_h$, $h \in [H]$.

Recall that we want to modify the MDP by skipping over states with range lower than a threshold α. The above lemma tells us that we can achieve this by skipping over states with $\mathrm{range}_{G^\star}(s)$ lower than $\alpha/\sqrt{2d}$. It will be useful to define a skipping function ω_G based on $G \in \mathbf{G}$, that indicates if we skip a state or not. Although this function can be defined as an indicator, for technical reasons[11], we define a smoothed version, which always skips states with $\mathrm{range}_G(s) < \alpha/(2\sqrt{2d})$, never skips for states with $\mathrm{range}_G(s) > \alpha/\sqrt{2d}$, and linearly interpolates between the two thresholds:

$$\omega_G(s) := \begin{cases} 1, & \text{if } s \neq s_\top \text{ and } \mathrm{range}_G(s) \le \alpha/(2\sqrt{2d}); \\ 2 - 2\sqrt{2d} \cdot \mathrm{range}_G(s)/\alpha, & \text{if } s \neq s_\top \text{ and } \alpha/(2\sqrt{2d}) \le \mathrm{range}_G(s) \le \alpha/\sqrt{2d}; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Since each $G \in \mathbf{G}$ defines a skipping function ω_G, which in turn defines a modified MDP, we will refer to G as a *modification* of the MDP, and call $G^\star$ the *correct* modification.

So far we have discussed explicitly modifying the MDP. Instead, we will do so implicitly through modified Bellman operators, which incorporate the skippy behavior into their choice of future q-values[12]. For $h \in [H]$, and $1 \le l \le h$, let $\mathrm{Traj} = (s_t, a_t, r_t)_{l \le t \le H+1}$ be any fixed trajectory that starts from some stage l. For $\tau \in [h+1 : H+1]$ let $F_{G,\mathrm{Traj},h+1}(\tau) = (1 - \omega_G(s_\tau)) \prod_{u=h+1}^{\tau-1} \omega_G(s_u)$ be the probability of stopping at stage τ when starting from state s_h and skipping subsequent states with probability $\omega_G(\cdot)$. For any sequence of objects $a_1, \ldots, a_{H+1}$, define $a_{h\to} := (a_{h+1}, \ldots, a_{H+1})$, for all $h \in [H]$. For a set $\mathcal{C}$ define $\mathcal{C}^{\mathcal{S}\mathcal{A}_{h\to}} := \mathcal{C}^{\mathcal{S}_{h+1}\times\mathcal{A}} \times \cdots \times \mathcal{C}^{\mathcal{S}_{H+1}\times\mathcal{A}}$.

For a policy π and modification G, the *skippy Bellman policy operator* and empirical version $T_G^\pi, \hat{T}_G^\pi : \cup_{h\in[2:H+1]} \mathbb{R}^{\mathcal{S}\mathcal{A}_{h\to}} \to \cup_{h\in[H]} \mathbb{R}^{\mathcal{S}_h\times\mathcal{A}}$ are defined as follows: for all $h \in [H]$, $q_{h\to} \in \mathbb{R}^{\mathcal{S}\mathcal{A}_{h\to}}$,

$$T_G^\pi\, q_{h\to}(s,a) := \mathbb{E}_{\mathrm{Traj}\sim\mathbb{P}_{\pi^b,s,a}} g_G^\pi(q_{h\to}, \mathrm{Traj}), \qquad \text{for all } (s,a) \in \mathcal{S}_h \times \mathcal{A}, \tag{7}$$

$$\hat{T}_G^\pi\, \hat{q}_{h\to} := \arg\min_{q_h \in \widetilde{\mathcal{F}}_h} \tfrac{1}{n} \textstyle\sum_{j=1}^n (q_h(S_h^j, A_h^j) - g_G^\pi(q_{h\to}, \mathrm{Traj}_{h\to}^j))^2, \text{ where} \tag{8}$$

$$g_G^\pi(q_{h\to}, \mathrm{Traj}) := \textstyle\sum_{\tau=h+1}^{H+1} (R_{h:\tau-1} + q_\tau(S_\tau, \pi))\, F_{G,\mathrm{Traj},h+1}(\tau)\,,$$

11. The smooth skipping behavior helps with covering arguments.
12. We find this makes the learners simpler to present (see Sections 4.3 and 4.4), and is also done by Tkachuk et al. [2024].

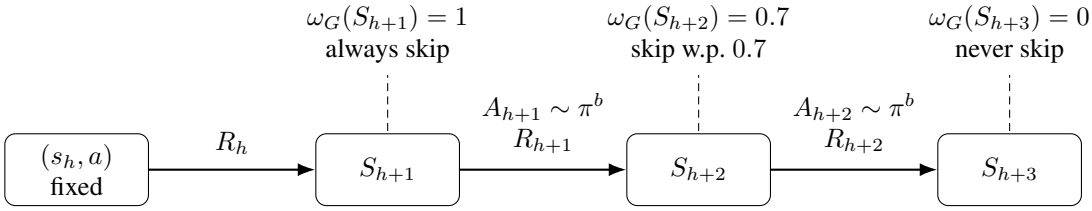

Figure 2: An example of computing the target $g^\pi_G(q_{h\to}, \text{Traj})$ for a trajectory $\text{Traj} = (s_h, a, R_h, S_{h+1}, A_{h+1}, \dots) \sim \mathbb{P}_{\pi^b, s_h, a}$. Starting from (s_h, a), the target is computed by following the trajectory and accumulating rewards while skipping continues, using the value estimate $q_\tau(S_\tau, \pi)$ at the first stage τ where skipping stops, and then averaging over the random stopping time τ. For the displayed trajectory, $\omega_G(S_{h+1}) = 1$, $\omega_G(S_{h+2}) = 0.7$, and $\omega_G(S_{h+3}) = 0$, so skipping stops at stage $h+1$ with probability $F_{G,\text{Traj},h+1}(h+1) = (1 - \omega_G(S_{h+1})) = 0$, stops at stage $h+2$ with probability $F_{G,\text{Traj},h+1}(h+2) = (1 - \omega_G(S_{h+2}))\omega_G(S_{h+1}) = 0.3$ and stops at stage $h+3$ with probability $F_{G,\text{Traj},h+1}(h+3) = (1 - \omega_G(S_{h+3}))\omega_G(S_{h+1})\omega_G(S_{h+2}) = 0.7$. The target is $g^\pi_G(q_{h\to}, \text{Traj}) = 0.3\big(R_h + R_{h+1} + q_{h+2}(S_{h+2}, \pi)\big) + 0.7\big(R_h + R_{h+1} + R_{h+2} + q_{h+3}(S_{h+3}, \pi)\big)$. The values of $\omega_G(S_t)$ were chosen for illustration purposes.

$\text{Traj}^j_{h\to} := (S^j_t, A^j_t, R^j_t)_{t\in[h+1:H+1]}$, and $\widetilde{\mathcal{F}}_h \supseteq \mathcal{F}_h$ is defined in Eq. (10)[13] . Figure 2 shows how the target $g^\pi_G(q_{h\to}, \text{Traj})$ is computed for a single sampled trajectory $\text{Traj} \sim \mathbb{P}_{\pi_b, s_h, a}$, and how the skipping probabilities $\omega_G(S_t)$ determine the distribution over the stopping time τ. If the modification G is such that no states are skipped (i.e., $\tau = h+1$ for all states), then $g^\pi_G(q_{h\to}, \text{Traj}) = R_h + q_{h+1}(S_{h+1}, \pi)$, which reduces Eq. (7) to the usual Bellman policy operator (Eq. (3)). Similarly, for a modification G, the *skippy Bellman optimality operator* and empirical version $T_G, \hat{T}_G : \cup_{h\in[2:H+1]}\mathbb{R}^{\mathcal{SA}_{h\to}} \to \cup_{h\in[H]}\mathbb{R}^{\mathcal{S}_h\times\mathcal{A}}$ are defined as follows: for all $h \in [H]$, $q_{h\to} \in \mathbb{R}^{\mathcal{SA}_{h\to}}$,

$$
\begin{aligned}
T_G\, q_{h\to}(s,a) &:= \mathbb{E}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} g_G(q_{h\to}, \text{Traj}), \quad \text{for all } (s,a) \in \mathcal{S}_h \times \mathcal{A}, \\
\hat{T}_G\, \hat{q}_{h\to} &:= \arg\min_{q_h\in\widetilde{\mathcal{F}}_h} \tfrac{1}{n}\textstyle\sum_{j=1}^n (q_h(S^j_h, A^j_h) - g_G(q_{h\to}, \text{Traj}^j_{h\to}))^2, \text{ where} \\
g_G(q_{h\to}, \text{Traj}) &:= \textstyle\sum_{\tau=h+1}^{H+1} (R_{h:\tau-1} + \max_{a'} q_\tau(S_\tau, a'))\, F_{G,\text{Traj},h+1}(\tau)\,.
\end{aligned} \tag{9}
$$

Remark 4 *Notice that the targets $g^\pi_G(q_{h\to}, \text{Traj}^j_{h\to})$ and $g_G(q_{h\to}, \text{Traj}^j_{h\to})$ are nonlinear functions of the trajectory $\text{Traj}^j_{h\to}$, due to the product form of $F_{G,\text{Traj}^j,h+1}(\tau)$. This is why trajectory data (i.e., samples from the distribution $\mathbb{P}_{\pi^b,s_1}$) are important, and samples from the marginal distributions $\mathbb{P}^h_{\pi^b,s_1}$ are likely not sufficient.*

The reason we will care about these skippy Bellman operators is that they will satisfy two properties that we needed for FQE/FQI to work in Section 4.1. First, by a slight modification to Lemma 4.2 in Tkachuk et al. [2024], if we have q^π-realizability (Assumption 1), then Bellman completeness holds (with a larger function class) for the skippy Bellman operators based on the correct modification $G^\star$ (proof: Appendix C.1).

13. The target $g^\pi_G(q_{h\to}, \text{Traj})$ is equivalent to a state-dependent λ-return [Sutton and Barto, 2018], where the continuation parameter λ is given by the skipping probability ω_G.

Lemma 5 (q^π-realizability $\implies$ skippy Bellman completeness) *Assume $(\mathcal{F}_h)_{h\in[H]}$ satisfies q^π-realizability (Assumption 1). Define a larger function class $\widetilde{\mathcal{F}}_h \supseteq \mathcal{F}_h$ as:*

$$\widetilde{\mathcal{F}}_h := \{f : \mathcal{S}_h \times \mathcal{A} \to \mathbb{R} \mid f(\cdot,\cdot) = \langle \phi(\cdot,\cdot), \theta\rangle, \theta \in \mathcal{B}(\tilde{L}_\theta)\}\,, \text{ for all } h \in [H+1], \tag{10}$$

where $\tilde{L}_\theta := L_\theta \cdot (8H^2 d_0 \sqrt{2d}/\alpha + 1)$. Then, for any $h \in [H]$, $q_{h\to} \in [0,H]^{\mathcal{SA}_{h\to}}$ with $q_{H+1}(s_\top,\cdot) = 0$,

$$T_{G^\star}\, q_{h\to} \in \widetilde{\mathcal{F}}_h \quad \text{and} \quad T^{\pi^e}_{G^\star} q_{h\to} \in \widetilde{\mathcal{F}}_h\,.$$

Notice that we have abused the Bellman completeness naming convention, since the above holds for arbitrary $q_{h+1},\dots,q_{H+1}$, that are not necessarily in $\widetilde{\mathcal{F}}_{h+1},\dots,\widetilde{\mathcal{F}}_{H+1}$ respectively. This is a stronger property than in the regular Bellman completeness assumptions (Assumptions 4 and 5), and will be important for the proofs of Lemmas 9 and 10.

Recall that with the regular Bellman policy operator $q^\pi_h = T^\pi q^\pi_{h+1}$. The second useful property is that a similar relationship holds for the skippy Bellman policy operators, except for skippy policies. In particular, for a policy π define its skippy version based on a modification G as the policy that at each state s chooses an action according to $\pi^b(a|s)$ (this is the skippy action) with probability $\omega_G(s)$ and according to $\pi(a|s)$ with probability $1-\omega_G(s)$[14]:

$$\pi_G(a|s) := \pi^b(a|s)\omega_G(s) + \pi(a|s)(1-\omega_G(s))\,. \tag{11}$$

The following useful property holds for the skippy Bellman policy operator (proof: Appendix C.2):

Lemma 6 $q^{\pi_G}_h = T^\pi_G\, q^{\pi_G}_{h\to}$, *for any policy $\pi \in \Pi$ and modification $G \in \mathbf{G}$.*

This reduces to $q^\pi_h = T^\pi q^\pi_{h+1}$, if G is such that no states are skipped. If we knew $G^\star$, then for policy evaluation, this motivates running FQE with the empirical skippy Bellman policy operator $\hat{T}^{\pi^e}_{G^\star}$.

For policy optimization, we do not have access to $\pi^\star$, and thus cannot estimate $q^{\pi^\star}$ using Eq. (8), since $g^{\pi^\star}_{G^\star}$ depends on $\pi^\star$. Instead, we define a skippy optimal policy $\bar{\pi}^\star_G$ that does not depend on $\pi^\star$: which at state s, with probability $1-\omega_G(s)$, takes the greedy action w.r.t. its future q-values,

$$\bar{\pi}^\star_G(a|s) := \pi^b(a|s)\omega_G(s) + \mathbb{I}\{a = \arg\max_{a'\in\mathcal{A}} q^{\bar{\pi}^\star_G}_h(s,a')\}(1-\omega_G(s))\,.$$

A similar result to $q^{\pi^\star}_h = T q^{\pi^\star}_{h+1}$ for the Bellman optimality operator holds for the skippy Bellman optimality operator with the skippy optimal policy (proof: Appendix C.3):

Lemma 7 $q^{\bar{\pi}^\star_G}_h = T_G\, q^{\bar{\pi}^\star_G}_{h\to}$, *for all modifications $G \in \mathbf{G}$.*

This motivates running FQI with the empirical skippy Bellman optimality operator $\hat{T}_{G^\star}$.

Finally, the error due to modifying a policy with $G^\star$ is controlled by α (proof: Appendix B):

Lemma 8 $|v^\pi - v^{\pi_{G^\star}}| \le H\alpha$, *for any $\pi \in \Pi$, and $v^{\pi^\star} - v^{\bar{\pi}^\star_{G^\star}} \le H\alpha$.*

This result makes intuitive sense: if we skip less states (i.e., make α smaller), then our modification error decreases. But, in Lemma 5 the size of the function class $\widetilde{\mathcal{F}}$ increases as we make α smaller. This shows how α trades off between modification error and function class size.

Unfortunately, the above approaches for policy evaluation and optimization require knowledge of $G^\star$, which we do not have. We are not aware of a sample efficient way to learn the correct modification $G^\star$, since learning the range function seems at least as hard as solving the original problem. Fortunately, in the next two sections we will see that we do not need to learn $G^\star$ directly.

14. There is a relationship between a skippy policy π_G and a modified MDP based on G. In particular, the q-function for the skippy policy q^{π_G} in the original MDP is equal to the q-function for π, q^π, in the modified MDP.

4.3. A Learner for Policy Optimization with q^π-realizability

In this subsection we present the policy optimization learner (Algorithm 1) by Tkachuk et al. [2024], which is sample efficient when q^π-realizability holds (see Theorem 2).

Recall that if we knew $G^\star$ we would simply run FQI, which estimates $q^{\bar{\pi}^\star_{G^\star}}$ well, and then greedifies. Since we do not know $G^\star$, we construct a set of estimates,

$$\mathbf{Q}_{\text{opt}} = \{q^G \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : q^G_h = \hat{T}_G\, q^G_{h\to},\ \text{for } h \in [H], G \in \mathbf{G}_{\text{opt}},\ q^G_{H+1} = \mathbf{0}\}\,,$$

that contains at least one good estimate of $q^{\bar{\pi}^\star_{G^\star}}$, and every element of $\mathbf{Q}_{\text{opt}}$ is a good estimate of some $q^{\bar{\pi}^\star_G}, G \in \mathbf{G}_{\text{opt}}$. The set $\mathbf{G}_{\text{opt}} \subseteq \mathbf{G}$ is defined in Section 4.5, and uses the existence of $G^\star$ to ensure the preceding claims hold. The next lemma formalizes the above claims (proof: Appendix C.4).

Lemma 9 ($\mathbf{Q}_{\text{opt}}$ guarantee) *For all $\delta \in (0,1)$, with probability at least $1-\delta$,*

1. *for all $q^G \in \mathbf{Q}_{opt}$, admissible distributions $\nu = (\nu_t)_{t\in[H]}$, and stages $h \in [H]$,*
   $\mathbb{E}_{(S,A)\sim\nu_h}|q^G(S,A) - q^{\bar{\pi}^\star_G}(S,A)| \le \tilde{\epsilon} = \tilde{\mathcal{O}}(\log(1/\alpha)C_0^{3/2}H^{5/2}d^{3/2}n^{-1/2})$, *defn Eq. (33)*,
2. *and there exists a $q^G \in \mathbf{Q}_{opt}$ such that $G = G^\star$.*

Although we use the $\tilde{\mathcal{O}}$ notation to hide logarithmic factors, we explicitly write the $\log(1/\alpha)$ factor, since it will be important to keep track of it when we set α at the end. For exposition only, we make two simplifying assumptions[15]: (i) in the first claim of the lemma we have perfect estimates (i.e., $\tilde{\epsilon} = 0$), and (ii) that the bound holds pointwise instead of in expectation. Together, these simplifications imply the following very useful properties: $\mathbf{Q}_{\text{opt}}$ **only** contains the q-functions of $\bar{\pi}^\star_G$ under different modifications (i.e., $\mathbf{Q}_{\text{opt}} \subseteq \{q^{\bar{\pi}^\star_G} : G \in \mathbf{G}\}$), and $q^{\bar{\pi}^\star_{G^\star}} \in \mathbf{Q}_{\text{opt}}$.

Based on these properties, this suggests a natural strategy for the learner. Recall that $q^{\pi^\star} \ge q^\pi$ for any policy π, and that $q^{\pi^\star} - q^{\bar{\pi}^\star_{G^\star}} \le H\alpha$ (Lemma 8). Thus, the q-function in $\mathbf{Q}_{\text{opt}}$ that has the highest value at the start state must estimate $q^{\pi^\star}(s_1)$ with error at most $H\alpha$. We have also seen at the end of Section 4.1 that greedifying w.r.t. a good estimate of $q^{\pi^\star}$ gives a near optimal policy. Thus, the learner should simply select the q-function in $\mathbf{Q}_{\text{opt}}$ that has the highest value at the start state, and greedify w.r.t. it. Fortunately, this same idea extends to the case without the simplifying assumptions (see Appendix B). The final error is $H\alpha + 2H\tilde{\epsilon}$ (see Eq. (25)), and α is set to balance the two terms. This learner is shown in Algorithm 1, and given the name LIN-q^π-FQI since the construction of $\mathbf{Q}_{\text{opt}}$ is related to FQI. The analysis in Appendix B formalizes the above steps.

Algorithm 1 LIN-q^π-FQI

input: accuracy ϵ, failure probability δ, concentrability coefficient C_0, trajectories $(\text{Traj}^j)_{j\in[n]}$.
1: $\hat{q} \leftarrow \arg\max_{q\in\mathbf{Q}_{\text{opt}}} \max_a q(s_1,a)$
2: **return** $\hat{\pi}(s) \leftarrow \arg\max_a \hat{q}(s,a)$

LIN-q^π-FQI is the learner defined by Tkachuk et al. [2024]. The only difference is in the presentation, since Tkachuk et al. [2024] define $\hat{\pi}$ as the solution to an optimization problem, where their feasible set is closely related to $\mathbf{Q}_{\text{opt}}$. We have explicitly defined the set $\mathbf{Q}_{\text{opt}}$ since it will be helpful for understanding the learner we present for policy evaluation in the next section.

We are not aware of a computationally efficient way to implement LIN-q^π-FQI. One of the difficulties is that the feasible set $\mathbf{Q}_{\text{opt}}$ is not convex, since it is defined based on $\hat{T}_G$, which depends on a product of the non-convex function ω_G (see Eq. (9) and the definition of $F_{G,\text{Traj},h+1}$).

15. The simplifying assumptions are not needed for the actual analysis, and are only made for presentation clarity.

4.4. Our Learner for Policy Evaluation with q^π-realizability

We are now ready to present our main contribution: Algorithm 2, which is a sample efficient learner for policy evaluation when q^π-realizability holds (see Theorem 1). Similar to the previous section, since we do not know $G^\star$, we construct a set,

$$\mathbf{Q}_{\text{eval}} = \{q^G \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : q_h^G = \hat{T}_G^{\pi^e} q_{h\to}^G, \text{ for } h \in [H], G \in \mathbf{G}_{\text{eval}},\, q_{H+1}^G = \mathbf{0}\}\,, \tag{12}$$

that contains at least one good estimate of $q^{\pi^e_{G^\star}}$, and every element of $\mathbf{Q}_{\text{eval}}$ is a good estimate of some $q^{\pi^e_G}, G \in \mathbf{G}_{\text{eval}}$. We defer the definition of $\mathbf{G}_{\text{eval}}$ to Section 4.5. The claims are formally stated in the following lemma (which parallels Lemma 9, proof: Appendix C.5).

Lemma 10 ($\mathbf{Q}_{\text{eval}}$ guarantee) *For all $\delta \in (0,1)$, with probability at least $1-\frac{\delta}{2}$,*

1. *for all $q^G \in \mathbf{Q}_{eval}$, admissible distributions $\nu = (\nu_t)_{t\in[H]}$, and stages $h \in [H]$,*
$$\mathbb{E}_{(S,A)\sim\nu_h} |q^G(S,A) - q^{\pi^e_G}(S,A)| \le \tilde{\epsilon}\,,$$
2. *and there exists a $q^G \in \mathbf{Q}_{eval}$ such that $G = G^\star$.*

Similar to the previous section, to streamline the explanation, we make the following simplifying assumptions[16]: (i) we have perfect estimates (i.e., $\tilde{\epsilon} = 0$), and (ii) that the bound holds pointwise instead of in expectation. These, in turn, imply the following useful properties: $\mathbf{Q}_{\text{eval}}$ **only** contains the q-functions of π^e_G under different modifications (i.e., $\mathbf{Q}_{\text{eval}} \subseteq \{q^{\pi^e_G} : G \in \mathbf{G}\}$), and $q^{\pi^e_{G^\star}} \in \mathbf{Q}_{\text{eval}}$. In other words, we have access to a set $\mathbf{Q}_{\text{eval}}$, which we know contains a good estimate of q^{π^e} (namely $q^{\pi^e_{G^\star}}$); however, it can also contain q-functions of other policies. The question is how to select a q-function from $\mathbf{Q}_{\text{eval}}$ that is a good estimate of q^{π^e} from the start state?

The main contribution of our work is providing a sample efficient answer to the above question. To build some intuition for how we can do this let us consider the difference between the policy optimization and evaluation objectives. In policy evaluation we know the policy π^e, which we are tasked with evaluating; however, we do not know the q-value of π^e. On the other hand, in policy optimization we do not know the optimal policy $\pi^\star$; however, we do know that the value function of $\pi^\star$ is the largest possible (i.e., $q^{\pi^\star} \ge q^\pi$ for any policy π). Importantly, we made use of the latter to design LIN-q^π-FQI when we selected the q-function in $\mathbf{Q}_{\text{opt}}$ that had the largest value at the start state. Unfortunately, the same strategy does not work for policy evaluation since we do not know the value of π^e. Instead, we make use of our knowledge of π^e.

If our learner outputs the value estimate $q^{\pi^e_G}(s_1, \pi^e_G) = v^{\pi^e_G}(s_1)$ for some $q^{\pi^e_G} \in \mathbf{Q}_{\text{eval}}$, then, by the performance difference lemma (Lemma 29),

$$|v^{\pi^e}(s_1) - v^{\pi^e_G}(s_1)| = |\textstyle\sum_{h=1}^H \mathbb{E}_{(S_h,A_h)\sim\mathbb{P}^h_{\pi^e,s_1}} (q^{\pi^e_G}(S_h,\pi^e) - q^{\pi^e_G}(S_h,\pi^e_G))|\,. \tag{13}$$

The quantity inside the expectation is often called the *advantage*, and a bound on the expected advantage for all h, clearly implies a bound on the error of our learner's value estimate. Importantly, the expected advantage only depends on $q^{\pi^e_G}$, π^e, and π^e_G, all of which are known, and does not depend on q^{π^e}, which we are trying to estimate and is thus unknown.

16. As before, the simplifying assumptions are not needed for the actual analysis, and just for presentation.

Remark 11 *Knowing π^e_G requires knowing π^b and G. This is why we assume that we know π^b in Theorem 1 and Algorithm 2. Removing this assumption is an interesting direction for future work, and has been studied in related settings [Zhan et al., 2022; Ozdaglar et al., 2023]. Furthermore, to have access to G we should have defined the set $\mathbf{Q}_{eval}$ to contain pairs (q^G, G); however, to simplify notation we have omitted G and assume we know it implicitly for each $q^G \in \mathbf{Q}_{eval}$.*

Notice that when $G = G^\star$, $\max_{s\in\mathcal{S}} \left|q^{\pi^e_{G^\star}}(s,\pi^e) - q^{\pi^e_{G^\star}}(s,\pi^e_{G^\star})\right| \le \alpha$, since π^e and $\pi^e_{G^\star}$ only differ on states s with range$(s) \le \alpha$. Under the simplifying assumptions (where we have pointwise access to $q^{\pi^e_G}$), the learner should select the $q^{\pi^e_G} \in \mathbf{Q}_{\text{eval}}$ that minimizes $\max_{s\in\mathcal{S}} \left|q^{\pi^e_G}(s,\pi^e) - q^{\pi^e_G}(s,\pi^e_G)\right|$, since the value estimate $v^{\pi^e_G}(s_1)$ will be at most $H\alpha$ away from $v^{\pi^e}(s_1)$ (by Eq. (13)).

Unfortunately, this idea does not quite work when we only have approximately good estimates of $q^{\pi^e_G}$ in expectation (as in Lemma 10). The reason for this is that the learner would now need to select the $q^G \in \mathbf{Q}_{\text{eval}}$ that minimizes $\max_{s\in\mathcal{S}} \left|q^G(s,\pi^e) - q^G(s,\pi^e_G)\right|$. With the simplifying assumptions, to show this quantity would not be too large we relied on the fact that $\max_{s\in\mathcal{S}} \left|q^{\pi^e_{G^\star}}(s,\pi^e) - q^{\pi^e_{G^\star}}(s,\pi^e_{G^\star})\right| \le \alpha$. To make use of this again we would need a pointwise bound on $|q^G(s,a) - q^{\pi^e_{G^\star}}(s,a)|$ for some $q^G \in \mathbf{Q}_{\text{eval}}$. But, by the second claim in Lemma 10 we only have a bound in expectation. Fortunately, this will be enough, since if we look back at Eq. (13) we see that we only need to bound the **expected** advantage.

As such, the learner (that does indeed work without any simplifying assumptions) minimizes the following estimate of the expected advantage over all $q^G \in \mathbf{Q}_{\text{eval}}$

$$\max_{h\in[H]} \tfrac{1}{n}\textstyle\sum_{j=1}^n |q^G(S^j_h,\pi^e) - q^G(S^j_h,\pi^e_G)|\,, \tag{14}$$

and then outputs the value estimate $\hat{v} = q^{\hat{G}}(s_1,\pi^e_{\hat{G}})$, where $\hat{G} \in \mathbf{G}_{\text{eval}}$ is such that $q^{\hat{G}}$ is the minimizer of Eq. (14) over $\mathbf{Q}_{\text{eval}}$. The final error is $\tilde{\mathcal{O}}(H\alpha + C_0 H\tilde{\epsilon})$ (see Eq. (21)), and α is set to balance the two terms. We call this learner LIN-q^π-FQE (Algorithm 2), since the construction of $\mathbf{Q}_{\text{eval}}$ (see Section 4.5) is based on FQE[17] . The analysis in Appendix A formalizes the above steps.

Algorithm 2 LIN-q^π-FQE

input: accuracy ϵ, fail prob δ, concentrability C_0, trajectories $(\text{Traj}^j)_{j\in[n]}$, behavior policy π^b, eval policy π^e

1: $q^{\hat{G}} \leftarrow \arg\min_{q^G\in\mathbf{Q}_{\text{eval}}} \max_{h\in[H]} \frac{1}{n}\sum_{j=1}^n |q^G(S^j_h,\pi^e) - q^G(S^j_h,\pi^e_G)|$

2: **return** $\hat{v} \leftarrow q^{\hat{G}}(s_1,\pi^e_{\hat{G}})$

4.5. Defining $\mathbf{G}_{\text{opt}}$ and $\mathbf{G}_{\text{eval}}$

In this section, we define the sets $\mathbf{G}_{\text{opt}}$ and $\mathbf{G}_{\text{eval}}$ which are used to define $\mathbf{Q}_{\text{opt}}$ and $\mathbf{Q}_{\text{eval}}$ in Sections 4.3 and 4.4 respectively. The set $\mathbf{G}_{\text{opt}}$ is the same as the feasible set in the optimization problem defined by Tkachuk et al. [2024]. Since the modified function class $(\widetilde{\mathcal{F}}_h)_{h\in[H]}$ contains bounded linear functions, it will be useful to define the following notation. For $x \in \mathbb{R}$, let $\text{clip}_{[0,H]}x \coloneqq \max\{0, \min\{H, x\}\}$. For $h \in [H], \theta \in \mathbb{R}^d, \theta_{h\to} \in (\mathbb{R}^d)^{[h+1:H+1]}$, let

$$q_\theta(\cdot,\cdot) \coloneqq \langle\phi(\cdot,\cdot),\,\theta\rangle, \quad q_{\theta_{h\to}} \coloneqq (q_{\theta_{h+1}},\ldots,q_{\theta_{H+1}}),$$
$$\bar{q}_\theta(\cdot,\cdot) \coloneqq \text{clip}_{[0,H]}q_\theta(\cdot,\cdot), \quad \bar{q}_{\theta_{h\to}} \coloneqq (\bar{q}_{\theta_{h+1}},\ldots,\bar{q}_{\theta_{H+1}}).$$

17. Similar to LIN-q^π-FQI, we are not aware of a computationally efficient way to implement LIN-q^π-FQE since the feasible set $\mathbf{Q}_{\text{eval}}$ is not convex for the same reasons as $\mathbf{Q}_{\text{opt}}$.

Let $\hat{X}_h := \lambda I + \sum_{j\in[n]} \phi(S_h^j, A_h^j)\phi(S_h^j, A_h^j)^\top$, with $\sqrt{\lambda} = H^{3/2}d/\tilde{L}_\theta$ (defn Eq. (35)), be the unnormalized empirical covariance matrix at stage $h \in [H]$. Now we define sets based on least squares estimates, that will be used to define $\mathbf{G}_{\text{opt}}$:

$$\hat{\Theta}^{\text{opt}}_{G,h} := \{\arg\min_{\hat{\theta}\in\mathcal{B}(\tilde{L}_\theta)} \tfrac{1}{n}\textstyle\sum_{j=1}^n (q_{\hat{\theta}}(S_h^j, A_h^j) - g_G(\bar{q}_{\theta_{h\to}}, \text{Traj}^j_{h\to}))^2 : \theta_{h\to} \in \bigtimes_{u=h+1}^{H+1} \Theta^{\text{opt}}_{G,u}\},$$
$$\Theta^{\text{opt}}_{G,h} := \{\theta_h \in \mathcal{B}(\tilde{L}_\theta) : \min_{\hat{\theta}_h\in\hat{\Theta}^{\text{opt}}_{G,h}} \|\theta_h - \hat{\theta}_h\|_{\hat{X}_h} \le \beta\},\ \Theta^{\text{opt}}_{G,H+1} = \{\mathbf{0}\},$$

where $\beta = \tilde{\mathcal{O}}(\log(1/\alpha)H^{3/2}d)$ (defn Eq. (36)). The sets $\Theta^{\text{opt}}_{G,h}$ should be thought of as expanding the least squares estimates in $\hat{\Theta}^{\text{opt}}_{G,h}$ to include parameters within a β neighborhood. The next lemma shows that the effect of this is that the parameter $\theta^\star_h(\bar{\pi}^\star_G)$ that realizes the q-function of policy $\bar{\pi}^\star_G$ will be in the set $\Theta^{\text{opt}}_{G,h}$ for all modifications $G \in \mathbf{G}$.

Lemma 12 (Lemma D.2 in [Tkachuk et al., 2024]) *For any $\delta \in (0,1)$, with probability at least $1-\delta/5$, for all $G \in \mathbf{G}$, and $h \in [H+1]$, it holds that $\theta^\star_h(\bar{\pi}^\star_G) \in \Theta^{opt}_{G,h}$.*

Due to the above lemma and skippy Bellman completeness (Lemma 5) holding for $G^\star$, the q-function based on any parameter in $\Theta^{\text{opt}}_{G^\star,h}$ will be a good estimate of $q_{\theta^\star_h(\bar{\pi}^\star_{G^\star})} = q_h^{\bar{\pi}^\star_{G^\star}}$. This implies that we can construct a non-empty set based on a data dependent check of this condition:

$$\mathbf{G}_{\text{opt}} := \{G \in \mathbf{G} : \tfrac{1}{n}\textstyle\sum_{j=1}^n (\max_{\theta\in\Theta^{\text{opt}}_{G,h}} \bar{q}_\theta(S_h^j, A_h^j) - \min_{\theta\in\Theta^{\text{opt}}_{G,h}} \bar{q}_\theta(S_h^j, A_h^j)) \le \bar{\epsilon},\ \forall h \in [H]\},$$

where $\bar{\epsilon} = \tilde{\mathcal{O}}(\log(1/\alpha)C_0^{1/2}H^{5/2}d^{3/2}n^{-1/2})$ (defn Eq. (41)).

Next, we define the set $\mathbf{G}_{\text{eval}}$. In a similar manner to $\hat{\Theta}^{\text{opt}}_{G,h}$ and $\Theta^{\text{opt}}_{G,h}$, we define:

$$\hat{\Theta}^{\text{eval}}_{G,h} := \{\arg\min_{\hat{\theta}\in\mathcal{B}(\tilde{L}_\theta)} \tfrac{1}{n}\textstyle\sum_{j=1}^n (q_{\hat{\theta}}(S_h^j, A_h^j) - g_G^{\pi^e}(\bar{q}_{\theta_{h\to}}, \text{Traj}^j_{h\to}))^2 : \theta_{h\to} \in \bigtimes_{u=h+1}^{H+1} \Theta^{\text{eval}}_{G,u}\},$$
$$\Theta^{\text{eval}}_{G,h} := \{\theta_h \in \mathcal{B}(\tilde{L}_\theta) : \min_{\hat{\theta}_h\in\hat{\Theta}^{\text{eval}}_{G,h}} \|\theta_h - \hat{\theta}_h\|_{\hat{X}_h} \le \beta\},\ \Theta^{\text{eval}}_{G,H+1} = \{\mathbf{0}\}.$$

The only difference between $\hat{\Theta}^{\text{opt}}_{G,h}$ and $\hat{\Theta}^{\text{eval}}_{G,h}$ is that the latter uses the target $g_G^{\pi^e}$ instead of g_G. The set $\mathbf{G}_{\text{eval}}$ is defined similarly to $\mathbf{G}_{\text{opt}}$, but instead checks if the q-values based on $\Theta^{\text{eval}}_{G,h}$ are close:

$$\mathbf{G}_{\text{eval}} := \{G \in \mathbf{G} : \tfrac{1}{n}\textstyle\sum_{j=1}^n (\max_{\theta\in\Theta^{\text{eval}}_{G,h}} \bar{q}_\theta(S_h^j, A_h^j) - \min_{\theta\in\Theta^{\text{eval}}_{G,h}} \bar{q}_\theta(S_h^j, A_h^j)) \le \bar{\epsilon},\ \forall h \in [H]\}$$

Acknowledgments

We thank Alex Ayoub and Kushagra Chandak for many helpful discussions. V.T. acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [PGS D - 600128 - 2025]. X.T. acknowledges funding from the Alberta Major Innovation Fund (Amii) and NSERC Discovery Grant [RGPIN-2022-03646]. C.S. acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC.

References

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, pages 3489–3489. PMLR, 2022.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

Zeyu Jia, Alexander Rakhlin, Ayush Sekhari, and Chen-Yu Wei. Offline reinforcement learning: Role of state aggregation and trajectory data. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2644–2719. PMLR, 2024.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2024.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

Haolin Liu, Braham Snyder, and Chen-Yu Wei. On the complexity of offline reinforcement learning with q-approximation and partial coverage. *arXiv preprint arXiv:2602.12107*, 2026.

Zakaria Mhammedi. Sample and oracle efficient reinforcement learning for MDPs with linearly-realizable value functions. In *Proceedings of Thirty Eighth Conference on Learning Theory*, pages 4078–4165. PMLR, 2025.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Asuman E Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Revisiting the linear-programming framework for offline RL with general function approximation. In *International Conference on Machine Learning*, pages 26769–26791. PMLR, 2023.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Volodymyr Tkachuk, Gellért Weisz, and Csaba Szepesvári. Trajectory data suffices for statistically efficient learning in offline RL with linear q^π-realizability and concentrability. In *Advances in Neural Information Processing Systems*, volume 37, pages 83268–83313, 2024.

Michael J Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Gellért Weisz, András György, and Csaba Szepesvári. Online RL in linearly q^π-realizable MDPs is as easy as in linear MDPs if you learn what to ignore. In *Advances in Neural Information Processing Systems*, volume 36, pages 59172–59205, 2023.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

In Appendix A we will prove our main result (Theorem 1). For completeness we also prove Theorem 2 in Appendix B, which follows the exact same steps as [Tkachuk et al., 2024]. The missing proofs and definitions from Section 4 are given in Appendix C.

Appendix A. Analysis of LIN-q^π-FQE

Before proceeding we introduce a useful lemma that makes use of the concentrability coefficient (Assumption 3) and will be used several times in this section.

Lemma 13 *Let $\nu = (\nu_h)_{h\in[H]}$ be an admissible distribution and let $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ be a non-negative function. Then, for all $h \in [H]$, $\mathbb{E}_{(S,A)\sim\nu_h} f(S,A) \le C_0 \cdot \mathbb{E}_{(S,A)\sim\mu_h} f(S,A)$.*

Proof $\mathbb{E}_{(S,A)\sim\nu_h} f(S,A) = \mathbb{E}_{(S,A)\sim\mu_h}[f(S,A)\cdot \nu_h(S,A)/\mu_h(S,A)] \le C_0 \cdot \mathbb{E}_{(S,A)\sim\mu_h} f(S,A)$. ■

Since we will make use of Lemma 10 multiple times, define $\bar{\mathcal{E}}$ to be the event (which occurs with probability at least $1-\delta/2$) on which the two claims of Lemma 10 hold. An important property that we will use throughout is that since $(\widetilde{\mathcal{F}}_h)_{h\in[H]}$ are linear function classes, each $q^G \in \mathbf{Q}^{\text{all}}_{\text{eval}}$, where

$$\mathbf{Q}^{\text{all}}_{\text{eval}} := \left\{ q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : q_h = \hat{T}^{\pi^e}_G\, q_{h\to},\ \text{ for } h\in[H], G\in\mathbf{G},\ q_{H+1} = \mathbf{0} \right\} \supseteq \mathbf{Q}_{\text{eval}}\,,$$

is based on H least squares solutions. Thus for each $h\in[H]$, $q^G_h = \bar{q}_{\theta_h}$ for some $\theta_h \in \hat{\Theta}^{\text{eval}}_{G,h}$. Before proving our main result (Theorem 1), we state some useful definitions and lemmas.

First, we show that the estimate of expected advantage in Eq. (14) is close to its expectation under the distribution μ_h. Define the event

$$\tilde{\mathcal{E}} := \bigcap_{h\in[H],a,a'\in\mathcal{A}} \tilde{\mathcal{E}}_{h,a,a'}, \quad \text{where} \tag{15}$$

$$\tilde{\mathcal{E}}_{h,a,a'} := \left\{ \left| \mathop{\mathbb{E}}_{(S,A)\sim\mu_h} \left|q^G(S,a) - q^G(S,a')\right| - \frac{1}{n}\sum_{j=1}^{n} \left|q^G(S^j_h,a) - q^G(S^j_h,a')\right| \right| \le \zeta_2,\ \forall q^G \in \mathbf{Q}^{\text{all}}_{\text{eval}} \right\},$$

$$\zeta_2 = \tilde{\mathcal{O}}\left(\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{d}H}{\sqrt{n}}\right), \quad \text{defined in Eq. (19)}\,.$$

Lemma 14 *For any $\delta\in(0,1)$, the probability of event $\tilde{\mathcal{E}}$ is at least $1-\delta/2$.*

Proof Recall that for each $q^G \in \mathbf{Q}^{\text{all}}_{\text{eval}}$ there exists a $\theta \in \hat{\Theta}^{\text{eval}}_{G,h}$ such that $q^G_h = \bar{q}_\theta$, for all $h\in[H]$. Since $\hat{\Theta}^{\text{eval}}_{G,h} \subseteq \mathcal{B}(\tilde{L}_\theta)$ the idea will be to show the result for a cover of $\mathcal{B}(\tilde{L}_\theta)$, and then relate any $\theta \in \hat{\Theta}^{\text{eval}}_{G,h}$ back to the cover.

By Lemma 28, we know there exists a set

$$C_\xi \subset \mathcal{B}(\tilde{L}_\theta) \quad \text{with} \quad |C_\xi| = (1+2\tilde{L}_\theta/\xi)^d \quad \text{where} \quad \xi > 0\,, \tag{16}$$

such that, for any $\theta \in \mathcal{B}(\tilde{L}_\theta)$ there exists a $\tilde{\theta} \in C_\xi$ such that $\|\theta - \tilde{\theta}\|_2 \le \xi$. Notice that for any $G \in \mathbf{G}$, $h \in [H]$ and $\theta \in \Theta^{\text{eval}}_{G,h} \subset \mathcal{B}(\tilde{L}_\theta)$, there exists a $\tilde{\theta} \in C_\xi$ such that, for any $(s,a) \in \mathcal{S}_h \times \mathcal{A}$,

$$\begin{aligned}
\left|\bar{q}_\theta(s,a) - \bar{q}_{\tilde{\theta}}(s,a)\right| &= \left|\text{clip}_{[0,H]}\langle \phi(s,a),\, \theta\rangle - \text{clip}_{[0,H]}\langle \phi(s,a),\, \tilde{\theta}\rangle\right| \\
&\le \left|\langle \phi(s,a),\, \theta - \tilde{\theta}\rangle\right| \\
&\le \|\phi(s,a)\|_2 \|\theta - \tilde{\theta}\|_2 \\
&\le L_\phi \xi\,.
\end{aligned}$$

Thus, for any $a, a' \in \mathcal{A}$,

$$\begin{aligned}
&\Big|\left|\bar{q}_\theta(s,a) - \bar{q}_\theta(s,a')\right| - \left|\bar{q}_{\tilde{\theta}}(s,a) - \bar{q}_{\tilde{\theta}}(s,a')\right|\Big| \\
&\le \Big|\left|\bar{q}_\theta(s,a) - \bar{q}_{\tilde{\theta}}(s,a)\right| + \left|\bar{q}_{\tilde{\theta}}(s,a) - \bar{q}_{\tilde{\theta}}(s,a')\right| + \left|\bar{q}_{\tilde{\theta}}(s,a') - \bar{q}_\theta(s,a')\right| - \left|\bar{q}_{\tilde{\theta}}(s,a) - \bar{q}_{\tilde{\theta}}(s,a')\right|\Big| \\
&= \left|\bar{q}_\theta(s,a) - \bar{q}_{\tilde{\theta}}(s,a)\right| + \left|\bar{q}_{\tilde{\theta}}(s,a') - \bar{q}_\theta(s,a')\right| \le 2L_\phi \xi\,. \qquad (17)
\end{aligned}$$

For any $\tilde{\theta} \in C_\xi$, $h \in [H]$, $a, a' \in \mathcal{A}$, define the event

$$\begin{aligned}
\check{\mathcal{E}}_{\tilde{\theta},h,a,a'} := \Bigg\{ &\left| \mathop{\mathbb{E}}_{(S,A)\sim\mu_h} \left|\bar{q}_{\tilde{\theta}}(S,a) - \bar{q}_{\tilde{\theta}}(S,a')\right| - \frac{1}{n}\sum_{j=1}^n \left|\bar{q}_{\tilde{\theta}}(S_h^j,a) - \bar{q}_{\tilde{\theta}}(S_h^j,a')\right| \right| \\
&\le \frac{H}{\sqrt{n}} \sqrt{\log\left(\frac{4H|\mathcal{A}|^2|C_\xi|}{\delta}\right)} \Bigg\}\,.
\end{aligned}$$

Then, since $|\bar{q}_{\tilde{\theta}}(s,a) - \bar{q}_{\tilde{\theta}}(s,a')| \in [0,H]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, by Hoeffding's inequality (Lemma 27), we have that, event $\check{\mathcal{E}}_{\tilde{\theta},h,a,a'}$ occurs with probability at least $1 - \delta/(2H|\mathcal{A}|^2|C_\xi|)$. Let

$$\check{\mathcal{E}}_{h,a,a'} := \bigcap_{\tilde{\theta}\in C_\xi} \check{\mathcal{E}}_{\tilde{\theta},h,a,a'}\,. \qquad (18)$$

Then, by applying a union bound over $\tilde{\theta} \in C_\xi$ we have that the probability of $\check{\mathcal{E}}_{h,a,a'}$ is at least $1 - \delta/(2H|\mathcal{A}|^2)$.

Assume we are on event $\check{\mathcal{E}}_{h,a,a'}$. Fix any $q^G \in \mathbf{Q}^{\text{all}}_{\text{eval}}$, and let $\theta \in \hat{\Theta}^{\text{eval}}_{G,h}$ be such that $q^G_h = \bar{q}_\theta$. Let $\xi > 0$ be a parameter to be chosen later (see Eq. (19)). By Eq. (17),

$$\left| \mathop{\mathbb{E}}_{(S,A)\sim\mu_h} \left|\bar{q}_\theta(S,a) - \bar{q}_\theta(S,a')\right| - \mathop{\mathbb{E}}_{(S,A)\sim\mu_h} \left|\bar{q}_{\tilde{\theta}}(S,a) - \bar{q}_{\tilde{\theta}}(S,a')\right| \right| \le 2L_\phi \xi\,.$$

Since we are on event $\check{\mathcal{E}}_{h,a,a'}$,

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{(S,A)\sim\mu_h} \left|\bar{q}_{\tilde{\theta}}(S,a) - \bar{q}_{\tilde{\theta}}(S,a')\right| - \frac{1}{n}\sum_{j=1}^n \left|\bar{q}_{\tilde{\theta}}(S_h^j,a) - \bar{q}_{\tilde{\theta}}(S_h^j,a')\right| \right| \\
&\le \frac{H}{\sqrt{n}} \sqrt{\log\left(\frac{4H|\mathcal{A}|^2|C_\xi|}{\delta}\right)}\,.
\end{aligned}$$

By Eq. (17),

$$\left|\frac{1}{n}\sum_{j=1}^{n}\left|\bar{q}_{\tilde{\theta}}(S_h^j,a)-\bar{q}_{\tilde{\theta}}(S_h^j,a')\right|-\frac{1}{n}\sum_{j=1}^{n}\left|\bar{q}_{\theta}(S_h^j,a)-\bar{q}_{\theta}(S_h^j,a')\right|\right|\le 2L_\phi\xi\,.$$

Putting the above results together via three triangle inequalities and recalling that $q_h^G=\bar{q}_\theta$, we have that with probability at least $1-\delta/(2H|\mathcal{A}|^2)$,

$$\begin{aligned}
&\Big|\mathop{\mathbb{E}}_{(S,A)\sim\mu_h}\left|q^G(S,a)-q^G(S,a')\right|-\frac{1}{n}\sum_{j=1}^{n}\left|q^G(S_h^j,a)-q^G(S_h^j,a')\right|\Big|\\
&\le\frac{H}{\sqrt{n}}\sqrt{\log\left(\frac{4H|\mathcal{A}|^2|C_\xi|}{\delta}\right)}+4L_\phi\xi\\
&\le\frac{\sqrt{d}H}{\sqrt{n}}\sqrt{\log\left(\frac{4H|\mathcal{A}|^2(1+8\sqrt{n}L_\phi\tilde{L}_\theta/(\sqrt{d}H))}{\delta}\right)}+\frac{\sqrt{d}H}{\sqrt{n}}=:\zeta_2\,,
\end{aligned}\tag{19}$$

where the last inequality follows by setting $\xi=\sqrt{d}H/(4\sqrt{n}L_\phi)$. Notice the above event is exactly $\tilde{\mathcal{E}}_{h,a,a'}$ as defined in Eq. (15). Finally, applying a union bound over $h\in[H]$ and $a,a'\in\mathcal{A}$ gives that the probability of event $\tilde{\mathcal{E}}$ is at least $1-\delta/2$. ■

Next, we show that there exists a $q^G\in\mathbf{Q}_{\text{eval}}$ (in particular $q^{G^\star}$) for which Eq. (14) is small, and thus the minimum must be no larger.

Lemma 15 *On event $\bar{\mathcal{E}}\cap\tilde{\mathcal{E}}$, the minimum over $q^G\in\mathbf{Q}_{eval}$ of Eq. (14) is at most $\alpha+2\tilde{\epsilon}+\zeta_2$, i.e.,*

$$\min_{q^G\in\mathbf{Q}_{eval}}\max_{h\in[H]}\frac{1}{n}\sum\nolimits_{j=1}^{n}\left|q^G(S_h^j,\pi^e)-q^G(S_h^j,\pi^e_G)\right|\le\alpha+2\tilde{\epsilon}+\zeta_2=\tilde{\mathcal{O}}\left(\alpha+\log\left(\frac{1}{\alpha}\right)\frac{C_0^{3/2}H^{5/2}d^{3/2}}{\sqrt{n}}\right).$$

Proof On event $\tilde{\mathcal{E}}$ for all $h\in[H]$,

$$\frac{1}{n}\sum_{j=1}^{n}\left|q^{G^\star}(S_h^j,\pi^e)-q^{G^\star}(S_h^j,\pi^e_{G^\star})\right|\le\mathop{\mathbb{E}}_{(S,A)\sim\mu_h}\left|q^{G^\star}(S,\pi^e)-q^{G^\star}(S,\pi^e_{G^\star})\right|+\zeta_2\,.$$

On event $\bar{\mathcal{E}}$, by the second claim in Lemma 10, $q^{G^\star}\in\mathbf{Q}_{\text{eval}}\subseteq\mathbf{Q}^{\text{all}}_{\text{eval}}$. Then, by the first claim in Lemma 10, and two triangle inequalities, for all $h\in[H]$,

$$\mathop{\mathbb{E}}_{(S,A)\sim\mu_h}\left|q^{G^\star}(S,\pi^e)-q^{G^\star}(S,\pi^e_{G^\star})\right|\le\mathop{\mathbb{E}}_{(S,A)\sim\mu_h}\left|q^{\pi^e_{G^\star}}(S,\pi^e)-q^{\pi^e_{G^\star}}(S,\pi^e_{G^\star})\right|+2\tilde{\epsilon}\,.$$

Since $\pi^e_{G^\star}$ only differs from π^e on states s where $\text{range}(s)\le\alpha$, and for any such state, $\max_{a,a'\in\mathcal{A}}q^{\pi^e_{G^\star}}(S,a)-q^{\pi^e_{G^\star}}(S,a')\le\alpha$ (by the definition of range), we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(S,A)\sim\mu_h}\left|q^{\pi^e_{G^\star}}(S,\pi^e)-q^{\pi^e_{G^\star}}(S,\pi^e_{G^\star})\right|\\
&=\mathop{\mathbb{E}}_{(S,A)\sim\mu_h}\left[\mathbb{I}\{\text{range}(S)\le\alpha\}\left|q^{\pi^e_{G^\star}}(S,\pi^e)-q^{\pi^e_{G^\star}}(S,\pi^e_{G^\star})\right|\right]\le\alpha\,.
\end{aligned}$$

Putting everything together and noting that the minimum over $q^G \in \mathbf{Q}_{\text{eval}}$ of Eq. (14) is at most the value at $q^{G^\star}$ (since on event $\bar{\mathcal{E}}$, $q^{G^\star} \in \mathbf{Q}_{\text{eval}}$), we get the desired result.

■

We state one final lemma, which makes use of the above lemmas, before proving Theorem 1.

Lemma 16 *On event $\bar{\mathcal{E}} \cap \tilde{\mathcal{E}}$, for all admissible distributions $\nu = (\nu_t)_{t\in[H]}$, and for all $h \in [H]$,*

$$\mathop{\mathbb{E}}_{(S,A)\sim\nu_h}\left|q^{\pi^e_{\hat{G}}}(S,\pi^e) - q^{\pi^e_{\hat{G}}}(S,\pi^e_{\hat{G}})\right| \le \tilde{\alpha} = \tilde{\mathcal{O}}\left(\alpha + \log\left(\frac{1}{\alpha}\right)\frac{C_0^{5/2}H^{5/2}d^{3/2}}{\sqrt{n}}\right).$$

Proof By two triangle inequalities and Lemma 10, on event $\bar{\mathcal{E}}$, for any admissible distribution $\nu = (\nu_t)_{t\in[H]}$, and for all $h \in [H]$,

$$\begin{aligned}&\mathop{\mathbb{E}}_{(S,A)\sim\nu_h}\left|q^{\pi^e_{\hat{G}}}(S,\pi^e) - q^{\pi^e_{\hat{G}}}(S,\pi^e_{\hat{G}})\right| \\ &\le \mathop{\mathbb{E}}_{(S,A)\sim\nu_h}\left|q^{\hat{G}}(S,\pi^e) - q^{\hat{G}}(S,\pi^e_{\hat{G}})\right| + 2\tilde{\epsilon}\,. \end{aligned} \tag{20}$$

We bound the expectation on the right side next. Since $q^{\hat{G}} \in \mathbf{Q}_{\text{eval}} \subset \mathbf{Q}^{\text{all}}_{\text{eval}}$, on event $\tilde{\mathcal{E}}$, for all $h \in [H]$,

$$\mathop{\mathbb{E}}_{(S,A)\sim\mu_h}\left|q^{\hat{G}}(S,\pi^e) - q^{\hat{G}}(S,\pi^e_{\hat{G}})\right| \le \frac{1}{n}\sum_{j=1}^n\left|q^{\hat{G}}(S_h^j,\pi^e) - q^{\hat{G}}(S_h^j,\pi^e_{\hat{G}})\right| + \zeta_2\,,$$

and by Lemma 15, on event $\tilde{\mathcal{E}} \cap \bar{\mathcal{E}}$,

$$\max_{h\in[H]}\frac{1}{n}\sum_{j=1}^n\left|q^{\hat{G}}(S_h^j,\pi^e) - q^{\hat{G}}(S_h^j,\pi^e_{\hat{G}})\right| \le \alpha + 2\tilde{\epsilon} + \zeta_2\,.$$

Combining the above two results, on event $\tilde{\mathcal{E}} \cap \bar{\mathcal{E}}$, for all $h \in [H]$,

$$\mathop{\mathbb{E}}_{(S,A)\sim\mu_h}\left|q^{\hat{G}}(S,\pi^e) - q^{\hat{G}}(S,\pi^e_{\hat{G}})\right| \le \alpha + 2\tilde{\epsilon} + 2\zeta_2\,.$$

Notice that since the expectation is over a non-negative function (due to the absolute value) we can apply Lemma 13 to get that for all admissible distributions $\nu = (\nu_t)_{t\in[H]}$, and for all $h \in [H]$,

$$\mathop{\mathbb{E}}_{(S,A)\sim\nu_h}\left|q^{\hat{G}}(S,\pi^e) - q^{\hat{G}}(S,\pi^e_{\hat{G}})\right| \le C_0\cdot(\alpha + 2\tilde{\epsilon} + 2\zeta_2) =: \tilde{\alpha}\,.$$

Combining the above with Eq. (20) completes the proof.

■

We are now ready to prove our main result (Theorem 1).

Proof [of Theorem 1] Recall that $\hat{G} \in \mathbf{G}_{\text{eval}}$ is such that $q^{\hat{G}} \in \mathbf{Q}_{\text{eval}}$ is the minimizer of Eq. (14). Our goal is to show that

$$\left|v^{\pi^e}(s_1) - \hat{v}\right| = \left|v^{\pi^e}(s_1) - q^{\hat{G}}(s_1,\pi^e_{\hat{G}})\right| \le \epsilon\,.$$

We relate things to $v^{\pi^e_{\hat{G}}}(\cdot) = q^{\pi^e_{\hat{G}}}(\cdot, \pi^e_{\hat{G}})$:

$$\left|v^{\pi^e}(s_1) - q^{\hat{G}}(s_1, \pi^e_{\hat{G}})\right| \le \left|v^{\pi^e}(s_1) - v^{\pi^e_{\hat{G}}}(s_1)\right| + \left|q^{\pi^e_{\hat{G}}}(s_1, \pi^e_{\hat{G}}) - q^{\hat{G}}(s_1, \pi^e_{\hat{G}})\right| .$$

By Lemma 10 (since $\mathbb{P}^h_{\pi^e_{\hat{G}}, s_1}$ is an admissible distribution), on event $\bar{\mathcal{E}}$,

$$\left|q^{\pi^e_{\hat{G}}}(s_1, \pi^e_{\hat{G}}) - q^{\hat{G}}(s_1, \pi^e_{\hat{G}})\right| = \mathop{\mathbb{E}}_{(S_1, A_1) \sim \mathbb{P}^1_{\pi^e_{\hat{G}}, s_1}} \left|\pi^e_{\hat{G}}(S_1, A_1) - q^{\hat{G}}(S_1, A_1)\right| \le \tilde{\epsilon} .$$

To bound the first term we will use Eq. (13), which we restate here for clarity:

$$\left|v^{\pi^e}(s_1) - v^{\pi^e_{\hat{G}}}(s_1)\right| = \left|\sum_{h=1}^{H} \mathop{\mathbb{E}}_{(S_h, A_h) \sim \mathbb{P}^h_{\pi^e, s_1}} \left(q^{\pi^e_{\hat{G}}}(S_h, \pi^e) - q^{\pi^e_{\hat{G}}}(S_h, \pi^e_{\hat{G}})\right)\right| .$$

Then by H triangle inequalities and Jensen's inequality we have that:

$$\left|\sum_{h=1}^{H} \mathop{\mathbb{E}}_{(S_h, A_h) \sim \mathbb{P}^h_{\pi^e, s_1}} \left(q^{\pi^e_{\hat{G}}}(S_h, \pi^e) - q^{\pi^e_{\hat{G}}}(S_h, \pi^e_{\hat{G}})\right)\right| \le \sum_{h=1}^{H} \mathop{\mathbb{E}}_{(S_h, A_h) \sim \mathbb{P}^h_{\pi^e, s_1}} \left|q^{\pi^e_{\hat{G}}}(S_h, \pi^e) - q^{\pi^e_{\hat{G}}}(S_h, \pi^e_{\hat{G}})\right| .$$

Noting that $\mathbb{P}^h_{\pi^e, s_1}$ is an admissible distribution, and applying Lemma 16, we have that on event $\bar{\mathcal{E}} \cap \tilde{\mathcal{E}}$,

$$\left|v^{\pi^e}(s_1) - v^{\pi^e_{\hat{G}}}(s_1)\right| \le H\tilde{\alpha} .$$

By combining the above two bounds,

$$\left|v^{\pi^e}(s_1) - \hat{v}\right| \le H\tilde{\alpha} + \tilde{\epsilon} = \tilde{\mathcal{O}}\left(H\alpha + \log(1/\alpha)\frac{C_0^{5/2} H^{7/2} d^{3/2}}{\sqrt{n}}\right) . \tag{21}$$

If we set

$$\alpha = \tilde{\mathcal{O}}\left(\frac{C_0^{5/2} H^{5/2} d^{3/2}}{\sqrt{n}}\right) , \tag{22}$$

then if $n = \tilde{\Theta}(\frac{C_0^5 H^7 d^3}{\epsilon^2})$, it implies that

$$\left|v^{\pi^e}(s_1) - \hat{v}\right| \le \epsilon .$$

Since the probability of event $\bar{\mathcal{E}}$ is at least $1 - \delta/2$ (Lemma 10), and the probability of event $\tilde{\mathcal{E}}$ is at least $1 - \delta/2$ (Lemma 14), the desired result follows by a union bound. ■

Appendix B. Analysis of LIN-q^π-FQI

The following analysis is basically identical to that in Tkachuk et al. [2024][Theorem 1]. However, we get an improved bound on n due to our improved bound in Lemma 9.

We begin by proving Lemma 8.

Proof [of Lemma 8] By the performance difference lemma (Lemma 29), for any $h \in [H+1]$, $s \in \mathcal{S}_h$, and $\pi \in \Pi$,

$$|v^\pi(s) - v^{\pi_{G^\star}}(s)| = |\textstyle\sum_{t=h}^H \mathbb{E}_{(S_t,A_t)\sim\mathbb{P}^t_{\pi,s}} (q^{\pi_{G^\star}}(S_t,\pi) - q^{\pi_{G^\star}}(S_t,\pi_{G^\star}))| \le H\alpha\,. \tag{23}$$

The inequality follows from the fact that π and $\pi_{G^\star}$ only differ on states s with $\mathrm{range}(s) \le \alpha$, which shows the first result of the lemma.

To show the second result we will show that for any $h \in [H]$, $s \in \mathcal{S}_h$,

$$v^{\pi^\star}(s) - v^{\bar\pi^\star_{G^\star}}(s) \le v^{\pi^\star}(s) - v^{\pi^\star_{G^\star}}(s)\,.$$

Then by applying first result we get second result.

It will be sufficient to show $v^{\bar\pi^\star_{G^\star}} \ge v^{\pi^\star_{G^\star}}$. The proof is taken verbatim from [Tkachuk et al., 2024][Lemma D.1]. We use induction. The base case of $h = H+1$ is immediately true by definition as v-values are 0 on $s_\top$, regardless of the policy. Assuming the inductive hypothesis holds for $h+1$, we continue by proving it for h. Let $(s,a) \in \mathcal{S}_h \times \mathcal{A}$ be arbitrary. Notice that

$$q^{\bar\pi^\star_{G^\star}}(s,a) - q^{\pi^\star_{G^\star}}(s,a) = \mathop{\mathbb{E}}_{S'\sim P(s,a)} v^{\bar\pi^\star_{G^\star}}(S') - v^{\pi^\star_{G^\star}}(S') \ge 0\,,$$

where the inequality is due to the inductive hypothesis. Next, for any $s \in \mathcal{S}_h$, by the above and the definition of the policies,

$$\begin{aligned}
v^{\bar\pi^\star_{G^\star}}(s) &= \omega_{G^\star}(s) q^{\bar\pi^\star_{G^\star}}(s,\pi^b) + (1-\omega_{G^\star}(s)) \max_{a\in\mathcal{A}} q^{\bar\pi^\star_{G^\star}}(s,a) \\
&\ge \omega_{G^\star}(s) q^{\pi^\star_{G^\star}}(s,\pi^b) + (1-\omega_{G^\star}(s)) \max_{a\in\mathcal{A}} q^{\pi^\star_{G^\star}}(s,a) \\
&\ge \omega_{G^\star}(s) q^{\pi^\star_{G^\star}}(s,\pi^b) + (1-\omega_{G^\star}(s)) q^{\pi^\star_{G^\star}}(s,\pi^\star) = v^{\pi^\star_{G^\star}}(s)\,,
\end{aligned}$$

finishing the induction. ■

Now, we show that for any policy π the value function of a policy π' that is greedy with respect to a good estimate of q^π cannot be much worse than π.

Lemma 17 *Let π, π' be any two policies. Let $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be such that for any admissible distribution $\nu = (\nu_t)_{t\in[H]}$, and for all $h \in [H]$,*

$$\mathop{\mathbb{E}}_{(S,A)\sim\nu_h} |q^\pi(S,A) - q(S,A)| \le \kappa\,.$$

If $\pi'(s) = \arg\max_{a\in\mathcal{A}} q(s,a)$ for all $s \in \mathcal{S}$, then for any admissible distribution $\nu = (\nu_t)_{t\in[H]}$, and for all $h \in [H+1]$,

$$\mathop{\mathbb{E}}_{(S_h,A)\sim\nu_h} [v^\pi(S_h) - v^{\pi'}(S_h)] \le 2(H-h+1)\kappa\,.$$

Proof We use induction. The base case of $h = H + 1$ is immediately true by definition as v-values are 0 on $s_\top$, regardless of the policy. Assuming the inductive hypothesis holds for $h+1$, we continue by proving it for h. Adding and subtracting q,

$$\mathop{\mathbb{E}}_{(S_h,A)\sim\nu_h}[v^{\pi}(S_h) - v^{\pi'}(S_h)] = \mathop{\mathbb{E}}_{(S_h,A)\sim\nu_h}[v^{\pi}(S_h) - q(S_h,\pi)] + \mathop{\mathbb{E}}_{(S_h,A)\sim\nu_h}[q(S_h,\pi) - v^{\pi'}(S_h)]\,.$$

The first term is bounded by κ by the assumption of the lemma. To bound the second term, notice that by the definition of π', $q(S_h,\pi) \le q(S_h,\pi')$. Thus, by making use of the assumption of the lemma and the inductive hypothesis,

$$\begin{aligned}
\mathop{\mathbb{E}}_{(S_h,A)\sim\nu_h}[q(S_h,\pi) - v^{\pi'}(S_h)] &\le \mathop{\mathbb{E}}_{(S_h,A)\sim\nu_h}[q(S_h,\pi') - q^{\pi'}(S_h,\pi')] \\
&\le \mathop{\mathbb{E}}_{(S_h,A)\sim\nu_h}[q^{\pi}(S_h,\pi') - q^{\pi'}(S_h,\pi')] + \kappa \\
&\le \mathop{\mathbb{E}}_{(S_h,A)\sim\nu_h}\left[\mathop{\mathbb{E}}_{S_{h+1}\sim P(\cdot|S_h,A)}[v^{\pi}(S_{h+1}) - v^{\pi'}(S_{h+1})]\right] + \kappa \\
&\le 2(H-h)\kappa + \kappa\,.
\end{aligned}$$

Putting everything together we get the desired result. ■

We are now ready to prove Theorem 2 .

Proof [of Theorem 2] We need to show that

$$v^{\pi^\star}(s_1) - v^{\hat\pi}(s_1) \le \epsilon\,.$$

By Lemma 8,

$$v^{\pi^\star}(s_1) - v^{\hat\pi}(s_1) \le v^{\bar\pi^\star_{G^\star}}(s_1) - v^{\hat\pi}(s_1) + H\alpha\,.$$

Recall that $\hat\pi(s) = \arg\max_a \hat q(s,a)$, where $\hat q = \arg\max_{q\in\mathbf{Q}_{\text{opt}}} \max_a q(s_1,a)$. Adding and subtracting $\hat q$,

$$v^{\bar\pi^\star_{G^\star}}(s_1) - v^{\hat\pi}(s_1) \le v^{\bar\pi^\star_{G^\star}}(s_1) - \hat q(s_1,\hat\pi) + \hat q(s_1,\hat\pi) - v^{\hat\pi}(s_1)\,. \tag{24}$$

For the remainder of the proof, assume we are on the event (which occurs with probability at least $1-\delta$) for which the two claims of Lemma 9 hold.

By the second claim of Lemma 9, $q^{G^\star} \in \mathbf{Q}_{\text{opt}}$. Thus, by the first claim of Lemma 9,

$$v^{\bar\pi^\star_{G^\star}}(s_1) = q^{\bar\pi^\star_{G^\star}}(s_1,\bar\pi^\star_{G^\star}) \le q^{G^\star}(s_1,\bar\pi^\star_{G^\star}) + \tilde\epsilon\,.$$

Since $\hat q$ is the maximizer of $\max_a q(s_1,a)$ over $q\in\mathbf{Q}_{\text{opt}}$, and $\max_a \hat q(s_1,a) = \hat q(s_1,\hat\pi)$ (by definition of $\hat\pi$),

$$q^{G^\star}(s_1,\bar\pi^\star_{G^\star}) \le \hat q(s_1,\hat\pi)\,.$$

Putting the above two inequalities together we get that following bound on the first term in Eq. (24):

$$v^{\bar\pi^\star_{G^\star}}(s_1) - \hat q(s_1,\hat\pi) \le \tilde\epsilon\,.$$

Next we bound the second term in Eq. (24). Let $\bar{G} \in \mathbf{G}_{\text{opt}}$ be the modification for which $\hat{q}$ is based on. By the first claim of Lemma 9,

$$\hat{q}(s_1, \hat{\pi}) \le q^{\bar{\pi}^\star_{\bar{G}}}(s_1, \hat{\pi}) + \tilde{\epsilon}\,.$$

Also,

$$q^{\bar{\pi}^\star_{\bar{G}}}(s_1, \hat{\pi}) - v^{\hat{\pi}}(s_1) = \mathop{\mathbb{E}}_{(S_2,A_2)\sim\mathbb{P}^2_{\hat{\pi},s_1}} \left[v^{\bar{\pi}^\star_{\bar{G}}}(S_2) - v^{\hat{\pi}}(S_2)\right].$$

Since (by the first claim of Lemma 9), for all admissible distributions $\nu = (\nu_t)_{t\in[H]}$, and for all $h \in [H]$,

$$\mathop{\mathbb{E}}_{(S,A)\sim\nu_h} \left|q^{\bar{G}}(S, A) - q^{\bar{\pi}^\star_{\bar{G}}}(S, A)\right| \le \tilde{\epsilon}\,,$$

we can apply Lemma 17 to get that

$$\mathop{\mathbb{E}}_{(S_2,A_2)\sim\mathbb{P}^2_{\hat{\pi},s_1}} \left[v^{\bar{\pi}^\star_{\bar{G}}}(S_2) - v^{\hat{\pi}}(S_2)\right] \le 2(H-1)\tilde{\epsilon}\,.$$

Putting the above three results together we get the following bound on the second term in Eq. (24):

$$\hat{q}(s_1, \hat{\pi}) - v^{\hat{\pi}}(s_1) \le 2(H-1)\tilde{\epsilon} + \tilde{\epsilon}\,.$$

Putting everything together we get that

$$v^{\bar{\pi}^\star_{\bar{G}^\star}}(s_1) - v^{\hat{\pi}}(s_1) \le H\alpha + 2H\tilde{\epsilon} = \tilde{\mathcal{O}}\left(H\alpha + \log(1/\alpha)\frac{C_0^{3/2}H^{7/2}d^{3/2}}{\sqrt{n}}\right). \tag{25}$$

If we set

$$\alpha := \tilde{\mathcal{O}}\left(\frac{C_0^{3/2}H^{5/2}d^{3/2}}{\sqrt{n}}\right), \tag{26}$$

then if $n = \tilde{\Theta}(\frac{C_0^3H^7d^3}{\epsilon^2})$, it implies that

$$v^{\pi^\star}(s_1) - v^{\hat{\pi}}(s_1) \le \epsilon\,.$$

■

Appendix C. Missing Proofs and Definitions from Section 4

Definition 18 (Definition F.1 in [Weisz et al., 2023]) *A finite set $G \subset \mathbb{R}^d$ is the basis of a near-optimal design for a set $\Theta \subseteq \mathbb{R}^d$, if there exists a probability distribution ρ over elements of G, such that for any $\theta \in \Theta$,*

$$\langle v,\, \theta\rangle = 0 \quad \textit{for all } v \in \mathrm{Ker}(V(G,\rho)),\ \textit{and} \tag{27}$$

$$\|\theta\|^2_{V(G,\rho)^\dagger} \le 2d, \tag{28}$$

$$\textit{where } V(G,\rho) = \sum_{x\in G} \rho(x)xx^\top\,, \tag{29}$$

where for a matrix X, $X^\dagger$ denotes the Moore–Penrose pseudoinverse of X, and $\mathrm{Ker}(X)$ denotes its kernel (null space).

C.1. Proof of Lemma 5

Proof The following lemma from [Tkachuk et al., 2024] will be useful (proof: Appendix C.1)[18].

Lemma 19 (Lemma 4.2 in [Tkachuk et al., 2024]) *For any $f : \mathcal{S} \to [0, H]$ with $f(s_\top) = 0$, policy $\pi \in \Pi$, and stage $h \in [H]$, there exists a parameter $\rho_h^\pi(f) \in \mathcal{B}(\tilde{L}_\theta)$ such that for all $(s, a) \in \mathcal{S}_h \times \mathcal{A}$,*

$$\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}} \sum_{\tau=h+1}^{H+1} \left(R_{h:\tau-1} + f_\tau(S_\tau)\right) F_{G^\star,\text{Traj},h+1}(\tau) = \langle \phi(s,a),\, \rho_h^\pi(f) \rangle \,.$$

Set the policy in the above lemma to the behavior policy π^b. Notice that for any $\rho_h^{\pi^b}(f) \in \mathcal{B}(\tilde{L}_\theta)$ in the lemma $\langle \phi(\cdot,\cdot),\, \rho_h^{\pi^b}(f) \rangle \in \widetilde{\mathcal{F}}_h$. Also, notice that for any $h \in [H]$, $q_{h\to} : [0, H]^{\mathcal{SA}_{h\to}}$ with $q_{H+1}(s_\top, \cdot) = 0$,

$$\begin{aligned} T_{G^\star}^\pi\, q_{h\to}(s,a) &= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} g_{G^\star}^\pi(q_{h\to}, \text{Traj}) \\ &= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} \sum_{\tau=h+1}^{H+1} \left(R_{h:\tau-1} + q_\tau(S_\tau, \pi)\right) F_{G^\star,\text{Traj},h+1}(\tau) \,, \end{aligned}$$

which matches the left-hand side of the equality in the above lemma with $f_t(\cdot) = q_t(\cdot, \pi)$ for all $t \in [h : H+1]$. Thus, by the above lemma,

$$T_{G^\star}^\pi\, q_{h\to} \in \widetilde{\mathcal{F}}_h \,.$$

Similarly, if we let $f_t(\cdot) = \max_a q_t(\cdot, a)$ for all $t \in [h : H+1]$, then by the above lemma,

$$T_{G^\star}\, q_{h\to} \in \widetilde{\mathcal{F}}_h \,.$$

This completes the proof of Lemma 5. ■

We provide the proof of Lemma 19 for completeness, and since we found some minor errors in the proof of the original lemma in [Tkachuk et al., 2024], after correcting them, the size of $\tilde{L}_\theta$ increases by a factor of $\sqrt{2d}$.

Proof [of Lemma 19] We first mention a useful definition and lemma from Weisz et al. [2023].

Definition 20 ($\bar{\alpha}$-admissible function (Definition 4.6 in Weisz et al. [2023])) *For any $h \in [H]$, $f : \mathcal{S}_h \to \mathbb{R}$ is $\bar{\alpha}$-admissible for some $\bar{\alpha} \geq 0$ if for all $s \in \mathcal{S}_h$, $\bar{\alpha}|f(s)| \leq \text{range}_G(s)$.*

Lemma 21 (Realizability of admissible functions (Lemma 4.7 in Weisz et al. [2023])) *For any $t \in [2 : H]$, $h \in [t-1]$, $\bar{\alpha} \geq 0$, if $f : \mathcal{S}_t \to \mathbb{R}$ is $\bar{\alpha}$-admissible, then for any $\pi \in \Pi$, there exists a $\tilde{\theta} \in \mathcal{B}(4d_0 L_\theta/\bar{\alpha})$ such that for all $(s, a) \in \mathcal{S}_h \times \mathcal{A}$,*

$$\mathbb{E}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}} f(S_t) = \langle \phi(s,a),\, \tilde{\theta} \rangle \,.$$

18. The actual lemma in [Tkachuk et al., 2024] is stated for the case with misspecification error; however, since in our case the misspecification error is 0, we state the lemma for that case only.

Now we prove the lemma. Fix any $f : \mathcal{S} \to [0, H]$ with $f(s_\top) = 0$, $\pi \in \Pi$. For $h \in [2 : H+1]$, define $\tilde{g}_h : \mathcal{S}_h \to [-H, H]$ as $\tilde{g}_{H+1}(\cdot) = 0$, and for all $h \in [2 : H]$, $s \in \mathcal{S}_h$,

$$\tilde{g}_h(s) \coloneqq \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}\left[\sum_{\tau=h}^{H+1}(-R_{\tau:H} + f_\tau(S_\tau))\, F_{G^\star,\text{Traj},h}(\tau)\right].$$

Notice that for any $h \in [H]$, $s \in \mathcal{S}_h$, and $a \in \mathcal{A}$,

$$\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}}\left[\sum_{\tau=h+1}^{H+1}(R_{h:\tau-1} + f_\tau(S_\tau))\, F_{G^\star,\text{Traj},h+1}(\tau)\right] = \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}}[\tilde{g}_{h+1}(S_{h+1}) + R_{h:H}] \tag{30}$$

$$= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}}[\tilde{g}_{h+1}(S_{h+1})] + q^\pi_h(s,a)\,. \tag{31}$$

The second term of the sum, $q^\pi_h(\cdot,\cdot)$ is equal to $\langle\phi(\cdot,\cdot),\, \theta^*_h(\pi)\rangle$ by Assumption 1. The first term needs more work before Lemma 21 can be applied. First we can write g_h in a different form. For all $h \in [2 : H]$, $s \in \mathcal{S}_h$,

$$\begin{aligned}
\tilde{g}_h(s) &= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}\left[\sum_{\tau=h}^{H+1}(-R_{\tau:H} + f_\tau(S_\tau))\, F_{G^\star,\text{Traj},h}(\tau)\right] \\
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}\left[\sum_{\tau=h}^{H+1}(-R_{\tau:H} + f_\tau(S_\tau))(1-\omega_{G^\star}(S_\tau))\prod_{u=h}^{\tau-1}\omega_{G^\star}(S_u)\right] \\
&= (1-\omega_{G^\star}(s))\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}[-R_{h:H} + f_h(s)] \\
&\quad + \omega_{G^\star}(s)\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}\left[\sum_{\tau=h+1}^{H+1}(-R_{\tau:H} + f_\tau(S_\tau))(1-\omega_{G^\star}(S_\tau))\prod_{u=h+1}^{\tau-1}\omega_{G^\star}(S_u)\right] \\
&= (1-\omega_{G^\star}(s))\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}[-R_{h:H} + f_h(s)] + \omega_{G^\star}(s)\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}[\tilde{g}_{h+1}(S_{h+1})]\,.
\end{aligned}$$

For $h \in [2 : H]$ define $g_h : \mathcal{S}_h \to \mathbb{R}$ as

$$g_h(s) = (1-\omega_{G^\star}(s))\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}[-R_{h:H} + f_h(s) - \tilde{g}_{h+1}(S_{h+1})]\,.$$

Then

$$\tilde{g}_h(s) = g_h(s) + \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}[\tilde{g}_{h+1}(S_{h+1})]\,.$$

Thus $\tilde{g}_h$ can be decomposed into a sum of g_t functions as for all $h \in [2 : H]$, $s \in \mathcal{S}_h$,

$$\tilde{g}_h(s) = \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s}}\left[\sum_{t=h}^{H} g_t(S_t)\right]. \tag{32}$$

The benefit of decomposing $\tilde{g}_h$ into g_t functions is that g_t are $\alpha/(2H\sqrt{2d})$-admissible (recall Definition 20). To see this, note that for any trajectory, $t \in [2 : H]$, $s \in \mathcal{S}_t$,

$$-R_{t:H} + f_t(s) - \tilde{g}_{t+1}(S_{t+1}) \in [-2H, 2H]\,.$$

Then, $g_t(s)$ takes the expected value (which is still in $[-2H, 2H]$) of the above display and multiplies it by $1 - \omega_{G^\star}(s)$ (which by Eq. (6) is in $[0, 1]$). Notice that

$$1 - \omega_{G^\star}(s) \le \frac{\sqrt{2d} \cdot \text{range}_{G^\star}(s)}{\alpha} .$$

This holds since: when $\text{range}_{G^\star}(s) \ge \alpha/\sqrt{2d}$ the right hand side is greater than 1, when $\text{range}_{G^\star}(s) \le \alpha/(2\sqrt{2d})$, $\omega_{G^\star}(s) = 1$, and when $\alpha/(2\sqrt{2d}) \le \text{range}_{G^\star}(s) < \alpha/\sqrt{2d}$, $1 - \omega_{G^\star}(s) = 2\sqrt{2d} \cdot \text{range}_{G^\star}(s)/\alpha - 1 \le \sqrt{2d} \cdot \text{range}_{G^\star}(s)/\alpha$. Therefore g_t is $\alpha/(2H\sqrt{2d})$-admissible. By using Lemma 21 we get that for any $h \in [H-1]$, $t \in [h+1 : H]$, there exists a $\tilde{\theta}_t \in \mathcal{B}(8Hd_0\sqrt{2d}L_\theta/\alpha)$ such that for all $(s, a) \in \mathcal{S}_h \times \mathcal{A}$,

$$\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}} g_t(S_t) = \langle \phi(s, a),\, \tilde{\theta}_t \rangle .$$

By combining this with Eq. (32), for all $h \in [H]$, the parameter $\tilde{\theta} = \sum_{t=h+1}^{H} \tilde{\theta}_t$ is in $\mathcal{B}(8H^2 d_0\sqrt{2d}L_\theta/\alpha)$, and for all $(s, a) \in \mathcal{S}_h \times \mathcal{A}$,

$$\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}} \tilde{g}_{h+1}(S_{h+1}) = \langle \phi(s, a),\, \tilde{\theta} \rangle .$$

Combining the above with Eq. (31), we have that for all $h \in [H]$, the parameter $\theta = \tilde{\theta} + \theta^*_h(\pi)$ is in $\mathcal{B}(8H^2 d_0\sqrt{2d}L_\theta/\alpha + L_\theta)$, and for all $(s, a) \in \mathcal{S}_h \times \mathcal{A}$,

$$\begin{aligned}\mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}} \left[\sum_{\tau=h+1}^{H+1} (R_{h:\tau-1} + f_\tau(S_\tau))\, F_{G^\star,\text{Traj},h+1}(\tau) \right] &= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi,s,a}} [\tilde{g}_{h+1}(S_{h+1})] + q^\pi_h(s, a) \\ &= \langle \phi(s, a),\, \theta \rangle .\end{aligned}$$

To finish the proof we define $\rho^\pi_h(f) = \theta$.

■

C.2. Proof of Lemma 6

Proof We need to show that $q^{\pi_G}_h = T^\pi_G\, q^{\pi_G}_{h\to}$, for any policy $\pi \in \Pi$ and modification $G \in \mathbf{G}$. We make use of the fact that $q^{\pi_G}_h = T^\pi q^{\pi_G}_{h+1}$, and the definition of π_G, to expand $q^{\pi_G}_h$ as follows, for any $(s, a) \in \mathcal{S}_h \times \mathcal{A}$, $h \in [H]$:

$$\begin{aligned} q^{\pi_G}_h(s, a) &= T^{\pi_G} q^{\pi_G}_{h+1} \\ &= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi_G,s,a}} [R_h + q^{\pi_G}_{h+1}(S_{h+1}, \pi_G)] \\ &= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi_G,s,a}} [R_h + (1 - \omega_G(S_{h+1}))q^{\pi_G}_{h+1}(S_{h+1}, \pi) + \omega_G(S_{h+1})q^{\pi_G}_{h+1}(S_{h+1}, \pi^b)] . \end{aligned}$$

Notice that if we change the expectation to be with respect to $\mathbb{P}_{\pi^b,s,a}$ instead of $\mathbb{P}_{\pi_G,s,a}$, then the value does not change. Thus, changing to $\mathbb{P}_{\pi^b,s,a}$ and expanding $q_\cdot^{\pi_G}(S_\cdot,\pi^b)$ repeatedly, we get that:

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} [R_h + (1-\omega_G(S_{h+1}))q_{h+1}^{\pi_G}(S_{h+1},\pi) \\
&\quad + \omega_G(S_{h+1})(R_{h+1} + (1-\omega_G(S_{h+2}))q_{h+2}^{\pi_G}(S_{h+2},\pi) + \omega_G(S_{h+2})q_{h+2}^{\pi_G}(S_{h+2},\pi^b))] \\
&\vdots \\
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} \left[\sum_{\tau=h+1}^{H+1} \prod_{u=h+1}^{\tau-1} \omega_G(S_u)(R_{\tau-1} + (1-\omega_G(S_\tau))q_\tau^{\pi_G}(S_\tau,\pi)) \right].
\end{aligned}
$$

Since $\sum_{\tau=h+1}^{H+1}\prod_{u=h+1}^{\tau-1}\omega_G(S_u)R_{\tau-1} = \sum_{\tau=h+1}^{H+1}(1-\omega_G(S_\tau))\prod_{u=h+1}^{\tau-1}\omega_G(S_u)R_{h:\tau-1}$, we have that:

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} \left[\sum_{\tau=h+1}^{H+1} (1-\omega_G(S_\tau)) \prod_{u=h+1}^{\tau-1} \omega_G(S_u)(R_{h:\tau-1} + q_\tau^{\pi_G}(S_\tau,\pi)) \right] \\
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} \sum_{\tau=h+1}^{H+1} F_{G,\text{Traj},h+1}(\tau)(R_{h:\tau-1} + q_\tau^{\pi_G}(S_\tau,\pi)) \\
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} g_G^\pi(q_{h\to}^{\pi_G},\text{Traj}) = T_G^\pi\, q_{h\to}^{\pi_G}(s,a)\,.
\end{aligned}
$$

Which completes the proof. ■

C.3. Proof of Lemma 7

Proof We need to show that $q_h^{\bar\pi_G^\star} = T_G\, q_{h\to}^{\bar\pi_G^\star}$, for any modification $G \in \mathbf{G}$. We make use of the fact that $q_h^{\bar\pi_G^\star} = T^{\bar\pi_G^\star} q_{h+1}^{\bar\pi_G^\star}$, and the definition of $\bar\pi_G^\star$, to expand $q_h^{\bar\pi_G^\star}$ as follows, for any $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, $h \in [H]$:

$$
\begin{aligned}
q_h^{\bar\pi_G^\star}(s,a) &= T^{\bar\pi_G^\star} q_{h+1}^{\bar\pi_G^\star} \\
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\bar\pi_G^\star,s,a}} [R_h + q_{h+1}^{\bar\pi_G^\star}(S_{h+1},\pi_G)] \\
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\bar\pi_G^\star,s,a}} [R_h + (1-\omega_G(S_{h+1}))\max_a q_{h+1}^{\bar\pi_G^\star}(S_{h+1},a) + \omega_G(S_{h+1})q_{h+1}^{\bar\pi_G^\star}(S_{h+1},\pi^b)]\,.
\end{aligned}
$$

Notice that if we change the expectation to be with respect to $\mathbb{P}_{\pi^b,s,a}$ instead of $\mathbb{P}_{\bar\pi_G^\star,s,a}$, then the value does not change. Thus, changing to $\mathbb{P}_{\pi^b,s,a}$ and expanding $q_\cdot^{\bar\pi_G^\star}(S_\cdot,\pi^b)$ repeatedly, we get that:

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} [R_h + (1-\omega_G(S_{h+1}))\max_a q_{h+1}^{\bar\pi_G^\star}(S_{h+1},a) \\
&\quad + \omega_G(S_{h+1})(R_{h+1} + (1-\omega_G(S_{h+2}))\max_a q_{h+2}^{\bar\pi_G^\star}(S_{h+2},a) + \omega_G(S_{h+2})q_{h+2}^{\bar\pi_G^\star}(S_{h+2},\pi^b))] \\
&\vdots \\
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} \left[\sum_{\tau=h+1}^{H+1} \prod_{u=h+1}^{\tau-1} \omega_G(S_u)(R_{\tau-1} + (1-\omega_G(S_\tau))\max_a q_\tau^{\bar\pi_G^\star}(S_\tau,a)) \right].
\end{aligned}
$$

Since $\sum_{\tau=h+1}^{H+1}\prod_{u=h+1}^{\tau-1}\omega_G(S_u)R_{\tau-1} = \sum_{\tau=h+1}^{H+1}(1-\omega_G(S_\tau))\prod_{u=h+1}^{\tau-1}\omega_G(S_u)R_{h:\tau-1}$, we have that:

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}}\left[\sum_{\tau=h+1}^{H+1}(1-\omega_G(S_\tau))\prod_{u=h+1}^{\tau-1}\omega_G(S_u)(R_{h:\tau-1}+\max_a q_\tau^{\bar\pi^\star_G}(S_\tau,a))\right]\\
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}}\sum_{\tau=h+1}^{H+1}F_{G,\text{Traj},h+1}(\tau)(R_{h:\tau-1}+\max_a q_\tau^{\bar\pi^\star_G}(S_\tau,a))\\
&= \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\pi^b,s,a}} g_G(q_{h\to}^{\bar\pi^\star_G},\text{Traj}) = T_G\, q_{h\to}^{\bar\pi^\star_G}(s,a)\,.
\end{aligned}
$$

Which completes the proof. ■

C.4. Proof of Lemma 9

We split the proof into two parts, showing each claim separately.
Proof

PROOF OF THE FIRST CLAIM:

To show the first claim we make use of the following lemma from Tkachuk et al. [2024].

Lemma 22 (Lemma 5.1 in [Tkachuk et al., 2024]) *For all $\delta\in(0,1)$, with probability at least $1-\delta/5$, for all $G\in\mathbf{G}_{opt}$, for all $h\in[H]$, for all $(\theta_{s,a})_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}$, $(\check\theta_{s,a})_{(s,a)\in\mathcal{S}_h\times\mathcal{A}}\in(\Theta^{opt}_{G,h})^{\mathcal{S}_h\times\mathcal{A}}$, and for all admissible distributions $\nu=(\nu_t)_{t\in[H]}$, it holds that*

$$
\mathop{\mathbb{E}}_{(S,A)\sim\nu_h}\left[\bar q_{\theta_{S,A}}(S,A)-\bar q_{\check\theta_{S,A}}(S,A)\right]\le\tilde\epsilon:=C_0\cdot(\bar\epsilon+\zeta_1)=\tilde{\mathcal{O}}\left(\log\left(\frac{1}{\alpha}\right)\frac{C_0^{3/2}H^{5/2}d^{3/2}}{\sqrt{n}}\right),
\tag{33}
$$

where ζ_1 is defined in Eq. (34).

Which gives the following corollary, which has an absolute value inside the expectation.

Corollary 23 *For all $\delta\in(0,1)$, with probability at least $1-\delta/5$, for all $G\in\mathbf{G}_{opt}$, for all $h\in[H]$, for all θ and $\check\theta\in\Theta^{opt}_{G,h}$, and for all admissible distributions $\nu=(\nu_t)_{t\in[H]}$, it holds that*

$$
\mathop{\mathbb{E}}_{(S,A)\sim\nu_h}\left|\bar q_\theta(S,A)-\bar q_{\check\theta}(S,A)\right|\le\tilde\epsilon\,.
$$

Proof Fix a stage $h\in[H]$, modification $G\in\mathbf{G}_{\text{opt}}$, parameters $\theta,\check\theta\in\Theta^{\text{opt}}_{G,h}$ and admissible distribution ν_h. Define

$$
\theta_{s,a}=\begin{cases}\theta & \text{if } \bar q_\theta(s,a)-\bar q_{\check\theta}(s,a)\ge 0;\\ \check\theta & \text{if } \bar q_{\check\theta}(s,a)-\bar q_\theta(s,a)>0;\end{cases}\qquad
\check\theta_{s,a}=\begin{cases}\check\theta & \text{if } \bar q_\theta(s,a)-\bar q_{\check\theta}(s,a)\ge 0;\\ \theta & \text{if } \bar q_{\check\theta}(s,a)-\bar q_\theta(s,a)>0\,.\end{cases}
$$

Clearly, this implies that

$$
\bar q_{\theta_{s,a}}(s,a)-\bar q_{\check\theta_{s,a}}(s,a)=\left|\bar q_\theta(s,a)-\bar q_{\check\theta}(s,a)\right|\qquad\forall(s,a)\in\mathcal{S}_h\times\mathcal{A}\,,
$$

which gives the desired result since by Lemma 22, $\mathbb{E}_{(S,A)\sim\nu_h}[\bar{q}_{\theta_{S,A}}(S,A) - \bar{q}_{\tilde{\theta}_{S,A}}(S,A)] \le \tilde{\epsilon}$. ■

Combining Lemma 12 and Corollary 23 with a union bound gives that with probability at least $1-2\delta/5$, for all $G \in \mathbf{G}_{\text{opt}}$, $h \in [H]$ and $\theta \in \Theta^{\text{opt}}_{G,h}$:

$$\mathop{\mathbb{E}}_{(S,A)\sim\nu_h}\left|\bar{q}_\theta(S,A) - \bar{q}_{\theta^\star_h(\bar{\pi}^\star_G)}(S,A)\right| \le \tilde{\epsilon}\,.$$

The above equation proves the first claim of the lemma, since for $q^G \in \mathbf{Q}_{\text{opt}}$, there exists a $\theta \in \Theta^{\text{opt}}_{G,h}$ such that $q^G_h = \bar{q}_\theta$, and $\bar{q}_{\theta^\star_h(\pi^e_G)} = q^{\bar{\pi}^\star_G}_h$ by linear q^π-realizability of $\bar{\pi}^\star_G$.

PROOF OF THE SECOND CLAIM:

It will suffice to show that $G^\star \in \mathbf{G}_{\text{opt}}$, since each $q^G \in \mathbf{Q}_{\text{opt}}$ is defined based on some $G \in \mathbf{G}_{\text{opt}}$. Lemma D.3 in Tkachuk et al. [2024] gave a weaker bound than ours, in the sense that their $\bar{\epsilon}$ in the definition of $\mathbf{G}_{\text{opt}}$ was larger by a factor of $\sqrt{C_0 d}$. Next, we prove the claim and show how we improved upon their result.

The following lemma shows that the data dependent condition in the definition of $\mathbf{G}_{\text{opt}}$ concentrates around its expectation.

Lemma 24 (Lemma I.1 in [Tkachuk et al., 2024]) *For all $\delta \in (0,1)$, with probability at least $1-\delta/5$, for all $G \in \mathbf{G}$, and for all $h \in [H]$,*

$$\left|\mathop{\mathbb{E}}_{(S,A)\sim\mu_h}\left[\max_{\theta\in\Theta_{G,h}}\bar{q}_\theta(S,A) - \min_{\theta\in\Theta_{G,h}}\bar{q}_\theta(S,A)\right] - \frac{1}{n}\sum_{j\in[n]}\left(\max_{\theta\in\Theta_{G,h}}\bar{q}_\theta(S^j_h,A^j_h) - \min_{\theta\in\Theta_{G,h}}\bar{q}_\theta(S^j_h,A^j_h)\right)\right|$$
$$\le \zeta_1 := \frac{2H}{\sqrt{n}}\sqrt{dH^2 d_0 \log\left(1 + 96\sqrt{n}\sqrt{2d}H^2 L_\phi L_\theta \alpha^{-1}\sqrt{n}L_\phi\tilde{L}_\theta/(H^{3/2}d)\right) + \log\left(\frac{20H}{\delta}\right)}\,. \tag{34}$$

Thus, it will sufficient to bound $\mathbb{E}_{(S,A)\sim\mu_h}[\max_{\theta\in\Theta_{G^\star,h}}\bar{q}_\theta(S,A) - \min_{\theta\in\Theta_{G^\star,h}}\bar{q}_\theta(S,A)]$. First we state a helpful lemma.

Lemma 25 (Lemma G.1 in [Tkachuk et al., 2024]) *Let*

$$\sqrt{\lambda} := H^{3/2}d/\tilde{L}_\theta\,, \tag{35}$$
$$\beta := \sqrt{\lambda}\tilde{L}_\theta + \bar{\beta} \tag{36}$$
$$\bar{\beta} := 2H\sqrt{2dH(d_0+1)\log\left(1 + 28\sqrt{2d}H^2 L_\theta\tilde{L}_\theta L_\phi\alpha^{-1}\right) + d\log(\lambda + nL^2_\phi/d) - d\log(\lambda) + \log\left(\frac{10H}{\delta}\right)}\,.$$

For any $\delta \in (0,1)$, with probability at least $1-\delta/5$ for all $h \in [H]$, $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, $\theta_h \in \Theta_{G^\star,h}$, it holds that

$$|\bar{q}_{\theta_h}(s,a) - q^{\bar{\pi}^\star_{G^\star}}(s,a)| \le 2\beta \mathop{\mathbb{E}}_{\text{Traj}\sim\mathbb{P}_{\bar{\pi},s,a}}\sum_{t=h}^{H}\min\{1, \|\phi(S_t,A_t)\|_{\hat{X}_t^{-1}}\}\,, \tag{37}$$

where, for all $(s', a') \in \mathcal{S} \times \mathcal{A}$,

$$\bar{\pi}(a'|s') = \pi^b(a'|s')\omega_{G^\star}(s') + \mathbb{I}\left\{\arg\max_{a'' \in \mathcal{A}} g^{\bar{\pi}}(s', a'') = a'\right\}(1 - \omega_{G^\star}(s'))\,, \tag{38}$$

and $g^{\bar{\pi}}$ *is a state-action value function of policy* $\bar{\pi}$ *(similar to* $q^{\bar{\pi}}$*), except in the alternative MDP that has the same state and action spaces, and transition distributions as the original MDP under consideration, but with a reward function modified as follows. For all* $(s', a') \in \mathcal{S} \times \mathcal{A}$*, the reward in this alternative MDP is deterministically* $\min\{1, \|\phi(s', a')\|_{\hat{X}_h^{-1}}\}$*. In particular for any* $h' \in [H], (s', a') \in \mathcal{S}_{h'} \times \mathcal{A}$*,*

$$g^{\bar{\pi}}(s', a') = \mathop{\mathbb{E}}_{\text{Traj} \sim \mathbb{P}_{\bar{\pi}, s', a'}} \sum_{t=h'}^{H} \min\{1, \|\phi(S_t, A_t)\|_{\hat{X}_t^{-1}}\}\,. \tag{39}$$

Let $\bar{\pi}$ be as defined in Lemma 25. Define a new policy $\check{\pi}_h$ as

$$\check{\pi}_h(a|s) = \begin{cases} \pi^b(a|s), & \text{if stage}(s) \le h; \\ \bar{\pi}(a|s), & \text{if stage}(s) > h; \end{cases} \quad \text{for all } (s, a) \in \mathcal{S} \times \mathcal{A}\,,$$

where $\text{stage}(s)$ gives the stage index of state s. Then, by Lemma 25, with probability at least $1 - \delta/5$, for all $h \in [H]$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(S_h, A_h) \sim \mu_h}\left[\max_{\theta \in \Theta_{G^\star, h}} \bar{q}_\theta(S_h, A_h) - \min_{\theta \in \Theta_{G^\star, h}} \bar{q}_\theta(S_h, A_h)\right] \\
&= \mathop{\mathbb{E}}_{(S_h, A_h) \sim \mu_h}\left[\max_{\theta \in \Theta_{G^\star, h}} \bar{q}_\theta(S_h, A_h) - v^{\bar{\pi}^\star_{G^\star}}(S_h) + v^{\bar{\pi}^\star_{G^\star}}(S_h) - \min_{\theta \in \Theta_{G^\star, h}} \bar{q}_\theta(S_h, A_h)\right] \\
&\le 4\beta \mathop{\mathbb{E}}_{\text{Traj} \sim \mathbb{P}_{\check{\pi}, s_1}} \sum_{t=h}^{H} \min\{1, \|\phi(S_t, A_t)\|_{\hat{X}_t^{-1}}\}\,.
\end{aligned}$$

So far the analysis has followed the same steps as in Tkachuk et al. [2024]. However, we will now deviate from their analysis to get a tighter bound. To do so we first use the linearity of expectation and Jensen's inequality to get that

$$4\beta \mathop{\mathbb{E}}_{\text{Traj} \sim \mathbb{P}_{\check{\pi}, s_1}} \sum_{t=h}^{H} \min\{1, \|\phi(S_t, A_t)\|_{\hat{X}_t^{-1}}\} \le 4\beta \sum_{t=h}^{H} \sqrt{\mathop{\mathbb{E}}_{\text{Traj} \sim \mathbb{P}_{\check{\pi}, s_1}} \min\{1, \|\phi(S_t, A_t)\|^2_{\hat{X}_t^{-1}}\}}\,.$$

Notice that for any $t \in [h : H]$, $\mathbb{E}_{\text{Traj} \sim \mathbb{P}_{\check{\pi}, s_1}} \min\{1, \|\phi(S_t, A_t)\|^2_{\hat{X}_t^{-1}}\}$ is an admissible distribution. By Lemma 13,

$$4\beta \sum_{t=h}^{H} \sqrt{\mathop{\mathbb{E}}_{\text{Traj} \sim \mathbb{P}_{\check{\pi}, s_1}} \min\{1, \|\phi(S_t, A_t)\|^2_{\hat{X}_t^{-1}}\}} \le 4\sqrt{C_0}\beta \sum_{t=h}^{H} \sqrt{\mathop{\mathbb{E}}_{(S_t, A_t) \sim \mu_t} \min\{1, \|\phi(S_t, A_t)\|^2_{\hat{X}_t^{-1}}\}}\,.$$

Since we applied Jensen's inequality first, and then used Lemma 13, the concentrability coefficient C_0 appears under the square root, which gives us an improvement over Tkachuk et al. [2024] by a factor of $\sqrt{C_0}$.

The following lemma shows that the terms under the square root concentrate at a rate of $\tilde{\mathcal{O}}(d/n)$, which improves upon the $\tilde{\mathcal{O}}(d^2/n)$ rate in Tkachuk et al. [2024]. The improvement comes from an elliptical potential and exponential supermartingale concentration argument.

Lemma 26 *For any $\delta \in (0,1)$, with probability at least $1-\delta/5$, for all $h \in [H]$,*

$$\mathop{\mathbb{E}}_{(S_h,A_h)\sim\mu_h} \min\Big\{1, \|\phi(S_h,A_h)\|^2_{\hat{X}_h^{-1}}\Big\} \le \check{\epsilon} = \tilde{\mathcal{O}}(d/n)\,,$$

where $\check{\epsilon}$ is defined in Eq. (40).

Proof Fix any $h \in [H]$. Let

$$\mathcal{H}_t \coloneqq \sigma\{(S_h^j, A_h^j)_{j\in[t]}\}$$

be the sigma algebra generated by the first t samples at stage h. Let

$$V_t \coloneqq \lambda I + \sum_{j\in[t]} \phi(S_h^j, A_h^j)\phi(S_h^j, A_h^j)^\top$$

be the unnormalized regularized empirical covariance matrix based on the first t samples at stage h. For any $t \in [n+1]$, let

$$q_t \coloneqq \|\phi(S_h^t, A_h^t)\|^2_{V_{t-1}^{-1}}, \quad Z_t \coloneqq \min\{1, q_t\}, \quad u_t \coloneqq \mathbb{E}[Z_t|\mathcal{H}_{t-1}] = \mathbb{E}\Big[\min\{1, \|\phi(S_h^t, A_h^t)\|^2_{V_{t-1}^{-1}}\} \mid \mathcal{H}_{t-1}\Big]\,,$$

where S_h^{n+1}, A_h^{n+1} is an independent sample from μ_h. Since $V_n = \hat{X}_h$, we have that

$$\mathop{\mathbb{E}}_{(S_h,A_h)\sim\mu_h} \min\Big\{1, \|\phi(S_h,A_h)\|^2_{\hat{X}_h^{-1}}\Big\} = \mathbb{E}\Big[\min\Big\{1, \|\phi(S_h^{n+1}, A_h^{n+1})\|^2_{V_n^{-1}}\Big\} \mid \mathcal{H}_n\Big] = u_{n+1}\,.$$

Thus, our goal is to bound u_{n+1} with high probability. The proof will proceed with the following steps:

1. We show that $u_{n+1} \le \frac{1}{n}\sum_{t\in[n]} u_t$, since u_t is non-increasing in t due to $V_t \succcurlyeq V_{t-1}$.
2. We show that $\sum_{t\in[n]} Z_t \le \tilde{\mathcal{O}}(d)$ deterministically by using the elliptical potential lemma.
3. We show that with high probability, $\sum_{t\in[n]} u_t \le 2\sum_{t\in[n]} Z_t + 2\log(1/\delta')$ by using a concentration argument.
4. Combining the above steps gives that with high probability, $u_{n+1} \le \tilde{\mathcal{O}}(d/n)$.

Step 1: Since $V_t \succcurlyeq V_{t-1}$ for all $t \in [n]$, we have that $\|\phi(S,A)\|^2_{V_t^{-1}} \le \|\phi(S,A)\|^2_{V_{t-1}^{-1}}$ for all $(S,A) \in \mathcal{S}_h \times \mathcal{A}$. For an independent sample $S, A \sim \mu_h$ we have that

$$\begin{aligned} u_{t+1} = \mathbb{E}\Big[\min\{1, \|\phi(S,A)\|^2_{V_t^{-1}}\} \mid \mathcal{H}_t\Big] &\le \mathbb{E}\Big[\min\{1, \|\phi(S,A)\|^2_{V_{t-1}^{-1}}\} \mid \mathcal{H}_t\Big] \\ &= \mathbb{E}\Big[\min\{1, \|\phi(S,A)\|^2_{V_{t-1}^{-1}}\} \mid \mathcal{H}_{t-1}\Big] = u_t\,. \end{aligned}$$

Thus, u_t is non-increasing in t, which gives that $u_{n+1} \le \frac{1}{n}\sum_{t\in[n]} u_t$.
Step 2: By the elliptical potential lemma (see, e.g., Lemma 19.4 in [Lattimore and Szepesvári, 2020]), we have that deterministically,

$$\sum_{t\in[n]} Z_t \le 2d\log\left(\frac{\mathrm{tr}(V_0) + nL_\phi^2}{d\det(V_0)^{1/d}}\right) = 2d\log\left(\frac{\lambda d + nL_\phi^2}{\lambda d}\right).$$

Step 3: For any $z \in [0,1]$, it holds that $e^{-z} \le 1 - cz \le e^{-cz}$ with $c \coloneqq 1 - e^{-1}$. Since $Z_t \in [0,1]$ for all $t \in [n]$, we have that

$$\mathbb{E}\big[e^{-Z_t} \mid \mathcal{H}_{t-1}\big] \le \mathbb{E}[1 - cZ_t \mid \mathcal{H}_{t-1}] = 1 - cu_t \le e^{-cu_t}\,.$$

This implies that

$$\mathbb{E}[\exp(cu_t - Z_t) \mid \mathcal{H}_{t-1}] \le 1\,.$$

Let

$$B_t \coloneqq \exp(cu_t - Z_t), \quad M_t \coloneqq \prod_{j\in[t]} B_j = \exp\left(c\sum_{j\in[t]} u_j - \sum_{j\in[t]} Z_j\right).$$

Since M_{t-1} is $\mathcal{H}_{t-1}$ measurable, we have that

$$\mathbb{E}[M_t \mid \mathcal{H}_{t-1}] = \mathbb{E}[M_{t-1}B_t \mid \mathcal{H}_{t-1}] = M_{t-1}\mathbb{E}[B_t \mid \mathcal{H}_{t-1}] \le M_{t-1}\,.$$

Thus, M_t is a nonnegative supermartingale. By the supermartingale property we have that $\mathbb{E}[M_n] \le \mathbb{E}[M_0] = 1$. Markov's inequality states that

$$\mathbb{P}\big(M_n \ge 1/\delta'\big) \le \delta'\mathbb{E}[M_n] \le \delta'\,.$$

Thus, with probability at least $1 - \delta'$,

$$c\sum_{t\in[n]} u_t - \sum_{t\in[n]} Z_t = \log(M_n) \le \log(1/\delta')\,.$$

Rearranging gives that with probability at least $1 - \delta'$,

$$\sum_{t\in[n]} u_t \le \frac{1}{c}\sum_{t\in[n]} Z_t + \frac{1}{c}\log(1/\delta')\,.$$

Step 4: Combining the above steps gives that with probability at least $1 - \delta'$,

$$u_{n+1} \le \frac{1}{n}\sum_{t\in[n]} u_t \le \frac{1}{(1-e^{-1})n}\left(\sum_{t\in[n]} Z_t + \log(1/\delta')\right) \le \frac{4}{n}\left(d\log\left(\frac{\lambda d + nL_\phi^2}{\lambda d}\right) + \log(1/\delta')\right) = \tilde{\mathcal{O}}(d/n)\,.$$

To conclude, we can set $\delta' = \delta/(5H)$ and apply a union bound over all $h \in [H]$ to get that with probability at least $1 - \delta/5$, for all $h \in [H]$,

$$\begin{aligned}\mathop{\mathbb{E}}_{(S_h,A_h)\sim\mu_h} \min\Big\{1, \|\phi(S_h,A_h)\|^2_{\hat{X}_h^{-1}}\Big\} &\le \frac{4}{n}\left(d\log\left(\frac{\lambda d + nL_\phi^2}{\lambda d}\right) + \log(5H/\delta)\right) \\ &=: \check{\epsilon} = \tilde{\mathcal{O}}(d/n)\,. \end{aligned} \tag{40}$$

Which is the desired result. ■

By Lemma 26, with probability at least $1-\delta/5$,

$$4\sqrt{C_0}\beta\sum_{t=h}^{H}\sqrt{\mathop{\mathbb{E}}_{(S_t,A_t)\sim\mu_t}\min\{1,\|\phi(S_t,A_t)\|^2_{\hat{X}_t^{-1}}\}} \le 4\sqrt{C_0}\beta\sum_{t=h}^{H}\sqrt{\tilde{\epsilon}}$$

Putting the results from the proof of the second claim together, and using a union bound, we get that with probability at least $1-3\delta/5$, for all $h\in[H]$,

$$\frac{1}{n}\sum_{j\in[n]}\left(\max_{\theta\in\Theta_{G,h}}\bar{q}_\theta(S_h^j,A_h^j)-\min_{\theta\in\Theta_{G,h}}\bar{q}_\theta(S_h^j,A_h^j)\right)\le\zeta_1+4\sqrt{C_0}H\beta\sqrt{\tilde{\epsilon}}=:\bar{\epsilon}\,. \tag{41}$$

Applying a union bound over the failure events of the two claims completes the proof of Lemma 9. ■

C.5. Proof of Lemma 10

Proof The proof follows the same steps as the proof of Lemma 9, except $\mathbf{G}_{\text{eval}}$, $\Theta^{\text{eval}}_{G,h}$ and π^e_G, are used instead of $\mathbf{G}_{\text{opt}}$, $\Theta^{\text{opt}}_{G,h}$ and $\bar{\pi}^\star_G$. Also, the lemmas need to be changed to hold with probability at least $1-\delta/10$ instead of $1-\delta/5$. ■

Appendix D. Useful Results

Lemma 27 (Hoeffding's Inequality (Theorem 2 in [Hoeffding, 1994])) *Let $(X_i)_{i\in\mathbb{N}}$ be independent random variables such that $X_i\in[a,b]$ for some $a,b\in\mathbb{R}$, and let $S_n=\frac{1}{n}\sum_{i=1}^n X_i$. Then, with probability at least $1-\zeta$ it holds that*

$$|\mathbb{E}\,S_n-S_n|\le\frac{(b-a)}{\sqrt{n}}\sqrt{\log\left(\frac{2}{\zeta}\right)}\,.$$

Lemma 28 (Covering number of the Euclidean ball) *Let $a>0,\epsilon>0,d\ge1$, and $\mathcal{B}_d(a)=\{x\in\mathbb{R}^d:\|x\|_2\le a\}$ denote the d-dimensional Euclidean ball of radius a centered at the origin. The covering number of $\mathcal{B}_d(a)$ is upper bounded by $\left(1+\frac{2a}{\epsilon}\right)^d$.*

Proof Same as the proof of Corollary 4.2.13 in [Vershynin, 2018] with $\mathcal{B}(1)$ replaced with $\mathcal{B}(a)$. ■

Lemma 29 (Performance Difference Lemma (Lemma 3.2 in [Cai et al., 2020])) *For any policies $\pi,\bar{\pi}$, it holds that*

$$v^\pi(s_1)-v^{\bar{\pi}}(s_1)=\sum_{h=1}^{H}\mathop{\mathbb{E}}_{(S_h,A_h)\sim\mathbb{P}^h_{\pi,s_1}}\left(q^{\bar{\pi}}(S_h,A_h)-v^{\bar{\pi}}(S_h)\right)\,.$$