IEEE CSS

Received 31 March 2022; revised 28 June 2022; accepted 29 July 2022. Date of publication 16 August 2022;
date of current version 7 September 2022. Recommended by Senior Editor Sonia Martinez.

Digital Object Identifier 10.1109/OJCSYS.2022.3198660

Finite Sample Analysis of Minmax Variant of Offline Reinforcement Learning for General MDPs

JAYANTH REDDY REGATTI (Student Member, IEEE), AND ABHISHEK GUPTA (Member, IEEE)

The Ohio State University, Columbus, OH 43210 USA

CORRESPONDING AUTHOR: JAYANTH REDDY REGATTI (e-mail: regatti.1@osu.edu).

This work was supported by the Army Research Lab under Grant W911NF1920256.

ABSTRACT In this work, we analyze the finite sample complexity bounds for offline reinforcement learning with general state, general function space and state-dependent action sets. The algorithm analyzed does not require the knowledge of the data-collection policy as compared to earlier works. We show that one can compute an ϵ-optimal Q function (state-action value function) using $O(1/\epsilon^4)$ i.i.d. samples of state-action-reward-next state tuples.

INDEX TERMS Machine learning, reinforcement learning, statistical learning.

I. INTRODUCTION

Reinforcement learning (RL) has attracted significant interest by the research community in the last decade, inspired by the early successes of deep reinforcement learning [1], [2]. However, online RL algorithms require access to the real environment throughout training and require large datasets, which are generated through interactions with the environment. This either requires a high fidelity simulator that mimics the environment, or requires the cost of interactions with the environment to be low. These are not practical for many real world problems. Building a high fidelity simulator for physical systems is very expensive [3], [4]. Many simulators also suffer from generalization issues due to the gap between the simulation and the real environment [5], [6]. Similarly, the cost of interactions with the environment in tasks related to healthcare and autonomous driving are very expensive and sometimes impractical [7]. Moreover, in such safety critical applications, it is not safe to deploy semi-trained policies in the real environments making online RL difficult to deploy for learning optimal policies. This has been a major hindrance to the adoption of reinforcement learning algorithms for deployment in sequential decision making problems [7]. One solution to this is to learn the policy from logged data and is often called as Batch Reinforcement learning or Offline Reinforcement learning (ORL) [7], [8].

Fitted value iteration is a class of approximate dynamic programming algorithms that approximate the value function [9], [10], [11], [12], [13] using a finite set of data and a suitable function approximating class. The finite dataset typically comprises state, action, next state and reward obtained over long periods of time. Fitted Q Iteration is a special case where the function approximated is the state-action value function, or the Q function. Several studies have been conducted on analyzing the finite sample properties of the fitted value iteration under various settings [10], [11], [14], [15], [16]. For Fitted Value Iteration in [11], the authors assume availability for multiple data points at a given state (or state-action pair if using for ORL) enabling sufficient data coverage. However, this is a strong assumption in practical offline RL setup where the data coverage may not be sufficient. More recently [17] provided a simpler method to compute the finite sample analysis for ORL and a min-max variant algorithm, and provided sharp convergence guarantees by using specialized concentration inequalities. Moreover, [17] do not assume availability of multiple samples as in [11]. However, [17] analysis only holds for the case when the function space is finite dimensional. Reference [14] also studies the min-max variant algorithm for general function space and single sample path case, but assume that the data collection policy is known (which is not always possible in real world

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

examples). Reference [18] and [19] studied sharp convergence guarantees of Fitted Q Iteration for general function space.

In the aforementioned analyses, the action space is assumed to be unconstrained by the state the agent is in. This does not hold true in several real world problems where there are physical or safety constraints that govern the permissible actions in any given state. These constraints are common place in robotics, economics, e-commerce, inventory management, [20], [21], [22], [23], [24], etc. Recent work by [25] addresses this by studying the asymptotic analysis of fitted Q iteration with multiple samples per state-action pair and under the state-dependent action set constraints by assuming some smoothness properties on the function approximation space and the MDP. In this work, we further generalize the work and derive the sample complexity guarantees for min-max variant of Fitted Q Iteration for general function space. We assume the availability of an i.i.d. dataset about state, action, reward, and next state, no knowledge of the policy used to collect the dataset, and some other mild assumptions on the MDP with denumerable state and action spaces.

A. EXAMPLES OF STATE-CONSTRAINED MDPS

We provide here four practical applications where state and action spaces are denumerable and the permissble actions are dependent on the state of the system.

1) ECO-DRIVING IN CONNECTED AND AUTOMATED VEHICLES (CAVS)

Consider an automated vehicle which can receive future signal phase and timing information and traffic information via vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communication respectively. Using this information, the CAV can optimize the vehicle speed and battery state-of-charge, which are the states of the optimal control problem aimed at minimizing the energy consumption [26], [27]. The control actions are the engine and electric motor torques which are generally computed from the non-linear engine and motor torque-speed curves respectively. Both engine and motor speed eventually depend on the vehicle speed as formulated in [28] and [29]. In this application, the set of actions is constrained by the current state of the system, the policy used for collecting data is generally not available due to complex interactions among the subsystems, and a huge amount of offline data is available.

2) VEHICLE REBALANCING IN RIDEHAILING SYSTEMS

Consider a vehicle rebalancing problem, where vehicles are relocated to meet customer's demands. In [30], each vehicle is modeled as an agent and the state of each vehicle consists of (i) vehicle state (empty, hasPassengers, full) (ii) presence of current active requests. The action set is {pickUp, rebalance, doNothing} and it is state-dependent. E.g., pickup action can only be executed when last passenger is dropped off or the relocation destination is reached. Offline reinforcement learning with state-dependent action constraint can be used to derive the optimal rebalancing policies for the vehicles [31].

3) ROBOTICS

Robotic vehicles and manipulators in industrial settings have to navigate tight spaces and meet safety regulations. Recent works have focused on robotic safety constraints in reinforcement learning, where the agent uses some constraint barrier functions or is constrained to explore within a safe set of policies [20], [21], [22], [23], among several others. Typically, these safety constraints appear as constraints on actions (and future states) and are dependent on the current state.

4) ONLINE ADVERTISEMENTS

Consider the problem of search based online advertising [32]. Here, a search platform displays ads relevant to queries entered by a user by allowing advertisers to bid on each query. An auto-bidding agent is an automated algorithm that determines a bid (dollar value) depending on the relevance of the user query to the advertiser's choice of bidded keywords. Whenever an ad is clicked by a user, the advertiser pays some amount to the search platform that is determined by the auction mechanism. The goal of the agent is to maximize the number of ad clicks on a given day where the spending is constrained by a fixed daily budget. The agent must therefore balance between aggressively bidding for the current search query and saving budget for future search queries. We studied this problem recently in the offline RL setting where the auto-bidding agent is trained from past data of the bids [33]. The auto-bidding agent has information about the past spend (on that particular day) along with several other features about the query and the likelihood of a click. The past spend is used to determine the budget remaining for the day which is a key factor in determining the bid amount for future queries on that day. The auto-bidding agent can not bid more than the daily budget and the participation of the auto-bidding agent stops when the daily budget is depleted. In this setting, the admissible actions are constrained by the current state.

B. NOTATION

Let $\mathcal{X}$ be a measurable space. We use $\Delta(\mathcal{X})$ to denote the set of all probability measures on the space $\mathcal{X}$. Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function and $\mu \in \Delta(\mathcal{X})$. We denote the (p, μ) norm of the function f as

$$\|f\|_{p,\mu} = \sqrt[p]{\int |f(x)|^p d\mu}.$$

Let X be a random variable taking values in $\mathcal{X}$ with distribution $\mu \in \Delta(\mathcal{X})$. Let $f : \mathcal{X} \to \mathbb{R}$. Define $\mathbb{V}$ as the variance

$$\mathbb{V}_{x\sim\mu}[f(X)] = \int (f(x) - \mathbb{E}[f(X)])^2 \, d\mu \tag{1}$$

The set of all continuous and bounded functions $f : \mathcal{X} \to \mathbb{R}$ is denoted by $\mathcal{C}_b(\mathcal{X})$ and measurable functions $f : \mathcal{X} \to \mathbb{R}$ is denoted by $\mathcal{M}(\mathcal{X})$.

II. PROBLEM FORMULATION

Let the MDP be defined by the tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$, where $\mathcal{S}$ is the state space (which can be finite or continuous), $\mathcal{A}$ be the action space (finite or continuous). Let η_{init} be the distribution of the starting state. At a state $s \in \mathcal{S}$, the set of feasible actions is given by $\Gamma(s) \subseteq \mathcal{A}$. We use $\mathcal{B}$ to denote the feasible state-action pairs: $\mathcal{B} = \{(s, a) \in \mathcal{S} \times \mathcal{A} \,|\, a \in \Gamma(s)\}$. The reward function is denoted by $R : \mathcal{B} \to [0, R_{\max}]$. This is common since in most practical applications, the reward is bounded. The transition kernel of the MDP which determines the state dynamics is denoted by $P : \mathcal{B} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ denotes the set of all probability distributions over $\mathcal{S}$. We use γ to denote the discount factor.

Let $V^\pi : \mathcal{S} \to \mathbb{R}$ denote the value function defined by

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h, a_h) \Big| s_1 = s, a_h \sim \pi(\cdot|s_h)\right]$$

The goal is to learn a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maximizes $v^\pi := \mathbb{E}_{s\sim\eta_{\text{init}}}[V^\pi(s)]$. Let $Q^\pi : \mathcal{B} \to \mathbb{R}$ denotes the state-action value function (also called Q function) as

$$Q^\pi(s, a) = \mathbb{E}\Bigg[\sum_{h=1}^{\infty} \gamma^{h-1} R(s_h, a_h) | s_1 = s, \\ a_1 = a, a_h \sim \pi(\cdot|s_h), h \geq 2\Bigg]$$

It is clear that since the reward is bounded by $R_{\max}$ and due to the discount factor, we have $\|V^\pi\|_\infty \leq V_{\max} = \frac{R_{\max}}{1-\gamma}$ and $\|Q^\pi\|_\infty \leq V_{\max}$. We make the following assumptions for our analysis.

Assumption 1: The following holds:

1) The set $\mathcal{B}$ is a compact subset of a Euclidean space.
2) The reward function R is continuous.
3) The correspondence $\Gamma : \mathcal{S} \to \mathcal{A}$ is upper hemicontinuous.
4) The state transition kernel P is weakly continuous when $\Delta(\mathcal{S})$ is endowed with the usual weak topology.

It is a common assumption to make when the state space and action spaces are denumerable; see, for example, Assumption 3.3.3 in [34], p. 28]. Under these assumptions, there exists an optimal policy π^* (see Chapter 4 [34] for more details). Let V^*, Q^* denote the corresponding value and state-action value functions. Denote by $v^* = \mathbb{E}_{s\sim\eta_{\text{init}}}[V^*(s)]$.

For a function $f \in \mathcal{C}_b(\mathcal{B})$, let $V_f(s') = \max_{a'\in\Gamma(s')} f(s', a')$. We define the Bellman operator $\mathcal{T} : \mathcal{M}(\mathcal{B}) \to \mathcal{M}(\mathcal{B})$ as

$$(\mathcal{T}f)(s, a) := R(s, a) + \gamma \mathbb{E}_{s'\sim P(s,a)}[V_f(s')].$$

The optimal Q function satisfies $Q^* = \mathcal{T}Q^*$. The goal of the agent is to design an algorithm to compute Q^* that reduces the Bellman error to 0, that is, satisfies $\|Q^* - \mathcal{T}Q^*\|_{2,\mu} = 0$.

A. DATA COLLECTION POLICY AND ORL PROBLEM

The offline dataset is constructed by using a stationary policy $\pi_b : \mathcal{S} \to \Delta(\mathcal{A})$ with the environment. We refer to this the data collection policy or the behaviour policy. The past interactions with the environment using the behavior policy are logged as the dataset $D^{1:n} := (s_i, a_i, r_i, s_i')_{i=1}^n$, which we assume is independent and identically distributed. We use $\mu \in \Delta(\mathcal{B})$ to denote the stationary distribution (occupation measure of state-action pair) of the MDP under the stationary policy π_b. Therefore, it follows that (s_i, a_i) is drawn i.i.d from μ for $i \in [n]$. Note that since in practical settings the behavior policy is not known, here we do not assume the knowledge of π_b.

Consider two function approximating classes $\mathcal{F}, \mathcal{G} \subset \{f : \mathcal{B} \to [0, V_{\max}] : f \in \mathcal{C}_b(\mathcal{B})\}$. These function classes could be neural networks, RKHS, non-parametric function approximator, etc. The ORL problem is learn a state-action value function $f \in \mathcal{F}$ using dataset D such that the Bellman residual $\|f - \mathcal{T}f\|_{2,\mu}$ is minimized. Fitted Q iteration attempts to solve the ORL problem by using a rich function approximating class and iterative use of Bellman residual minimization on the empirical risk

$$\mathcal{L}_D(f, f') = \frac{1}{n}\sum_{i=1}^{n} \left(f(s_i, a_i) - r_i - \gamma V_{f'}(s_i')\right)^2.$$

We now define the operator $\widehat{\mathcal{T}}_{\mathcal{G}} : \mathcal{F} \to \mathcal{G}$ such that

$$\widehat{\mathcal{T}}_{\mathcal{G}} f = \arg\min_{g\in\mathcal{G}} \mathcal{L}_D(g, f), \text{ where } f \in \mathcal{F}. \tag{2}$$

Fitted Q Iteration (FQI) using the function approximating class $\mathcal{F}$ involves iteratively applying the operator $\widehat{\mathcal{T}}_{\mathcal{F}}$, i.e.,

$$f_{t+1} = \widehat{\mathcal{T}}_{\mathcal{F}} f_t.$$

Remark 1: In some cases, the dataset $D^{1:n}$ may be very large. In this case, at iteration t, techniques can be used to create a smaller dataset $D_t' \subset D^{1:n}$, which is used for evaluating the operator in (2). We do not analyze this setting in this paper.

B. KEY DIFFICULTIES AND SOLUTION APPROACH

Unlike supervised learning problems, Bellman residual minimization can not be solved using empirical risk minimization. In other words, the expectation of the empirical risk does not equal to $\|f - \mathcal{T}f\|_{2,\mu}^2$ – it over-estimates the Bellman error by a variance term as has been demonstrated in [14], [17]. To see this, let us define $\mathcal{L}_\mu(f; f') = \mathbb{E}[\mathcal{L}_D(f, f')]$ where the expectation is taken with respect to the draw of the dataset $D^{1:n}$, i.e., $(s, a) \sim \mu$, $s' \sim P(s, a)$. For completeness, we show in Appendix C that

$$\mathcal{L}_\mu(f; f) = \|f - \mathcal{T}f\|_{2,\mu}^2 + \mathbb{E}_{(s,a)\sim\mu}\left[\mathbb{V}_{s'\sim P(s,a)}[V_f(s')]\right].$$

One approach to addressing this is to draw two uncorrelated samples in computation of $L_D(f, f)$ [14], i.e., for every state action pair, two next states should be sampled according to $P(s, a)$. However, this assumption is not practical in the continuous-state continuous-action ORL setting since we can not guarantee that multiple next states can be sampled for a given state-action pair or that the same state-action pair is visited twice while collecting the dataset.

TABLE 1 A Summary of Prior Works on the Analysis of the Min-Max Variant.

Work	Function Space	Knowledge of data collection policy	Sample Complexity
[14]	General	$\surd$	$\mathcal{O}\left(\frac{1}{\epsilon^4}\right)$
[17]	Finite	$\times$	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
This paper	General	$\times$	$\mathcal{O}\left(\frac{1}{\epsilon^4}\right)$

Another approach followed is to estimate the variance and subtract it from the empirical objective, and this approach results in the following min-max formulation of the offline RL algorithm [17]

$$\widehat{f} := \arg \inf_{f\in\mathcal{F}} \sup_{g\in\mathcal{G}} \mathcal{L}_D(f; f) - \mathcal{L}_D(g; f) \tag{3}$$

where $\mathcal{G} \subset \{g : \mathcal{B} \to [0, V_{\max}] | g \in \mathcal{C}_b(\mathcal{B})\}$ is another rich function class that is continuous in the actions. In this work, we study the finite sample complexity of this algorithm in the general state space and general function space setting.

Reference [17] study this algorithm under the finite state space, finite action space and when the function space is finite. The finite function space assumption allows them to use a simple union bound argument along with the Bernstein inequality to get the sharp sample complexity. In this paper, we assume a general function space and use a covering number argument to achieve the sample complexity bound. Reference [14] study a similar algorithm, however, the algorithm there requires the knowledge of the behavior policy (another difference being that they study the case where the data is generated from a single sample path and for finite action space). In particular, their objective is

$$\arg \inf_{f\in\mathcal{F}} \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\pi_b(a_i|s_i)} \Big[\left(f\left(s_i, a_i\right) - r_i - \gamma V_{f'}\left(s_i'\right)\right)^2 - \left(g\left(s_i, a_i\right) - r_i - \gamma V_{f'}\left(s_i'\right)\right)^2\Big]$$

where π_b is the behavior policy used to collect the single sample path data. In contrast, here we analyze the algorithm when such knowledge of the data collection policy is unknown. For a general state space and general action space, estimating the data collection policy from the finite data leads to high variance estimates, subsequently affecting the fitted learning objective. The results of previous works studying the min-max objective are presented in Table 1.

C. PRELIMINARIES

We now introduce the following two quantities that capture the strength of the function approximation spaces $\mathcal{F}$ and $\mathcal{G}$:

$$\epsilon_{\mathcal{F}} = \inf_{f\in\mathcal{F}} \|f - \mathcal{T}f\|_{2,\mu}^2, \quad \epsilon_{\mathcal{F},\mathcal{G}} = \sup_{f\in\mathcal{F}} \inf_{g\in\mathcal{G}} \|g - \mathcal{T}f\|_{2,\mu}^2.$$

If Q^* is realizable in $\mathcal{F}$, i.e., $Q^* \in \mathcal{F}$, then $\epsilon_{\mathcal{F}} = 0$. When $\mathcal{F} = \mathcal{G}$, $\epsilon_{\mathcal{F},\mathcal{G}}$ is called the inherent Bellman error. We provide the finite sample guarantees based on $\epsilon_{\mathcal{F}}$ and $\epsilon_{\mathcal{F},\mathcal{G}}$, therefore it is inherently assumed that these quantities are small in order to control the bounds.

A distribution $\nu \in \Delta(\mathcal{B})$ is admissible in MDP, if there exists $h \geq 1$, and a potentially non-stationary and stochastic policy $\pi := (\pi_1, \pi_2, \ldots)$ such that

$$\nu(ds, da) = \mathbb{P}[s_h \in ds, a_t \in da | s_1 \sim \eta_{\text{init}}, a_t \sim \pi_t(\cdot|s_t)]$$

We denote $s' \sim P(\nu)$ as a shorthand for $(s, a) \sim \nu, s' \sim P(s, a)$. Also, we define $\pi_{f,f'}(s)$ as

$$\pi_{f,f'}(s) = \arg \max_{a\in\Gamma(s)} \max\left\{f(s, a), f'(s, a)\right\}. \tag{4}$$

For every given $f \in \mathcal{F}$, denote $g_f^* = \arg\min_{g\in\mathcal{G}} \|g - \mathcal{T}f\|_{2,\mu}$ and observe that $\|g_f^* - \mathcal{T}f\|_{2,\mu}^2 \leq \epsilon_{\mathcal{F},\mathcal{G}}$.

Definition 1: Define the class of functions $Z_{\mathcal{F}} = \{Z_f : \mathcal{B} \times \mathbb{R} \times \mathcal{S} \to \mathbb{R} \,|\, f \in \mathcal{F}\}$ such that

$$Z_f\left(s, a, r, s'\right) := \left(f(s, a) - r - \gamma V_f\left(s'\right)\right)^2 - \left((\mathcal{T}f)(s, a) - r - \gamma V_f\left(s'\right)\right)^2 \tag{5}$$

Definition 2: Define a function class $X_{\mathcal{F}} : \{X_{g,f,g_f^*} : \mathcal{B} \times \mathbb{R} \times \mathcal{S} \to \mathbb{R} \,|\, f, g \in \mathcal{F}\}$ where

$$X_{g,f,g_f^*}\left(s, a, r, s'\right) := \left(g(s, a) - r - \gamma V_f\left(s'\right)\right)^2 - \left(g_f^*(s, a) - r - \gamma V_f\left(s'\right)\right)^2. \tag{6}$$

Definition 3: Define the function class $Y_{\mathcal{F},\mathcal{G}} : \{Y_{g,f} : \mathcal{B} \times \mathbb{R} \times \mathcal{S} \to \mathbb{R} \,|\, g \in \mathcal{G}, f \in \mathcal{F}\}$ such that

$$Y_{g,f}\left(s, a, r, s'\right) = \left(g(s, a) - r - \gamma V_f\left(s'\right)\right)^2 - \left((\mathcal{T}f)(s, a) - r - \gamma V_f\left(s'\right)\right)^2. \tag{7}$$

We use $\mathrm{d}_{X_{\mathcal{F}}}$, $\mathrm{d}_{Z_{\mathcal{F}}}$ and $\mathrm{d}_{Y_{\mathcal{F},\mathcal{G}}}$ to be the pseudo-dimension of the function classes described above. These notations are introduced in Definition 7 (Appendix B).

III. ASSUMPTIONS AND MAIN RESULTS

Our main assumption on the MDP is that we assume the existence of finite concentrability coefficient from [17] (for the case of finite state and finite action space). We now state the assumption adapted to the current setup of state dependent action sets.

Assumption 2 (Concentrability coefficient): For all admissible $\nu \in \Delta(\mathcal{B})$, we assume that $C < \infty$ such that $\|\frac{d\nu}{d\mu}\|_\infty \leq C$.

The above assumption implies that the transitions are sufficiently stochastic and $\nu(\cdot, \cdot) \leq C\mu(\cdot, \cdot)$, $\forall (s, a) \in \mathcal{B}$. Note that this assumption is much stronger than the usual discounted average concentrability of future states [11].

We next assume that the finiteness of the capacity of the function approximation class since our sample complexity bounds depend on the function class capacity.

Assumption 3 (Finite capacity of function classes): The pseudo-dimensions $\mathtt{d}_{X_\mathcal{F}}$, $\mathtt{d}_{Z_\mathcal{F}}$ and $\mathtt{d}_{Y_{\mathcal{F},\mathcal{G}}}$ are all assumed to be finite.

Remark 2: A sufficient condition that ensures that the function classes have finite capacity is discussed in [25], [35]. In [35], the author shows that the optimal value/ Q function (under state dependent action constraints) is Lipschitz continuous under the following assumptions: the transition function is Lipschitz continuous in (s, a), the reward function is Lipschitz in (s, a), the correspondence Γ is Lipschitz continuous under the Hausdorff metric, and the Bellman operator is a contraction. In addition, if we assume $\mathcal{F}$ and $\mathcal{G}$ are Lipschitz continuous function classes and Γ is Lipschitz continuous correspondence, then it can be shown that $Z_\mathcal{F}, X_\mathcal{F}, Y_{\mathcal{F},\mathcal{G}}$ are also Lipschitz continuous using Lemma 3.2 in [35]. The finite capacity of the Lipschitz and uniformly bounded function class follows from Theorem 2.7.1 and 2.7.11 of [36].

We now state the finite sample analysis result of the offline RL algorithm (3).

Theorem 1 (Error bound for min-max): Suppose Assumptions 1, 2, and 3 hold. Given a dataset $D = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$, two classes of bounded functions $\mathcal{F}, \mathcal{G}$ and $\epsilon, \delta > 0$, then with probability at least $1 - \delta$, the output policy of (3), $\pi_{\widehat{f}}$ satisfies

$$v^* - v^{\pi_{\widehat{f}}} \leq \frac{2\sqrt{C}}{(1-\gamma)^2}\left(\sqrt{\epsilon + \epsilon_\mathcal{F} + \epsilon_{\mathcal{F},\mathcal{G}}}\right) \tag{8}$$

when

$$n \geq \frac{K_1 V_{\max}^4}{\epsilon^2}\Bigg[\log\frac{16e}{\delta} + \log\left(2(\mathtt{d}_{X_\mathcal{F}}+1)\left(\frac{K_2 e V_{\max}^2}{\epsilon}\right)^{\mathtt{d}_{X_\mathcal{F}}} + (\mathtt{d}_{Y_{\mathcal{F},\mathcal{G}}}+1)\left(\frac{K_2 e V_{\max}^2}{\epsilon}\right)^{\mathtt{d}_{Y_{\mathcal{F},\mathcal{G}}}} + (\mathtt{d}_{Z_\mathcal{F}}+1)\left(\frac{K_2 e V_{\max}^2}{\epsilon}\right)^{\mathtt{d}_{Z_\mathcal{F}}}\right)\Bigg],$$

where $\mathtt{d}_{X_\mathcal{F}}, \mathtt{d}_{Y_{\mathcal{F},\mathcal{G}}}, \mathtt{d}_{Z_\mathcal{F}}$ are the pseudo-dimensions of the spaces $X_\mathcal{F}, Y_{\mathcal{F},\mathcal{G}}, Z_\mathcal{F}$ respectively, and $K_1 = 64 \times 128 \times 36$ and $K_2 = 6 \times 64$.

Remark 3: The sample complexity does not get affected by arbitrary re-scaling of the reward function as long as ϵ is scaled by the square of the scaling factor for the reward, and each function in the function class $\mathcal{F}$ and $\mathcal{G}$ also get rescaled by the same factor. Observe that when we rescale the reward function R by $p > 0$ and ϵ by p^2, (i) $V_{\max}$, v^*, and $v^{\pi_{\widehat{f}}}$ also get rescaled by p, (ii) $\frac{V_{\max}^2}{\epsilon}$ does not get rescaled, and (iii) the $\epsilon_\mathcal{F}$ and $\epsilon_{\mathcal{F},\mathcal{G}}$ terms get rescaled as by p^2. Thus, if we scale ϵ to $p^2\epsilon$, then scaling term appears on sides of the inequality in (8) and the lower bound on n remains the same. Consequently, the sample complexity result remains unchanged.

1) DEPENDENCE ON FUNCTION CLASS

We can observe that, the error depends on $\epsilon_\mathcal{F}, \epsilon_{\mathcal{F},\mathcal{G}}$. When the function class considered is sufficiently rich (such as a neural network class or RKHS), we can assume that $Q^* \in \mathcal{F}$ and $\mathcal{T}f \in \mathcal{F}$, which results in $\epsilon_\mathcal{F} = 0$ and $\epsilon_{\mathcal{F},\mathcal{G}} = 0$.

When $\mathcal{F} = \mathcal{G}$, observe that the function classes $Z_\mathcal{F}$ and $Y_{\mathcal{F},\mathcal{G}}$ are the same, i.e. $(\mathtt{d}_{Z_\mathcal{F}} = \mathtt{d}_{Y_{\mathcal{F},\mathcal{G}}})$. In addition, when the function class is sufficiently rich where $\epsilon_{\mathcal{F},\mathcal{G}} = 0$, then $g_f^* = \mathcal{T}f$ which implies that the function class $X_\mathcal{F}$ is equal to $Z_\mathcal{F}$ and $Y_{\mathcal{F},\mathcal{G}}$. The result then is simplified as follows: The following holds with probability at least $1 - \delta$,

$$v^* - v^{\pi_{\widehat{f}}} \leq \frac{2\sqrt{C}}{(1-\gamma)^2}\left(\sqrt{\epsilon}\right)$$

when

$$n \geq \frac{K_1 V_{\max}^4}{\epsilon^2}\left[\log\frac{64e(\mathtt{d}_{X_\mathcal{F}}+1)}{\delta} + \mathtt{d}_{X_\mathcal{F}}\left(\frac{K_2 e V_{\max}^2}{\epsilon}\right)\right].$$

2) DEPENDENCE ON ϵ

Ignoring log terms in the sample complexity bound, we observe that

$$v^* - v^{\pi_{\widehat{f}}} \leq \mathcal{O}(\sqrt{\epsilon})$$

when $n \geq \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$. This shows that, we can achieve an error of ϵ by using approximately $\mathcal{O}(\frac{1}{\epsilon^4})$ data samples. This result is consistent with the earlier results [14].

IV. PROOF OF THEOREM 1

In this section, we prove the main result of the paper. We follow the analysis of [17], however, unlike [17] the focus is not to obtain fast rates of convergence ($\frac{1}{\epsilon^2}$ dependence instead of $\frac{1}{\epsilon^4}$), but to obtain convergence rates for the general state space and state dependent action space setting. Although challenging, with some additional effort, it may be possible to obtain the fast rates such as those in [17], [18] using specialized concentration inequalities (see Section V).

A. PROOF OUTLINE

It is evident that $v^* - v^{\pi_{\widehat{f}}}$ is related to $\mathcal{L}_\mu(\widehat{f}; \widehat{f}) - \mathcal{L}_\mu(\mathcal{T}\widehat{f}; \widehat{f})$. Thus, we need to bound $\mathcal{L}_\mu(\widehat{f}; \widehat{f}) - \mathcal{L}_\mu(\mathcal{T}\widehat{f}; \widehat{f})$ in terms of its empirical counterpart $\mathcal{L}_D(\widehat{f}; \widehat{f}) - \mathcal{L}_D(\mathcal{T}\widehat{f}; \widehat{f})$ using concentration inequalities. Accordingly, we first derive a decomposition of the empirical term $\mathcal{L}_D(\widehat{f}; \widehat{f}) - \mathcal{L}_D(\mathcal{T}\widehat{f}; \widehat{f})$ into three terms I, II, III and bound each of these terms separately. We state the decomposition lemma below.

Lemma 1 (Decomposition Lemma): For $f^* \in \mathcal{F}$ s.t, $\|f^* - \mathcal{T}f^*\|_{2,\mu}^2 \leq \epsilon_\mathcal{F}$, we have

$$\mathcal{L}_D(\widehat{f}; \widehat{f}) - \mathcal{L}_D(\mathcal{T}\widehat{f}, \widehat{f}) \leq \underbrace{\mathcal{L}_D(f^*; f^*) - \mathcal{L}_D(\mathcal{T}f^*, f^*)}_{I} + \underbrace{|\mathcal{L}_D(\mathcal{T}\widehat{f}, \widehat{f}) - \mathcal{L}_D(\widehat{\mathcal{T}_\mathcal{G}}\widehat{f}, \widehat{f})|}_{II}$$

$$+\underbrace{|\mathcal{L}_D(\mathcal{T}f^*, f^*) - \mathcal{L}_D(\widehat{\mathcal{T}}_{\mathcal{G}} f^*; f^*)|}_{III} \tag{9}$$

Proof: This result is established in Appendix A. ■

In what follows, we divide the proof of Theorem 1 into three steps. We derive an upper bound on $v^* - v^{\pi_{\hat{f}}}$ as a function of $\mathcal{L}_\mu(\widehat{f}; \widehat{f}) - \mathcal{L}_\mu(\mathcal{T}\widehat{f}; \widehat{f})$ in Subsection IV-B. We then bound the three terms noted in Subsection IV-D using the concentration of measures results derived in Subsection IV-C. Finally, we prove Theorem 1 in Subsection IV-E.

B. RELATION BETWEEN $v^* - V^{\pi_{\hat{f}}}$ AND $\mathcal{L}_\mu(\widehat{f}; \widehat{f}) - \mathcal{L}_\mu(\mathcal{T}\widehat{f}; \widehat{f})$

We shall first show the relation between, $v^* - v^{\pi_{\hat{f}}}$ and $\mathcal{L}_\mu(\widehat{f}; \widehat{f}) - \mathcal{L}_\mu(\mathcal{T}\widehat{f}; \widehat{f})$, that will be used to prove the main theorem.

Lemma 2: The following hold true,

1) Let ν be any admissible distribution. Then $\forall f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,
$$\|f\|_{2,\nu} \leq \sqrt{C}\, \|f\|_{2,\mu}\,.$$
2) We denote $\eta_h^\pi := \mathbb{P}[s_h = s | s_1 \sim \eta_1, \pi]$, and π_f as the policy greedy with respect to the state-action value function $f : \mathcal{B} \to \mathbb{R}$, i.e., $\pi_f(s) := \arg\max_{a \in \Gamma(s)} f(s, a)$. Then we have
$$v^* - v^{\pi_f} \leq \sum_{h=1}^{\infty} \gamma^{h-1} \left(\left\| Q^* - f \right\|_{2, \eta_h^{\pi_f} \times \pi^*} + \left\| Q^* - f \right\|_{2, \eta_h^{\pi_f} \times \pi_f} \right).$$
3) Let $f, f' : \mathcal{B} \to \mathbb{R}$. Then we have $\forall\, \nu \in \Delta(\mathcal{B})$,
$$\left\| V_f - V_{f'} \right\|_{2,P(\nu)} \leq \left\| f - f' \right\|_{2,P(\nu) \times \pi_{f,f'}}.$$
4) For an exploratory distribution $\mu \in \Delta(\mathcal{B})$, any distribution $\nu \in \Delta(\mathcal{B})$, policy π, and $f, f' : \mathcal{B} \to \mathbb{R}$, we have
$$\left\| f - Q^* \right\|_{2,\nu} \leq \frac{\sqrt{C}}{1-\gamma} \|f - \mathcal{T}f\|_{2,\mu}\,.$$

Proof: The results can be adapted directly from [17] to the general state space setting in this paper using Assumptions 1 and 2. ■

Lemma 3: For $f, g \in \mathcal{F}$, we have $\|g - \mathcal{T}f\|_{2,\mu}^2 = \mathcal{L}_\mu(g; f) - \mathcal{L}_\mu(\mathcal{T}f; f)$.

Proof: From the definitions in $\mathcal{L}_\mu$ and $\mathcal{L}_D$, we have

$$\begin{aligned}
\mathcal{L}_\mu(g; f) - \mathcal{L}_\mu(\mathcal{T}f; f) &= \mathbb{E}\left[\mathcal{L}_D(g, f) - \mathcal{L}_D(\mathcal{T}f; f)\right] \\
&= \mathop{\mathbb{E}}_{\substack{(s,a)\sim\mu, \\ s'\sim P(s,a)}} \big[(g(s,a) - r - \gamma V_f(s'))^2 \\
&\quad -(\mathcal{T}f(s,a) - r - \gamma V_f(s'))^2\big] \\
&= \mathop{\mathbb{E}}_{\substack{(s,a)\sim\mu, \\ s'\sim P(s,a)}} \big[(g(s,a)^2 - \mathcal{T}f(s,a)^2) \\
&\quad +2(r + \gamma V_f(s'))(\mathcal{T}f(s,a) - g(s,a))\big] \\
&\overset{i}{=} \mathop{\mathbb{E}}_{(s,a)\sim\mu} \big[(g(s,a)^2 - \mathcal{T}f(s,a)^2) \\
&\quad +2\mathcal{T}f(s,a)(\mathcal{T}f(s,a) - g(s,a))\big] \\
&= \mathop{\mathbb{E}}_{(s,a)\sim\mu,} \left[(g(s,a) - \mathcal{T}f(s,a))^2\right] \\
&= \|g - \mathcal{T}f\|_{2,\mu}^2
\end{aligned}$$

where (i) follows from the definition of the operator $\mathcal{T}$. The proof is complete. ■

Lemma 4:
$$v^* - v^{\pi_{\hat{f}}} \leq \frac{2\sqrt{C}}{(1-\gamma)^2} \left(\sqrt{\mathcal{L}_\mu\left(\widehat{f}; \widehat{f}\right) - \mathcal{L}_\mu\left(\mathcal{T}\widehat{f}, \widehat{f}\right)} \right). \tag{10}$$

Proof: Substituting $f = \widehat{f}$ in the result of Lemma 4, we get
$$\left\| \widehat{f} - Q^* \right\|_{2,\nu} \leq \frac{\sqrt{C}}{1-\gamma} \left\| \widehat{f} - \mathcal{T}\widehat{f} \right\|_{2,\mu}.$$
The proof then follows by applying Lemmas 2 and 3 to the above equation. ■

From the above Lemma, we observe that it is sufficient to bound $\mathcal{L}_\mu(\widehat{f}, \widehat{f}) - \mathcal{L}_\mu(\mathcal{T}\widehat{f}, \widehat{f})$ to prove the main theorem.

C. USING CONCENTRATION INEQUALITY

Recall the definition of $Z_{\mathcal{F}}$ from Definition 1. It is straight forward that the class of functions $Z_{\mathcal{F}}$ is bounded, and using the bounds on $R(\cdot,\cdot)$ and $\mathcal{F}$, we can observe that $\forall f \in \mathcal{F}$, $|Z_f(\cdot,\cdot,\cdot,\cdot)| \leq 3V_{\max}^2$. For $f \in \mathcal{F}$ and $(s_i, a_i, r_i, s_i') \in D$, we can denote an i.i.d random variable $Z_f^i := Z_f(s_i, a_i, r_i, s_i')$. Now, we can observe from Lemma 3 that

$$\frac{1}{n}\sum_{i=1}^{n} Z_{\widehat{f}}^i = \mathcal{L}_D\left(\widehat{f}; \widehat{f}\right) - \mathcal{L}_D\left(\mathcal{T}\widehat{f}, \widehat{f}\right);$$
$$\mathbb{E}\left[Z_{\widehat{f}}^1\right] = \mathcal{L}_\mu\left(\widehat{f}; \widehat{f}\right) - \mathcal{L}_\mu\left(\mathcal{T}\widehat{f}; \widehat{f}\right) = \left\| \widehat{f} - \mathcal{T}\widehat{f} \right\|_{2,\mu}^2.$$

Reference [17] assume a finite function space $\mathcal{F}$ and proceed to bound $\mathbb{E}[Z_{\widehat{f}}^1] - \frac{1}{n}\sum_{i=1}^n Z_{\widehat{f}}^i$ using Bernstein's inequality and apply a union bound over the function space $\mathcal{F}$ (consequently the bound depends on $|\mathcal{F}|$). However, we can not do this for a general function space. Instead, we use the Pollard's concentration inequality (Lemma 11) to bound $\mathbb{E}[Z_{\widehat{f}}^1] - \frac{1}{n}\sum_{i=1}^n Z_{\widehat{f}}^i$.

Lemma 5: With probability at least $1 - \delta_1$, we have
$$\mathbb{E}\left[Z_{\widehat{f}}^1\right] \leq \frac{1}{n}\sum_{i=1}^{n} Z_{\widehat{f}}^i + \epsilon/8,$$
where $\delta_1 = 8\mathbb{E}[N_1(\frac{\epsilon}{64}, Z_{\mathcal{F}}, D^{1:n})] \exp\left(\frac{-n\epsilon^2}{64\times 128\times 36\, V_{\max}^4}\right)$.

Proof: We can directly apply Lemma 11 (along with Remark 4) on the class of functions $Z_{\mathcal{F}}$ with $B = 6V_{\max}^2$ and $\epsilon/8$

instead of ϵ. We get

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left(\mathbb{E}[Z_f^1]-\frac{1}{n}\sum_{i=1}^n Z_f^i\right)>\epsilon/8\right\}$$
$$\leq 8\mathbb{E}[N_1(\epsilon/64,Z_{\mathcal{F}},D^{1:n})]\exp\left(\frac{-n\epsilon^2}{64\times 128\times 36\,V_{\max}^4}\right)$$

Since the above inequality holds for all $f\in\mathcal{F}$, it certainly holds for a given $\hat{f}\in\mathcal{F}$. Note that, every $Z_f\in Z_{\mathcal{F}}$ is defined by a $f\in\mathcal{F}$. The result follows by taking the value of δ_1 equal to the right hand side of the above equation. ■

We now bounded $\mathcal{L}_\mu(\widehat{f},\widehat{f})-\mathcal{L}_\mu(\mathcal{T}\widehat{f},\widehat{f})$ in terms of $\mathcal{L}_D(\widehat{f};\widehat{f})-\mathcal{L}_D(\mathcal{T}\widehat{f},\widehat{f})$ by using Pollard's concentration inequality. In the next subsections, we will continue to bound $\mathcal{L}_D(\widehat{f};\widehat{f})-\mathcal{L}_D(\mathcal{T}\widehat{f},\widehat{f})$ using the Lemma 1 and repeated use of the Pollard's concentration inequality.

D. BOUNDING TERMS I, II, III

In this section, we bound the terms in Lemma 1.

Lemma 6 (Term I in (9)): With probability at least $1-\delta_1$, we have

$$\mathcal{L}_D(f^*;f^*)-\mathcal{L}_D(\mathcal{T}f^*;f^*)=\frac{1}{n}\sum_{i=1}^n Z_{f^*}^i\leq\frac{\epsilon}{8}+\epsilon_{\mathcal{F}}\quad(11)$$

where $\delta_1=8N_1(\frac{\epsilon}{64},Z_{\mathcal{F}},D^{1:n})\exp(\frac{-n\epsilon^2}{64\times128\times36\,V_{\max}^4})$.

Proof: From Lemma 3, $\mathbb{E}[Z_{f^*}^1]=\mathcal{L}_\mu(f^*;f^*)-\mathcal{L}_\mu(\mathcal{T}f^*;f^*)=\|f^*-\mathcal{T}f^*\|_{2,\mu}^2\leq\epsilon_{\mathcal{F}}$. Therefore, applying Lemma 11 (along with Remark 4) and since $\mathbb{E}[Z_{f^*}^1]\leq\epsilon_{\mathcal{F}}$,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n Z_{f^*}^i>\frac{\epsilon}{8}+\epsilon_{\mathcal{F}}\right\}$$
$$\leq 8\mathbb{E}\left[N_1\left(\frac{\epsilon}{64},Z_{\mathcal{F}},D^{1:n}\right)\right]\exp\left(\frac{-n\epsilon^2}{64\times128\times36\,V_{\max}^4}\right).$$

The result follows similarly by taking the value of δ_1 as the right hand side of the above equation. This bounds term (I) in (9). ■

We are left to compute bounds on the terms II, III. Observe that both the terms II, III in Lemma 1 are of the form $|\mathcal{L}_D(\mathcal{T}f,f)-\mathcal{L}_D(\widehat{\mathcal{T}}_{\mathcal{G}}f;f)|$ where term II considers $\widehat{f}$ and term III considers f^*. Therefore, in the following Lemma, we want to bound for any $f\in\mathcal{F}$,

$$|\mathcal{L}_D(\mathcal{T}f,f)-\mathcal{L}_D\left(\widehat{\mathcal{T}}_{\mathcal{G}}f;f\right)|.$$

This can then be used to bound each of the terms II and III. Before stating the Lemma, we first discuss about the function classes $X_{\mathcal{F}}$ and $Y_{\mathcal{F},\mathcal{G}}$.

Recall the definition of $X_{\mathcal{F}}$ from Definition 2. Similar to $Z_{\mathcal{F}}$, we can show that $X_{\mathcal{F}}$ is a bounded class of functions, and $\forall f,g\in\mathcal{F}$, $|X_{g,f,g_f^*}(\cdot,\cdot,\cdot,\cdot)|\leq 3V_{\max}^2$. For each $(s_i,a_i,r_i,s_i')\in D$, we denote an i.i.d random variable $X_{g,f,g_f^*}^i:=X_{g,f,g_f^*}(s_i,a_i,r_i,s_i')$. Now, from the definition of $\mathcal{L}_D$ and $\mathcal{L}_\mu$,

$$\frac{1}{n}\sum_i^n X_{g,f,g_f^*}^i=\mathcal{L}_D(g;f)-\mathcal{L}_D(g_f^*,f);$$
$$\mathbb{E}\left[X_{g,f,g_f^*}^1\right]=\mathcal{L}_\mu(g;f)-\mathcal{L}_\mu\left(g_f^*,f\right).$$

Recall the definition of $Y_{\mathcal{F},\mathcal{G}}$ from Definition 3. Similarly, we can again show that $Y_{\mathcal{F},\mathcal{G}}$ is a bounded class of functions, and $\forall f\in\mathcal{F},g\in\mathcal{G}$, $|Y_{g,f}(\cdot,\cdot,\cdot,\cdot)|\leq 3V_{\max}^2$. For each $(s_i,a_i,r_i,s_i')\in D$, we denote an i.i.d random variable $Y_{g,f}^i:=Y_{g,f}(s_i,a_i,r_i,s_i')$. Now, from the definition of $\mathcal{L}_D$ and $\mathcal{L}_\mu$,

$$\frac{1}{n}\sum_i^n Y_{g,f}^i=\mathcal{L}_D(g;f)-\mathcal{L}_D(\mathcal{T}f,f);$$
$$\mathbb{E}[Y_{g,f}^1]=\mathcal{L}_\mu(g;f)-\mathcal{L}_\mu(\mathcal{T}f,f).$$

We are now ready for the next Lemma.

Lemma 7 (Terms II, III in (9)): With probability at least $1-2\delta_2-\delta_3$, we have

$$|\mathcal{L}_D(\mathcal{T}f;f)-\mathcal{L}_D(\widehat{\mathcal{T}}_{\mathcal{G}}f;f)|\leq\epsilon_{\mathcal{F},\mathcal{G}}+\frac{3\epsilon}{8}$$

where $\delta_2=8\mathbb{E}[N_1\left(\frac{\epsilon}{64},X_{\mathcal{F}},D^{1:n}\right))]\exp\left(\frac{-n\epsilon^2}{64\times128\times36\,V_{\max}^4}\right)$ and $\delta_3=8\mathbb{E}[N_1\left(\frac{\epsilon}{64},Y_{\mathcal{F},\mathcal{G}},D^{1:n}\right))]\exp\left(\frac{-n\epsilon^2}{64\times128\times36\,V_{\max}^4}\right)$.

Proof: Observe that

$$|\mathcal{L}_D(\mathcal{T}f;f)-\mathcal{L}_D(\widehat{\mathcal{T}}_{\mathcal{G}}f;f)|$$
$$=\left|\frac{1}{n}\sum_{i=1}^n\left(X_{\widehat{\mathcal{T}}_{\mathcal{G}}f,f,g_f^*}^i+Y_{g_f^*,f}^i\right)\right|$$
$$\leq\left|\frac{1}{n}\sum_{i=1}^n X_{\widehat{\mathcal{T}}_{\mathcal{G}}f,f,g_f^*}^i\right|+\left|\sum_{i=1}^n Y_{g_f^*,f}^i\right|\quad(12)$$

Using Lemma 11 for the function space $X_{\mathcal{F}}$ with $B=6V_{\max}^2$ and $\epsilon/8$ instead of ϵ, we get

$$\mathbb{P}\left\{\sup_{f,g\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n X_{\widehat{\mathcal{T}}_{\mathcal{G}}f,f,g_f^*}^i-\mathbb{E}[X_{\widehat{\mathcal{T}}_{\mathcal{G}}f,f,g_f^*}^1]\right|>\frac{\epsilon}{8}\right\}$$
$$\leq 8\mathbb{E}[N_1(\epsilon/64,X_{\mathcal{F}},D^{1:n}))]\exp\left(\frac{-n\epsilon^2}{64\times128\times36\,V_{\max}^4}\right).$$

Observe that for $a,b\in\mathbb{R}$, $|a|-|b|\leq|a-b|$. From this (and using the same argument as in Remark 4), we get with probability greater than $1-\delta_2$,

$$\left|\frac{1}{n}\sum_{i=1}^n X_{\widehat{\mathcal{T}}_{\mathcal{G}}f,f,g_f^*}^i\right|\leq\left|\mathbb{E}[X_{\widehat{\mathcal{T}}_{\mathcal{G}}f,f,g_f^*}^1]\right|+\frac{\epsilon}{8}\quad(13)$$

where $\delta_2=8\mathbb{E}[N_1\left(\frac{\epsilon}{64},X_{\mathcal{F}},D^{1:n}\right))]\exp\left(\frac{-n\epsilon^2}{64\times128\times36\,V_{\max}^4}\right)$.

Now, observe that $\frac{1}{n}\sum_{i=1}^n X_{\widehat{\mathcal{T}}_{\mathcal{G}}f,f,g_f^*}^i\leq 0$, since $\frac{1}{n}\sum_{i=1}^n X_{\widehat{\mathcal{T}}_{\mathcal{G}}f,f,g_f^*}^i=\mathcal{L}_D(\widehat{\mathcal{T}}_{\mathcal{G}}f,f)-\mathcal{L}_D(g_f^*,f)\leq 0$ where the inequality follows due to the optimality of $\widehat{\mathcal{T}}_{\mathcal{G}}f$ given

dataset D. Therefore, applying Lemma 11 to the function space $X_{\mathcal{F}}$ again we get,

$$\left|\mathbb{E}\left[X^1_{\widehat{\mathcal{T}}_{\mathcal{G}} f, f, g^*_f}\right]\right| \leq \frac{\epsilon}{8} \tag{14}$$

with probability at least $1-\delta_2$.

Now by combining (13) and (14), we have with probability at least $1-2\delta_2$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} X^i_{\widehat{\mathcal{T}}_{\mathcal{G}} f, f, g^*_f}\right| \leq \epsilon/4$$

where $\delta_2 = 8\mathbb{E}[N_1(\frac{\epsilon}{64}, X_{\mathcal{F}}, D^{1:n}))]\exp\left(\frac{-n\epsilon^2}{64\times 128\times 36 V^4_{\max}}\right)$.

Now, we bound the second term $\left|\frac{1}{n}\sum_{i=1}^{n} Y^i_{g^*_f, f}\right|$ in (12). We have, $\mathbb{E}[Y^i_{g^*_f, f}] = \|g^*_f - \mathcal{T} f\|^2_{2,\mu} \leq \epsilon_{\mathcal{F},\mathcal{G}}$.

Applying Lemma 11 to the function space $Y_{\mathcal{F},\mathcal{G}}$ similarly as above, we have with probability at least $1-\delta_3$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} Y^i_{g^*_f, f}\right| \leq \mathbb{E}\left[Y^1_{g^*_f, f}\right] + \frac{\epsilon}{8} \leq \epsilon_{\mathcal{F},\mathcal{G}} + \frac{\epsilon}{8}$$

for $\delta_3 = 8\mathbb{E}[N_1(\frac{\epsilon}{64}, Y_{\mathcal{F},\mathcal{G}}, D^{1:n}))]\exp(\frac{-n\epsilon^2}{64\times 128\times 36 V^4_{\max}})$. The result follows by combining the bounds on $|\frac{1}{n}\sum_{i=1}^{n} X^i_{\widehat{\mathcal{T}}_{\mathcal{G}} f, f, g^*_f}|$ and $|\sum_{i=1}^{n} Y^i_{g^*_f, f}|$. ■

Lemma 8: With probability atleast $1-\delta_1-4\delta_2-2\delta_3$,

$$\mathcal{L}_D\left(\widehat{f}; \widehat{f}\right) - \mathcal{L}_D\left(\mathcal{T}\widehat{f}, \widehat{f}\right) \leq \epsilon_{\mathcal{F}} + \epsilon_{\mathcal{F},\mathcal{G}} + \frac{7\epsilon}{8} \tag{15}$$

where δ_1 is as defined in Lemma 6; δ_2 and δ_3 are defined as in Lemma 7.

Proof: Let us recall that, both the terms *II* and *III* can be bounded using Lemma 7. The result follows from Lemmas 1, 6 and 7. Note that we apply Lemma 7 twice, therefore the coefficients of δ_2 and δ_3 are multiplied by 2.

E. PROOF OF THEOREM 1

We now have all the intermediate results to prove the main theorem. From Lemmas 1, 5, and 8, with probability atleast $1-2\delta_1-4\delta_2-2\delta_3$,

$$\mathbb{E}\left[Z^1_{\widehat{f}}\right] \leq \epsilon_{\mathcal{F}} + \epsilon_{\mathcal{F},\mathcal{G}} + \epsilon.$$

Substituting this result in (10), with $K_1 = 64\times 128\times 36$, we get

$$\begin{aligned}
&\mathbb{P}\left\{v^* - v^{\pi_{\widehat{f}}} > \frac{2\sqrt{C}}{(1-\gamma)^2}\sqrt{\epsilon_{\mathcal{F}} + \epsilon_{\mathcal{F},\mathcal{G}} + \epsilon}\right\} \\
&\leq \exp\left(\frac{-n\epsilon^2}{K_1 V^4_{\max}}\right)\Big(16\mathbb{E}[N_1(\epsilon/64, Z_{\mathcal{F}}, D^{1:n})] \\
&\quad + 32\mathbb{E}[N_1(\epsilon/64, X_{\mathcal{F}}, D^{1:n})] \\
&\quad + 16\mathbb{E}[N_1(\epsilon/64, Y_{\mathcal{F},\mathcal{G}}, D^{1:n})]\Big).
\end{aligned}$$

We then apply Lemma 10 in the Appendix and let $K_2 = 6\times 64$, we get

$$\begin{aligned}
&\mathbb{P}\left\{v^* - v^{\pi_{\widehat{f}}} > \frac{2\sqrt{C}}{(1-\gamma)^2}\sqrt{\epsilon_{\mathcal{F}} + \epsilon_{\mathcal{F},\mathcal{G}} + \epsilon}\right\} \\
&\leq \exp\left(\frac{-n\epsilon^2}{K_1 V^4_{\max}}\right)\Bigg(16e(d_{Z_{\mathcal{F}}}+1)\left(\frac{K_2\, eV^2_{\max}}{\epsilon}\right)^{d_{Z_{\mathcal{F}}}} \\
&\quad + 32e(d_{X_{\mathcal{F}}}+1)\left(\frac{K_2\, eV^2_{\max}}{\epsilon}\right)^{d_{X_{\mathcal{F}}}} \\
&\quad + 16e(d_{Y_{\mathcal{F},\mathcal{G}}}+1)\left(\frac{K_2\, eV^2_{\max}}{\epsilon}\right)^{d_{Y_{\mathcal{F},\mathcal{G}}}}\Bigg).
\end{aligned}$$

Now consider the right hand side term to be δ, applying log and rearranging the terms, we get with probability at least $1-\delta$,

$$v^* - v^{\pi_{\widehat{f}}} \leq \frac{2\sqrt{C}}{(1-\gamma)^2}\left(\sqrt{\epsilon + \epsilon_{\mathcal{F}} + \epsilon_{\mathcal{F},\mathcal{G}}}\right)$$

when

$$\begin{aligned}
n \geq \frac{K_1 V^4_{\max}}{\epsilon^2}\Bigg[&\log\frac{16e}{\delta} + \log\Bigg(2(\mathrm{d}_{X_{\mathcal{F}}}+1)\left(\frac{K_2 eV^2_{\max}}{\epsilon}\right)^{\mathrm{d}_{X_{\mathcal{F}}}} \\
&+ (\mathrm{d}_{Y_{\mathcal{F},\mathcal{G}}}+1)\left(\frac{K_2 eV^2_{\max}}{\epsilon}\right)^{\mathrm{d}_{Y_{\mathcal{F},\mathcal{G}}}} \\
&+ (\mathrm{d}_{Z_{\mathcal{F}}}+1)\left(\frac{K_2 eV^2_{\max}}{\epsilon}\right)^{\mathrm{d}_{Z_{\mathcal{F}}}}\Bigg)\Bigg].
\end{aligned}$$

The proof is complete.

V. DISCUSSION

A. EXPERIMENTAL RESULTS

In this section, we perform numerical simulations to study how the sample complexity materializes in practice. We adopt the *optimal charging schedule for a battery pack* example from [25]. Here, the battery pack is used to serve some random user demands and is charged using a random renewable source. The maximum capacity of a battery pack is given by a real value $B \in \mathbb{R}^+$. The state of charge of the battery pack is denoted by $G_t \in [0, 1]$ which is the fraction of charge as compared to the maximum capacity. We represent the net generation at time t (renewable generation minus demand) as G_t. We assume that G_t is a bounded random variable, that is uniformly distributed between $[G_{\max}, G_{\min}]$. The state of the system is given by $s_t = [\mathrm{SoC}_t, G_{t-1}]^T$, and the action is given by a_t, which determines what fraction of the net demand (generation) is served. At every time step, the state $[\mathrm{SoC}_t, G_{t-1}$ is observed and an action a_t is taken. When $G_{t-1} > 0$, since there is net generation, the battery charge is increased by $a_t G_{t-1}$. When $G_{t-1} \leq 0$, since there is net demand, the battery charge decreases by $a_t G_{t-1}$. Note that, even when $G_{t-1} > 0$, a_t need not be very high, since the battery can get damaged due to overheating. At every time step, the battery also self discharges determined by some parameter

$\beta \in [0, 1]$. The state update equation is given by,

$$\texttt{SoC}_{t+1} := \min\left\{\beta \texttt{SoC}_t + a_t \frac{G_{t-1}}{B}, 1\right\}.$$

For every action taken, the reward is specified by

$$r(s_t, a_t) = a_t\left(\tanh(\xi G_{t-1})\frac{(r2 - r1)}{2} + \frac{(r_1 + r_2)}{2}\right),$$

where $\xi > 0$ is a scale parameter, r_1 is the reward of using renewable energy, r_2 is the utility of serving the user demand and $0 \leq r_1 < r_2$. Also, observe that the action space is constrained by the current state. Here, it depends on the current state of charge, i.e., a_t needs to be chosen such that $\texttt{SoC}_t \geq 0$ since more charge can not be extracted from the battery than what is present. Formally,

$$\Gamma(s_t) = \left\{a_t \in [0, 1] : \beta \texttt{SoC}_t + a_t \frac{G_{t-1}}{B} \geq 0\right\}.$$

The goal of the RL problem is to maximize the reward until the battery doesn't run out of charge.

1) OFFLINE DATASET

Here, we outline the method used to collect the offline dataset for this environment. We train a policy in the online setting using the TD3 algorithm [37]. We used the following parameters of the environment: $G_t \sim \text{Unif}(-10, 10)$, $B = 10$, $\xi = 0.01$, $r_1 = 5$, $r_2 = 15$, $\beta = 0.97$. For online training, we used $\gamma = 0.9$ and used the same parameters as the original TD3 paper. We used the same function approximation class for $\mathcal{F}$ and $\mathcal{G}$: we used a neural network with two hidden layers each of width 4 and `ReLU` activation functions.

We use the output of the above algorithm and deploy it in the environment and log the interactions (s, a, r, s'). We now, use the logged dataset to solve (3).

2) ALGORITHM

To solve the min max optimization in (3), we perform a bilevel optimization routine by alternating the updates on f, g.

This is written as

$$g_t = \arg\min_{g \in \mathcal{G}} \mathcal{L}_D(g, f_{t-1})$$

$$f_t = \arg\min_{f \in \mathcal{F}} \mathcal{L}_D(f, f) - \mathcal{L}_D(g_t, f)$$

At each iteration t, we sample a mini-batch of size 256 from the offline dataset D. We then use the samples to update the neural network weights (of g_t and f_t) using Stochastic Gradient Descent (SGD) algorithm with a learning rate of `1e-5`. We iterate until $t = 10^5$.

3) RESULTS

To show the impact of the dataset size, we take different dataset sizes $\in [100, 1000, 10000]$ and train three different algorithms using these datasets. The loss trajectory for the n=10000 case is presented in Fig. 1. Since we do not have the true V^*, it is difficult to compute v^* for this environment.

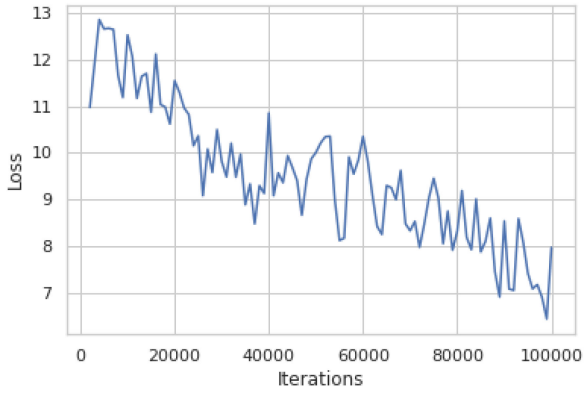

FIGURE 1. We plot the loss $\mathcal{L}_D(f_t; f_t) - \mathcal{L}_D(g_t; f_t)$ at every iteration t of the bi-level optimization algorithm.

TABLE 2 Comparison of Sub-Optimality With the Number of Data Points n. Note That π_b is Fixed for All Three Cases.

n	$v^{\pi_{\hat{f}}}$	v^{π_b}
100	4.3	2.9
1000	5.0	2.9
10000	5.7	2.9

Therefore, instead of measuring $v^* - v^{\pi_{\hat{f}}}$, we simply measure $v^{\pi_{\hat{f}}}$ where $\pi_{\hat{f}}$ is the policy greedy with respect to the output of (3). To compute $v^{\pi_{\hat{f}}}$, we used the initial state distribution, where $G_t \sim \text{Unif}(-10, 10)$ and $\texttt{SoC}_0 = 1.0$. We present the results of the algorithm in Table 2. We can observe that $v^* - v^{\pi_{\hat{f}}}$ reduces as the number of data points in the offline dataset increases which is inline with the theoretical results.

B. SHARP CONCENTRATION RESULTS

While earlier works [11], [14] obtain a dependence on the number of samples as $\mathcal{O}(\frac{1}{\epsilon^4})$, recent works were able to improve on the sample complexity to $\mathcal{O}(\frac{1}{\epsilon^2})$ [17], [18], [19]. The key result that leads to an improvement in the sample complexity is the sharper concentration inequality (stated below) as compared to using the Pollard's concentration inequality (Lemma 11).

Lemma 9 (Lemma 11.6 [38]): Let $B \geq 1$ and let $\mathcal{G}$ be a set of functions $g : \mathbb{R}^d \to [0, B]$. Let $Z, Z_1, \ldots, Z_n$ be i.i.d $\mathbb{R}^d$ valued random variables. Assume $\epsilon > 0$, $0 < \alpha < 1$ and $n \geq 1$. Then

$$\mathbb{P}\left\{\sup_{g \in \mathcal{G}} \frac{\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)]}{\epsilon + \frac{1}{n}\sum_{i=1}^n g(Z_i) + \mathbb{E}[g(Z)]} > \alpha\right\}$$

$$\leq 4\mathbb{E}\left[N_1\left(\frac{\alpha\epsilon}{5}, \mathcal{G}, Z_1^n\right)\right]\exp\left(-\frac{3\epsilon\alpha^2 n}{40B}\right).$$

Reference [18], [19] decompose the error $v^* - v^{\pi_{\hat{f}}}$ in a way that exploits this sharp concentration result (notice the exponent of ϵ in exp term as compared to Pollard's theorem reviewed in Theorem 11), thereby deriving a sharper sample complexity. Decomposing the error for the min-max variant

studied in this paper to apply this result is challenging and we leave it as future work.

C. SINGLE SAMPLE PATH

In this work, we assumed that the offline data available with the agent is sampled independently from a distribution $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$. This assumption can be practical in several applications where the agent is highly scalable and deployed in a large number of environments to collect data. For example, in e-commerce/web-applications, millions of users may query in parallel independent of one another and the deployed agent also responds independently of the other queries. However, there are several applications where a single controller/agent collects a single (long) trajectory of data in applications such as building energy management. Here the deployed agent is highly customized to the particular environment (e.g., building, factory) because each environment has different dynamics than the others. In these situations, one needs to deal with a long trajectory of data where the (s, a) pairs are no longer independent. [14] studies such a setting by assuming some mixing properties on the process. The sample complexity itself remains the same, however the bound includes an additional term dependent on the mixing coefficient of the process.

D. REMOVING CONCENTRABILITY ASSUMPTION

In this work, we consider the stricter condition (Assumption 2) of concentrability where every admissible $\nu \in \Delta(\mathcal{B})$ is absolutely continuous with respect to μ. A more general condition is the discounted concentrability of future state-action distributions where the distribution of future state-action pairs is assumed to be absolutely continuous with respect to μ. Let P^π be an operator acting on $f : \mathcal{B} \to \mathbb{R}$ s.t. $(P^\pi f)(s, a) = \int_{\mathcal{B}} f(s', \pi(s'))P(ds'|s, a)$. For $m \geq 0$, and any arbitrary sequence of stationary policies $\{\pi_1, \ldots, \pi_m\}$

$$C_\mu(m) = \sup_{\pi_1,\ldots,\pi_m} \left\| \frac{d\left(\eta_{\text{init}} P^{\pi_1} P^{\pi_2} \cdots P^{\pi_m}\right)}{d\mu} \right\|_\infty .$$

The discounted concentrability assumption requires $C_\mu = (1-\gamma^2)\sum_{m\geq 1} m\gamma^{m-1} C_\mu(m) < \infty$. The sample complexity analysis under discounted concentrability assumption requires substantially different arguments, and thus is left as a future work.

VI. CONCLUSION

In this work, we studied the finite sample analysis of the min-max formulation of the ORL algorithm for general state and function spaces under the state dependent action constraints. The sample complexity of the algorithm depends as $\mathcal{O}(1/\epsilon^4)$.

More sophisticated concentration inequalities can be utilized to further sharpen the sample complexity and will be considered as future work. The data is assumed to be drawn according to a distribution μ, however in some practical scenarios, the data is available as a single trajectory. Another direction of this work can be to extend the analysis to the single sample path case by using the $\beta-$mixing properties of the data distribution.

APPENDIX A
PROOF OF LEMMA 1

By the optimality of $\widehat{f}$, we can write

$$\mathcal{L}_D(\widehat{f}; \widehat{f}) - \mathcal{L}_D(\widehat{\mathcal{T}}_{\mathcal{G}}\widehat{f}, \widehat{f}) \leq \mathcal{L}_D(f^*; f^*) - \mathcal{L}_D(\widehat{\mathcal{T}}_{\mathcal{G}} f^*; f^*)$$

Adding and subtracting $\mathcal{L}_D(\mathcal{T}\widehat{f}, \widehat{f})$ on LHS and $\mathcal{L}_D(\mathcal{T}f^*, f^*)$ on RHS, we get

$$\begin{aligned}\mathcal{L}_D(\widehat{f}; \widehat{f}) - \mathcal{L}_D(\mathcal{T}\widehat{f}, \widehat{f}) + \mathcal{L}_D(\mathcal{T}\widehat{f}, \widehat{f}) - \mathcal{L}_D(\widehat{\mathcal{T}}_{\mathcal{G}}\widehat{f}, \widehat{f}) \\ \leq \mathcal{L}_D(f^*; f^*) - \mathcal{L}_D(\mathcal{T}f^*, f^*) \\ + \mathcal{L}_D(\mathcal{T}f^*, f^*) - \mathcal{L}_D(\widehat{\mathcal{T}}_{\mathcal{G}} f^*; f^*)\end{aligned}$$

The result is obtained by rearranging the terms.

APPENDIX B
CAPACITY OF FUNCTION CLASSES AND RESULTS FROM EMPIRICAL PROCESS THEORY

*Definition 4 (Covering Number):*Let $(\mathcal{M}, d)$ be a pseudo-metric space, and let $\epsilon > 0$. Let $M_1, \ldots, M_k$ be balls of radius $\epsilon > 0$ in $\mathcal{M}$. We say that $\{M_i\}_{i=1}^k$ is a covering of $(\mathcal{M}, d)$ if $\mathcal{M} \subseteq \cup_{i=1}^k M_i$. The covering number $N(\epsilon, \mathcal{M}, d)$ is defined as the smallest k such that the set of ϵ balls $\{M\}_1^k$ is a covering of $(\mathcal{M}, d)$. If no such finite k exists, then the covering number is ∞.

Definition 5 (Empirical Covering Number): Let $\mathcal{H}$ be a class of functions with domain $\mathcal{R}$, and let points $R^{1:n} := (R_1, \ldots, R_n)$ be points in $\mathcal{R}$. The empirical covering number is defined with respect to the pseudo metric $l_{R^{1:N}}(f, g) = \frac{1}{N}\sum_{i=1}^N |f(R_i) - g(R_i)|; \quad g \in \mathcal{F}$ and denoted by $N_1(\epsilon, \mathcal{F}, R^{1:N})$.

Definition 6 (VC dimension [39]): Let $\mathcal{H}$ denote a class of functions from $\mathcal{X} \to \{0, 1\}$. The growth function is defined as, for any non-negative m,

$$s(\mathcal{H}, m) := \max_{x_1,\ldots,x_m \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}\}|.$$

If for any $\{x_1, \ldots, x_m\}$, $|\{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}\}| = 2^m$, we say $\mathcal{H}$ shatters the set $\{x_1, \ldots, x_m\}$. The VC dimension of $\mathcal{H}$ is defined as the largest number of points m that it can shatter, i.e.,

$$\texttt{VC-dim}(\mathcal{H}) := \sup\{m \in \mathbb{N} : s(\mathcal{H}, m) = 2^m\}.$$

For a real valued function class, the capacity is defined in terms of the pseudo-dimension.

Definition 7 (pseudo-dimension [39]): Let $\mathcal{F} : \{f : \mathcal{X} \to \mathbb{R}\}$. The pseudo-dimension $\mathtt{d}_{\mathcal{F}}$ is defined as the largest integer m for which there exists $(x_1, \ldots, x_m, y_1, \ldots, y_m) \in \mathcal{X}^m \times \mathbb{R}^m$ such that for any $(b_1, \ldots, b_m) \in \{0, 1\}^m$ there exists $f \in \mathcal{F}$ such that $\forall i : f(x_i) > y_i \iff b_i = 1$. For cases where the function class $\mathcal{F}$ is generated by a neural network with a fixed architecture and activation function, we can also write $d_{\mathcal{F}} = \texttt{VC-dim}(\text{sign}(\mathcal{F}))$, where $\text{sign}(\mathcal{F}) = \{\text{sign}(f) : f \in \mathcal{F}\}$ and $\text{sign}(x) = 1$ if $x \geq 0$ and $\text{sign(x)} = 1$ if $x < 0$.

The following lemma relates the pseudo dimension of a function class to the empirical covering number.

Lemma 10 (see [40]): Consider a set Z and a class $\mathcal{F} \subset \{f : Z \to [0, \bar{C}]\}$ of functions on Z with pseudo-dimension $\mathtt{d}_{\mathcal{F}} < \infty$. For any points $z^{1:n} \in Z^n$ and $\epsilon > 0$,

$$N_1\left(\epsilon, \mathcal{F}, z^{1:n}\right) \leq e\left(\mathtt{d}_{\mathcal{F}} + 1\right)\left(\frac{2e\bar{C}}{\epsilon}\right)^{\mathtt{d}_{\mathcal{F}}}.$$

Lemma 11 (Pollard's tail inequality: Theorem 9.1 [38]): Let $\mathcal{H}$ be a class of functions that map $\mathcal{R}$ into $[-B/2, B/2]$, and let μ be a probability measure on $\mathcal{R}$. Let $R_1, \ldots, R_n$ are i.i.d with distribution μ. For every $\epsilon > 0$,

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} h(R_i) - \mathbb{E}[h(R)]\right| > \epsilon\right\} \leq 8\mathbb{E}\left[N_1\left(\epsilon/8, \mathcal{H}, R^{1:n}\right)\right]\exp\left(-\frac{n\epsilon^2}{128B^2}\right)$$

where $N_1(\cdot, \mathcal{H}, R^{1:n})$ is the empirical covering number of $\mathcal{H}$ given data points $R^{1:n}$.

Remark 4: Note that, during the proofs, we often consider

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n} h(R_i) - \mathbb{E}[h(R)]\right) > \epsilon\right\}$$

instead of the absolute value. Observe that,

$$\sup_{h\in\mathcal{H}}\left(\sum_{i=1}^{n}\frac{h(R_i) - \mathbb{E}[h(R)]}{n}\right) \leq \sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\frac{h(R_i) - \mathbb{E}[h(R)]}{n}\right|$$

and this shows that

$$\left\{\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n} h(R_i) - \mathbb{E}[h(R)]\right) > \epsilon\right\} \subseteq \left\{\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} h(R_i) - \mathbb{E}[h(R)]\right| > \epsilon\right\}.$$

Therefore, we simply use

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n} h(R_i) - \mathbb{E}[h(R)]\right) > \epsilon\right\} \leq 8\mathbb{E}\left[N_1(\epsilon/8, \mathcal{H}, R^{1:n})\right]\exp\left(-\frac{n\epsilon^2}{128B^2}\right).$$

APPENDIX C
DIFFERENCE BETWEEN $\mathcal{L}_\mu(f; f)$ AND $\|f - \mathcal{T}f\|^2_{2,\mu}$

In this section, we show that $\mathcal{L}_\mu(f; f)$ overestimates $\|f - \mathcal{T}f\|^2_{2,\mu}$ by a variance term. First, we have

$$\begin{aligned}
\mathcal{L}_\mu(f, f) &= \mathbb{E}[\mathcal{L}_D(f, f)] \\
&= \mathbb{E}\left[f(s,a) - R(s,a) - \gamma V_f(s')\right]^2 \\
&= \mathbb{E}\Big[(f(s,a) - R(s,a))^2 + \gamma^2 V_f^2(s') \\
&\quad -2\gamma f(s,a)V_f(s') + 2\gamma R(s,a)V_f(s')\Big] \\
&= \mathbb{E}_\mu\left[(f(s,a) - R(s,a))^2\right] + \gamma^2\mathbb{E}[V_f^2(s')] \\
&\quad - 2\gamma\mathbb{E}[f(s,a)V_f(s')] + 2\gamma\mathbb{E}[R(s,a)V_f(s')].
\end{aligned}$$

Further, we note that $\|f - \mathcal{T}f\|^2_{2,\mu}$ is expanded as

$$\begin{aligned}
&\|f - \mathcal{T}f\|^2_{2,\mu} \\
&= \int \left(f(s,a) - R(s,a) - \gamma\mathbb{E}_{s'\sim P(s,a)}[V_f(s')]\right)^2 d\mu \\
&= \int (f(s,a) - R(s,a))^2 d\mu \\
&\quad + \gamma^2\int \left(\mathbb{E}_{s'\sim P(s,a)}[V_f(s')]\right)^2 d\mu \\
&\quad - 2\gamma\int f(s,a)\,\mathbb{E}_{s'\sim P(s,a)}[V_f(s')]\,d\mu \\
&\quad + 2\gamma\int R(s,a)\,\mathbb{E}_{s'\sim P(s,a)}[V_f(s')]d\mu \\
&= \int (f(s,a) - R(s,a))^2 d\mu \\
&\quad + \gamma^2\int \left(\mathbb{E}_{s'\sim P(s,a)}[V_f(s')]\right)^2 d\mu \\
&\quad - 2\gamma\mathbb{E}[f(s,a)\,V_f(s')] + 2\gamma\mathbb{E}[R(s,a)\,V_f(s')].
\end{aligned}$$

The two equations above yield the desired results.

ACKNOWLEDGMENT

The authors would like to thank Shobhit Gupta, Yuntian Deng, and Shiping Shao for useful discussions regarding the work.

REFERENCES

[1] V. Mnih et al., "Playing atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.
[2] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[3] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 5026–5033.
[4] "Opening up a physics simulator for robotics." Accessed: Oct. 10, 2021. [Online]. Available: https://www.deepmind.com/blog/opening-up-a-physics-simulator-for-robotics
[5] S. Desai, H. Karnan, J. P. Hanna, G. Warnell, and a. P. Stone, "Stochastic grounded action transformation for robot learning in simulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 6106–6111.
[6] A. Afzal, D. S. Katz, C. L. Goues, and C. S. Timperley, "A study on the challenges of using robotics simulators for testing," 2020, *arXiv:2004.07368*.
[7] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020, *arXiv:2005.01643*.
[8] S. Lange, T. Gabel, and M. Riedmiller, "Batch reinforcement learning," in *Reinforcement Learning*. Berlin, Germany: Springer, 2012, pp. 45–73.
[9] G. J. Gordon, "Approximate solutions to Markov decision processes," Ph.D. dissertation, Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 1999. [Online]. Available: https://www.proquest.com/dissertations-theses/approximate-solutions-markov-decision-processes/docview/304499958/se-2
[10] C. Szepesvári and R. Munos, "Finite time bounds for sampling based fitted value iteration," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 880–887.

[11] R. Munos and C. Szepesvári, "Finite-time bounds for fitted value iteration," *J. Mach. Learn. Res.*, vol. 9 no. 27, pp. 815–857, 2008.
[12] D. Ernst, P. Geurts, and L. Wehenkel, "Tree-based batch mode reinforcement learning," *J. Mach. Learn. Res.*, vol. 6, pp. 503–556, 2005.
[13] M. Riedmiller, "Neural fitted Q iteration-first experiences with a data efficient neural reinforcement learning method," in *Proc. Eur. Conf. Mach. Learn.*, Springer, 2005, pp. 317–328.
[14] A. Antos, C. Szepesvári, and R. Munos, "Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path," *Mach. Learn.*, vol. 71, no. 1, pp. 89–129, 2008.
[15] A. Lazaric, M. Ghavamzadeh, and R. Munos, "Finite-sample analysis of least-squares policy iteration," *J. Mach. Learn. Res.*, vol. 13, pp. 3041–3074, 2012.
[16] A. M. Farahmand, R. Munos, and C. Szepesvári, "Error propagation for approximate policy and value iteration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 568–576.
[17] J. Chen and N. Jiang, "Information-theoretic considerations in batch reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1042–1051.
[18] H. Le, C. Voloshin, and Y. Yue, "Batch policy learning under constraints," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3703–3712.
[19] T. Nguyen-Tang et al., "Sample complexity of offline reinforcement learning with deep ReLU networks," 2021, *arXiv:2103.06671*.
[20] M. J. Khojasteh, V. Dhiman, M. Franceschetti, and N. Atanasov, "Probabilistic safety constraints for learned high relative degree system dynamics," in *Proc. Learn. Dyn. Control*, 2020, pp. 781–792.
[21] D. D. Fan, J. Nguyen, R. Thakker, N. Alatur, A.-a. Agha-mohammadi, and E. A. Theodorou, "Bayesian learning-based adaptive control for safety critical systems," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 4093–4099.
[22] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 3387–3395.
[23] R. Cheng, M. J. Khojasteh, A. D. Ames, and J. W. Burdick, "Safe multi-agent interaction through robust control barrier functions with learned uncertainties," in *Proc. 59th IEEE Conf. Decis. Control*, 2020, pp. 777–783.
[24] O. Shelke, V. Baniwal, and H. Khadilkar, "Anticipatory decisions in retail e-commerce warehouses using reinforcement learning," in *Proc. 8th ACM IKDD CODS 26th COMAD*, 2021, pp. 272–280.
[25] H. Li, S. Shao, and A. Gupta, "Fitted value iteration in continuous MDPs with state dependent action sets," *IEEE Contr. Syst. Lett.*, vol. 6, pp. 1310–1315, 2022.
[26] Z. Zhu, S. Gupta, A. Gupta, and M. Canova, "A deep reinforcement learning framework for ECO-driving in connected and automated hybrid electric vehicles," 2021, *arXiv:2101.05372*.
[27] Z. Zhu, N. Pivaro, S. Gupta, A. Gupta, and M. Canova, "Safe model-based off-policy reinforcement learning for ECO-driving in connected and automated hybrid electric vehicles," *IEEE Trans. Intell. Veh.*, vol. 7, no. 2, pp. 387–398, Jun. 2022.
[28] S. R. Deshpande et al., "In-vehicle test results for advanced propulsion and vehicle system controls using connected and automated vehicle information," *SAE Int. J. Adv. Curr. Practices Mobility*, vol. 3, no. 2021–01-0430, pp. 2915–2930, 2021.
[29] S. R. Deshpande, S. Gupta, A. Gupta, and M. Canova, "Real-time ecodriving control in electrified connected and autonomous vehicles using approximate dynamic programing," *J. Dyn. Syst., Meas., Control*, vol. 144, no. 1, 2022, Art. no. 011111.
[30] M. Guériau and I. Dusparic, "SAMoD: Shared autonomous mobility-on-demand using decentralized reinforcement learning," in *Proc. 21st Int. Conf. Intell. Transp. Syst.*, 2018, pp. 1558–1563.
[31] Y. Deng, H. Chen, S. Shao, J. Tang, J. Pi, and A. Gupta, "Multi-objective vehicle rebalancing for ridehailing system using a reinforcement learning approach," *J. Manage. Sci. Eng.*, vol. 7, pp. 346–364, 2022.
[32] Léon Bottou et al., "Counterfactual reasoning and learning systems: The example of computational advertising," *J. Mach. Learn. Res.*, vol. 14 no. 1, pp. 3207–3260, 2013.
[33] J. R. Regatti, A. A. Deshmukh, F. Cheng, Y. H. Jung, A. Gupta, and U. Dogan, "Offline RL with resource constrained online deployment," 2021, *arXiv:2110.03165*.
[34] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, vol. 30. Berlin, Germany: Springer, 2012.
[35] K. Hinderer, "Lipschitz continuity of value functions in Markovian decision processes," *Math. Methods Operations Res.*, vol. 62, no. 1, pp. 3–22, 2005.
[36] W. A. Vaart and J. A. Wellner, "Weak convergence," in *Weak Convergence and Empirical Processes*. Berlin, Germany: Springer, 1996, pp. 16–28.
[37] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
[38] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. Berlin, Germany: Springer, 2006.
[39] L. Peter, N. Bartlett, C. H.Liaw, and A. Mehrabian, "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 2285–2301, 2019.
[40] D. Haussler, "Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension," *J. Combinatorial Theory, Ser. A*, vol. 69, no. 2, pp. 217–232, 1995.