Robust Reinforcement Learning using Offline Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

The goal of robust reinforcement learning (RL) is to learn a policy that is robust against the uncertainty in model parameters. Parameter uncertainty commonly occurs in many real-world RL applications due to simulator modeling errors, changes in the real-world system dynamics over time, and adversarial disturbances. Robust RL is typically formulated as a max-min problem, where the objective is to learn the policy that maximizes the value against the worst possible models that lie in an uncertainty set. In this work, we propose a robust RL algorithm called Robust Fitted Q-Iteration (RFQI), which uses only an offline dataset to learn the optimal robust policy. Robust RL with offline data is significantly more challenging than its non-robust counterpart because of the minimization over all models present in the robust Bellman operator. This poses challenges in offline data collection, optimization over the models, and unbiased estimation. In this work, we propose a systematic approach to overcome these challenges, resulting in our RFQI algorithm. We prove that RFQI learns a near-optimal robust policy under standard assumptions and demonstrate its superior performance on standard benchmark problems.

1 Introduction

Reinforcement learning (RL) algorithms often require a large number of data samples to learn a control policy. As a result, training them directly on the real-world systems is expensive and potentially dangerous. This problem is typically overcome by training them on a simulator (online RL) or using a pre-collected offline dataset (offline RL). The offline dataset is usually collected either from a sophisticated simulator of the real-world system or from the historical measurements. The trained RL policy is then deployed assuming that the training environment, the simulator or the offline data, faithfully represents the model of the real-world system. This assumption is often incorrect due to multiple factors such as the approximation errors incurred while modeling, changes in the real-world parameters over time and possible adversarial disturbances in the real-world. For example, the standard simulator settings of the sensor noise, action delay, friction, and mass of a mobile robot can be different from that of the actual real-world robot, in addition to changes in the terrain, weather conditions, lighting, and obstacle densities of the testing environment. Unfortunately, the current RL control policies can fail dramatically when faced with even mild changes in the training and testing environments (Sünderhauf et al., 2018; Tobin et al., 2017; Peng et al., 2018).

The goal in robust RL is to learn a policy that is robust against the model parameter mismatches between the training and testing environments. The robust planning problem is formalized using the framework of Robust Markov Decision Process (RMDP) (Iyengar, 2005; Nilim and El Ghaoui, 2005). Unlike the standard MDP which considers a single model (transition probability function), the RMDP formulation considers a set of models which is called the *uncertainty set*. The goal is to find

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

an optimal robust policy that performs the best under the worst possible model in this uncertainty set. The minimization over the uncertainty set makes the robust MDP and robust RL problems significantly more challenging than their non-robust counterparts.

In this work, we study the problem of developing a robust RL algorithm with provably optimal performance for an RMDP with arbitrarily large state spaces, using only offline data with function approximation. Before stating the contributions of our work, we provide a brief overview of the results in offline and robust RL that are directly related to ours. We leave a more thorough discussion on related works to Appendix D.

Offline RL: Offline RL considers the problem of learning the optimal policy only using a pre-collected (offline) dataset. Offline RL problem has been addressed extensively in the literature (Antos et al., 2008; Bertsekas, 2011; Lange et al., 2012; Chen and Jiang, 2019; Xie and Jiang, 2020; Levine et al., 2020; Xie et al., 2021). Many recent works develop deep RL algorithms and heuristics for the offline RL problem, focusing on the algorithmic and empirical aspects (Fujimoto et al., 2019; Kumar et al., 2019, 2020; Yu et al., 2020; Zhang and Jiang, 2021). A number of theoretical work focus on analyzing the variations of Fitted Q-Iteration (FQI) algorithm (Gordon, 1995; Ernst et al., 2005), by identifying the necessary and sufficient conditions for the learned policy to be approximately optimal and characterizing the performance in terms of sample complexity (Munos and Szepesvári, 2008; Farahmand et al., 2010; Lazaric et al., 2012; Chen and Jiang, 2019; Liu et al., 2020; Xie et al., 2021). All these works assume that the offline data is generated according to a single model and the goal is to find the optimal policy for the MDP with the same model. In particular, none of these works consider the *offline robust RL problem* where the offline data is generated according to a (training) model which can be different from the one in testing, and the goal is to learn a policy that is robust w.r.t. an uncertainty set.

Robust RL: The RMDP framework was first introduced in Iyengar (2005); Nilim and El Ghaoui (2005). The RMDP problem has been analyzed extensively in the literature (Xu and Mannor, 2010; Wiesemann et al., 2013; Yu and Xu, 2015; Mannor et al., 2016; Russel and Petrik, 2019) providing computationally efficient algorithms, but these works are limited to the planning problem. Robust RL algorithms with provable guarantees have also been proposed (Lim et al., 2013; Tamar et al., 2014; Roy et al., 2017; Panaganti and Kalathil, 2021; Wang and Zou, 2021), but they are limited to tabular or linear function approximation settings and only provide asymptotic convergence guarantees. Robust RL problem has also been addressed using deep RL methods (Pinto et al., 2017; Derman et al., 2018, 2020; Mankowitz et al., 2020; Zhang et al., 2020a). However, these works do not provide any theoretical guarantees on the performance of the learned policies.

The works that are closest to ours are by Zhou et al. (2021); Yang et al. (2021); Panaganti and Kalathil (2022) that address the robust RL problem in a tabular setting under the generative model assumption. Due to the generative model assumption, the offline data has the same uniform number of samples corresponding to each and every state-action pair, and tabular setting allows the estimation of the uncertainty set followed by solving the planning problem. Our work is significantly different from these in the following way: (i) we consider a robust RL problem with arbitrary large state space, instead of the small tabular setting, (ii) we consider a true offline RL setting where the state-action pairs are sampled according to an arbitrary distribution, instead of using the generative model assumption, (iii) we focus on a function approximation approach where the goal is to directly learn optimal robust value/policy using function approximation techniques, instead of solving the tabular planning problem with the estimated model. *To the best of our knowledge, this is the first work that addresses the offline robust RL problem with arbitrary large state space using function approximation, with provable guarantees on the performance of the learned policy.*

Offline Robust RL: Challenges and Our Contributions: Offline robust RL is significantly more challenging than its non-robust counterpart mainly because of the following key difficulties.
(i) Data generation: The optimal robust policy is computed by taking the infimum over all models in the uncertainty set $\mathcal{P}$. However, generating data according to all models in $\mathcal{P}$ is clearly infeasible. It may only be possible to get the data from a nominal (training) model P^o. *How do we use the data from a nominal model to account for the behavior of all the models in the uncertainty set $\mathcal{P}$?*
(ii) Optimization over the uncertainty set $\mathcal{P}$: The robust Bellman operator (defined in (3)) involves a

minimization over $\mathcal{P}$, which is a significant computational challenge. Moreover, the uncertainty set $\mathcal{P}$ itself is unknown in the RL setting. *How do we solve the optimization over $\mathcal{P}$?*
(iii) Function approximation: Approximation of the robust Bellman update requires a modified target function which also depends on the approximate solution of the optimization over the uncertainty set. *How do we perform the offline RL update accounting for both approximations?*

As the *key technical contributions* of this work, we first derive a dual reformulation of the robust Bellman operator which replaces the expectation w.r.t. all models in the uncertainty set $\mathcal{P}$ with an expectation only w.r.t. the nominal (training) model P^o. This enables using the offline data generated by P^o for learning, without relying on high variance importance sampling techniques to account for all models in $\mathcal{P}$. Following the same reformulation, we then show that the optimization problem over $\mathcal{P}$ can be further reformulated as functional optimization. We solve this functional optimization problem using empirical risk minimization and obtain performance guarantees using the Rademacher complexity based bounds. We then use the approximate solution obtained from the empirical risk minimization to generate modified target samples that are then used to approximate robust Bellman update through a generalized least squares approach with provably bounded errors. Performing these operations iteratively results in our proposed Robust Fitted Q-Iteration (RFQI) algorithm, for which we prove that its learned policy achieves non-asymptotic and approximately optimal performance guarantees.

Notations: For a set $\mathcal{X}$, we denote its cardinality as $|\mathcal{X}|$. The set of probability distribution over $\mathcal{X}$ is denoted as $\Delta(\mathcal{X})$, and its power set sigma algebra as $\Sigma(\mathcal{X})$. For any $x \in \mathbb{R}$, we denote $\max\{x, 0\}$ as $(x)_+$. For any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, state-action distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$, and real number $p \geq 1$, the ν-weighted p-norm of f is defined as $\|f\|_{p,\nu} = \mathbb{E}_{s,a\sim\nu}[|f(s,a)|^p]^{1/p}$.

2 Preliminaries

A Markov Decision Process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, r, P, \gamma, d_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\gamma \in (0, 1)$ is the discount factor, and $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution. The transition probability function $P_{s,a}(s')$ is the probability of transitioning to state s' when action a is taken at state s. In the literature, P is also called the *model* of the MDP. We consider a setting where $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite but can be arbitrarily large. We will also assume that $r(s, a) \in [0, 1]$, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, without loss of generality. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a conditional distribution over actions given a state. The value function $V_{\pi,P}$ and the state-action value function $Q_{\pi,P}$ of a policy π for an MDP with model P are defined as

$$V_{\pi,P}(s) = \mathbb{E}_{\pi,P}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s], \quad Q_{\pi,P}(s, a) = \mathbb{E}_{\pi,P}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a],$$

where the expectation is over the randomness induced by the policy π and model P. The optimal value function V_P^* and the optimal policy π_P^* of an MDP with the model P are defined as $V_P^* = \max_\pi V_{\pi,P}$ and $\pi_P^* = \arg\max_\pi V_{\pi,P}$. The optimal state-action value function is given by $Q_P^* = \max_\pi Q_{\pi,P}$. The optimal policy can be obtained as $\pi_P^*(s) = \arg\max_a Q_P^*(s, a)$. The discounted state-action occupancy of a policy π for an MDP with model P, denoted as $d_{\pi,P} \in \Delta(\mathcal{S} \times \mathcal{A})$, is defined as $d_{\pi,P}(s, a) = (1 - \gamma)\mathbb{E}_{\pi,P}[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}(s_t = s, a_t = a)]$.

Robust Markov Decision Process (RMDP): Unlike the standard MDP which considers a single model (transition probability function), the RMDP formulation considers a set of models. We refer to this set as the *uncertainty set* and denote it as $\mathcal{P}$. We consider $\mathcal{P}$ that satisfies the standard (s, a)-*rectangularity condition* (Iyengar, 2005). We note that a similar uncertainty set can be considered for the reward function at the expense of additional notations. However, since the analysis will be similar and the sample complexity guarantee will be identical up to a constant factor, without loss of generality, we assume that the reward function is known and deterministic.

We specify an RMDP as $M = (\mathcal{S}, \mathcal{A}, r, \mathcal{P}, \gamma, d_0)$, where the uncertainty set $\mathcal{P}$ is typically defined as

$$\mathcal{P} = \otimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{s,a}, \quad \text{where } \mathcal{P}_{s,a} = \{P_{s,a} \in \Delta(\mathcal{S}) \ : \ D(P_{s,a}, P_{s,a}^o) \leq \rho\}, \tag{1}$$

$P^o = (P_{s,a}^o, (s, a) \in \mathcal{S}\times\mathcal{A})$ is the *nominal model*, $D(\cdot, \cdot)$ is a distance metric between two probability distributions, and $\rho > 0$ is the radius of the uncertainty set that indicates the level of robustness. The

nominal model P^o can be thought as the model of the training environment. It is either the model of the simulator on which the (online) RL algorithm is trained, or in our setting, it is the model according to which the offline data is generated. The uncertainty set $\mathcal{P}$ (1) is the set of all valid transition probability functions (valid testing models) in the neighborhood of the nominal model P^o, which by definition satisfies (s,a)-rectangularity condition (Iyengar, 2005), where the neighborhood is defined using the distance metric $D(\cdot,\cdot)$ and radius ρ. In this work, we consider the *Total Variation (TV) uncertainty set* defined using the TV distance, i.e., $D(P_{s,a}, P^o_{s,a}) = (1/2)\|P_{s,a} - P^o_{s,a}\|_1$.

The RMDP problem is to find the optimal robust policy which maximizes the value against the worst possible model in the uncertainty set $\mathcal{P}$. The *robust value function* V^π corresponding to a policy π and the *optimal robust value function* V^* are defined as (Iyengar, 2005; Nilim and El Ghaoui, 2005)

$$V^\pi = \inf_{P\in\mathcal{P}} V_{\pi,P}, \qquad V^* = \sup_\pi \inf_{P\in\mathcal{P}} V_{\pi,P}. \tag{2}$$

The *optimal robust policy* π^* is such that the robust value function corresponding to it matches the optimal robust value function, i.e., $V^{\pi^*} = V^*$. It is known that there exists a deterministic optimal policy (Iyengar, 2005) for the RMDP. The *robust Bellman operator* is defined as (Iyengar, 2005)

$$(TQ)(s,a) = r(s,a) + \gamma \inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}[\max_b Q(s',b)]. \tag{3}$$

It is known that T is a contraction mapping in the infinity norm and hence it has a unique fixed point Q^* with $V^*(s) = \max_a Q^*(s,a)$ and $\pi^*(s) = \arg\max_a Q^*(s,a)$ (Iyengar, 2005). The *Robust Q-Iteration (RQI)* can now be defined using the robust Bellman operator as $Q_{k+1} = TQ_k$. Since T is a contraction, it follows that $Q_k \to Q^*$. So, RQI can be used to compute (solving the planning problem) Q^* and π^* in the tabular setting with a known $\mathcal{P}$. Due to the optimization over the uncertainty set $\mathcal{P}_{s,a}$ for each (s,a) pair, solving the planning problem in RMDP using RQI is much more computationally intensive than solving it in MDP using Q-Iteration.

Offline RL: Offline RL considers the problem of learning the optimal policy of an MDP when the algorithm does not have direct access to the environment and cannot generate data samples in an online manner. For learning the optimal policy π^*_P of an MDP with model P, the algorithm will only have access to an offline dataset $\mathcal{D}_P = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$, where $(s_i,a_i) \sim \mu$, $\mu \in \Delta(\mathcal{S}\times\mathcal{A})$ is some distribution, and $s'_i \sim P_{s_i,a_i}$. *Fitted Q-Iteration (FQI)* is a popular offline RL approach which is amenable to theoretical analysis while achieving impressive empirical performance. In addition to the dataset $\mathcal{D}_P$, FQI uses a function class $\mathcal{F} = \{f : \mathcal{S}\times\mathcal{A} \to [0, 1/(1-\gamma))]\}$ to approximate Q^*_P. The typical FQI update is given by $f_{k+1} = \arg\min_{f\in\mathcal{F}} \sum_{i=1}^N (r(s_i,a_i) + \gamma \max_b f_k(s'_i, b) - f(s_i,a_i))^2$, which aims to approximate the non-robust Bellman update using offline data with function approximation. Under suitable assumptions, it is possible to obtain provable performance guarantees for FQI (Szepesvári and Munos, 2005; Chen and Jiang, 2019; Liu et al., 2020).

3 Offline Robust Reinforcement Learning

The goal of an offline robust RL algorithm is to learn the optimal robust policy π^* using a pre-collected offline dataset $\mathcal{D}$. The data is typically generated according to a nominal (training) model P^o, i.e., $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$, where $(s_i,a_i) \sim \mu, \mu \in \Delta(\mathcal{S}\times\mathcal{A})$ is some data generating distribution, and $s'_i \sim P^o_{s_i,a_i}$. The uncertainty set $\mathcal{P}$ is defined around this nominal model P^o as given in (1) w.r.t. the total variation distance metric. We emphasize that the learning algorithm does not know the nominal model P^o as it has only access to $\mathcal{D}$, and hence it also does not know $\mathcal{P}$. Moreover, the learning algorithm does not have data generated according to any other models in $\mathcal{P}$ and has to rely only on $\mathcal{D}$ to account for the behavior w.r.t. all models in $\mathcal{P}$.

Learning policies for RL problems with large state-action spaces is computationally intractable. RL algorithms typically overcome this issue by using function approximation. In this paper, we consider two function classes $\mathcal{F} = \{f : \mathcal{S}\times\mathcal{A} \to [0, 1/(1-\gamma)]\}$ and $\mathcal{G} = \{g : \mathcal{S}\times\mathcal{A} \to [0, 2/(\rho(1-\gamma))]\}$. We use $\mathcal{F}$ to approximate Q^* and $\mathcal{G}$ to approximate the dual variable functions which we will introduce in the next section. For simplicity, we will first assume that these function classes are

finite but exponentially large, and we will use the standard log-cardinality to characterize the sample complexity results, as given in Theorem 1. We note that, at the cost of additional notations and analysis, infinite function classes can also be considered where the log-cardinalities are replaced by the appropriate notions of covering number.

Similar to the non-robust offline RL, we make the following standard assumptions about the data generating distribution μ and the representation power of $\mathcal{F}$.

Assumption 1 (Concentratability)**.** *There exists a finite constant $C > 0$ such that for any $\nu \in \{d_{\pi,P^o} \mid \text{any policy } \pi\} \subseteq \Delta(\mathcal{S} \times \mathcal{A})$, we have $\|\nu/\mu\|_\infty \leq \sqrt{C}$.*

Assumption 1 states that the ratio of the distribution ν and the data generating distribution μ, $\nu(s,a)/\mu(s,a)$, is uniformly bounded. This assumption is widely used in the offline RL literature (Munos, 2003; Agarwal et al., 2019; Chen and Jiang, 2019; Wang et al., 2021; Xie et al., 2021) in many different forms. We borrow this assumption from Chen and Jiang (2019), where they used it for non-robust offline RL. In particular, we note that the distribution ν is in the collection of discounted state-action occupancies on model P^o alone for the robust RL.

Assumption 2 (Approximate completeness)**.** *Let $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ be the data distribution. Then, $\sup_{f \in \mathcal{F}} \inf_{f' \in \mathcal{F}} \|f' - Tf\|_{2,\mu}^2 \leq \varepsilon_{\mathrm{c}}$.*

Assumption 2 states that the function class $\mathcal{F}$ is approximately closed under the robust Bellman operator T. This assumption has also been widely used in the offline RL literature (Agarwal et al., 2019; Chen and Jiang, 2019; Wang et al., 2021; Xie et al., 2021).

One of the most important properties that the function class $\mathcal{F}$ should have is that there must exist a function $f' \in \mathcal{F}$ which well-approximates Q^*. This assumption is typically called *approximate realizability* in the offline RL literature. This is typically formalized by assuming $\inf_{f \in \mathcal{F}} \|f - Tf\|_{2,\mu}^2 \leq \varepsilon_{\mathrm{r}}$ (Chen and Jiang, 2019). It is known that the approximate completeness assumption and the concentratability assumption imply the realizability assumption (Chen and Jiang, 2019; Xie et al., 2021).

4 Robust Fitted Q-Iteration: Algorithm and Main Results

In this section, we give a step-by-step approach to overcome the challenges of the offline robust RL outlined in Section 1. We then combine these intermediate steps to obtain our proposed RFQI algorithm. We then present our main result about the performance guarantee of the RFQI algorithm, followed by a brief description about the proof approach.

4.1 Dual Reformulation of Robust Bellman Operator

One key challenge in directly using the standard definition of the optimal robust value function given in (2) or of the robust Bellman operator given in (3) for developing and analyzing robust RL algorithms is that both involve computing an expectation w.r.t. each model $P \in \mathcal{P}$. Given that the data is generated only according to the nominal model P^o, estimating these expectation values is really challenging. We show that we can overcome this difficulty through the dual reformulation of the robust Bellman operator, as given below.

Proposition 1. *Let M be an RMDP with the uncertainty set $\mathcal{P}$ specified by (1) using the total variation distance $D(P_{s,a}, P^o_{s,a}) = (1/2)\|P_{s,a} - P^o_{s,a}\|_1$. Then, for any $Q : \mathcal{S} \times \mathcal{A} \to [0, 1/(1-\gamma)]$, the robust Bellman operator T given in (3) can be equivalently written as*

$$(TQ)(s,a) = r(s,a) - \gamma \inf_{\eta \in [0, \frac{2}{\rho(1-\gamma)}]} (\mathbb{E}_{s' \sim P^o_{s,a}}[(\eta - V(s'))_+] - \eta + \rho(\eta - \inf_{s''} V(s''))_+), \quad (4)$$

where $V(s) = \max_{a \in \mathcal{A}} Q(s,a)$. Moreover, the inner optimization problem in (4) is convex in η.

Note that in (4), the expectation is now only w.r.t. the nominal model P^o, which opens up the possibility of using empirical estimates obtained from the data generated according to P^o. This avoids the need to use importance sampling based techniques to account for all models in $\mathcal{P}$, which often have high variance, and thus, are not desirable.

While (4) provides a form that is amenable to estimation using offline data, it involves finding $\inf_{s''} V(s'')$. Though this computation is straightforward in a tabular setting, it is infeasible in a function approximation setting. In order to overcome this issue, we make the following assumption.

Assumption 3 (Fail-state)**.** *The RMDP M has a 'fail-state' s_f, such that $r(s_f, a) = 0$ and $P_{s_f,a}(s_f) = 1,\ \forall a \in \mathcal{A},\ \forall P \in \mathcal{P}$.*

We note that this is not a very restrictive assumption because such a 'fail-state' is quite natural in most simulated or real-world systems. For example, a state where a robot collapses and not able to get up, either in a simulation environment like MuJoCo or in real-world setting, is such a fail state.

Assumption 3 immediately implies that $V_{\pi,P}(s_f) = 0,\ \forall P \in \mathcal{P}$, and hence $V^*(s_f) = 0$ and $Q^*(s_f, a) = 0,\ \forall a \in \mathcal{A}$. It is also straightforward to see that $Q_{k+1}(s_f, a) = 0,\ \forall a \in \mathcal{A}$, where Q_k's are the RQI iterates given by the robust Bellman update $Q_{k+1} = TQ_k$ with the initialization $Q_0 = 0$. By the contraction property of T, we have $Q_k \to Q^*$. So, under Assumption 3, without loss of generality, we can always keep $Q_k(s_f, a) = 0,\ \forall a \in \mathcal{A}$ and for all k in RQI (and later in RFQI). So, in the light of the above description, for the rest of the paper we will use the robust Bellman operator T by setting $\inf_{s''} V(s'') = 0$. In particular, for any function $f : \mathcal{S} \times \mathcal{A} \to [0, 1/(1-\gamma)]$ with $f(s_f, a) = 0$, the robust Bellman operator T is now given by

$$(Tf)(s,a) = r(s,a) - \gamma \inf_{\eta \in [0, \frac{2}{(\rho(1-\gamma))}]} (\mathbb{E}_{s' \sim P^o_{s,a}}[(\eta - \max_{a'} f(s', a'))_+] - (1-\rho)\eta). \quad (5)$$

4.2 Approximately Solving the Dual Optimization using Empirical Risk Minimization

Another key challenge in directly using the standard definition of the optimal robust value function given in (2) or of the robust Bellman operator given in (3) for developing and analyzing robust RL algorithms is that both involve an optimization over $\mathcal{P}$. The dual reformulation given in (5) partially overcomes this challenge also, as the optimization over $\mathcal{P}$ is now replaced by a convex optimization over a scalar $\eta \in [0, 2/(\rho(1-\gamma))]$. However, this still requires solving an optimization for each $(s, a) \in \mathcal{S} \times \mathcal{A}$, which is clearly infeasible even for moderately sized state-action spaces, not to mention the function approximation setting. Our key idea to overcome this difficulty is to reformulate this as a functional optimization problem instead of solving it as multiple scalar optimization problems. This functional optimization method will make it amenable to approximately solving the dual problem using an empirical risk minimization approach with offline data.

Consider the probability (measure) space $(\mathcal{S} \times \mathcal{A}, \Sigma(\mathcal{S} \times \mathcal{A}), \mu)$ and let $L^1(\mathcal{S} \times \mathcal{A}, \Sigma(\mathcal{S} \times \mathcal{A}), \mu)$ be the set of all absolutely integrable functions defined on this space.[1] In other words, L^1 is the set of all functions $g : \mathcal{S} \times \mathcal{A} \to \mathcal{C} \subset \mathbb{R}$, such that $\|g\|_{1,\mu}$ is finite. We set $\mathcal{C} = [0, 2/\rho(1-\gamma)]$, anticipating the solution of the dual optimization problem (5). We also note μ is the data generating distribution which is a σ-finite measure.

For any given function $f : \mathcal{S} \times \mathcal{A} \to [0, 1/(1-\gamma)]$, we define the loss function $L_{\text{dual}}(\cdot; f)$ as

$$L_{\text{dual}}(g; f) = \mathbb{E}_{s,a \sim \mu}[\mathbb{E}_{s' \sim P^o_{s,a}}[(g(s,a) - \max_{a'} f(s', a'))_+] - (1-\rho)g(s,a)], \quad \forall g \in L^1. \quad (6)$$

In the following lemma, we show that the scalar optimization over η for each (s, a) pair in (5) can be replaced by a single functional optimization w.r.t. the loss function L_{dual}.

Lemma 1. *Let L_{dual} be the loss function defined in (6). Then, for any function $f : \mathcal{S} \times \mathcal{A} \to [0, 1/(1-\gamma)]$, we have*

$$\inf_{g \in L^1} L_{\text{dual}}(g; f) = \mathbb{E}_{s,a \sim \mu}\Big[\inf_{\eta \in [0, \frac{2}{(\rho(1-\gamma))}]} \Big(\mathbb{E}_{s' \sim P^o_{s,a}}\big[\big(\eta - \max_{a'} f(s', a')\big)_+\big] - (1-\rho)\eta\Big)\Big]. \quad (7)$$

Note that the RHS of (7) has minimization over η for each (s, a) pair and minimization is inside the expectation $\mathbb{E}_{s,a \sim \mu}[\cdot]$. However, the LHS of (7) has a single functional minimization over $g \in L^1$ and this minimization is outside the expectation. For interchanging the expectation and minimization,

[1] In the following, we will simply denote $L^1(\mathcal{S} \times \mathcal{A}, \Sigma(\mathcal{S} \times \mathcal{A}), \mu)$ as L^1 for conciseness.

and for moving from point-wise optimization to functional optimization, we use the result from Rockafellar and Wets (2009, Theorem 14.60), along with the fact that L^1 is a decomposable space. We also note that this result has been used in many recent works on distributionally robust optimization (Shapiro, 2017; Duchi and Namkoong, 2018) (see Appendix A for more details).

We can now define the empirical loss function $\widehat{L}_{\text{dual}}$ corresponding to the true loss L_{dual} as

$$\widehat{L}_{\text{dual}}(g; f) = \frac{1}{N}\sum_{i=1}^{N}(g(s_i, a_i) - \max_{a'} f(s_i', a'))_+ - (1-\rho)g(s_i, a_i). \tag{8}$$

Now, for any given f, we can find an approximately optimal dual function through the *empirical risk minimization* approach as $\inf_{g\in L^1} \widehat{L}_{\text{dual}}(g; f)$.

As we mentioned in Section 3, our offline robust RL algorithm is given an input function class $\mathcal{G} = \{g : \mathcal{S}\times\mathcal{A} \to [0, 2/(\rho(1-\gamma))]\}$ to approximate the dual variable functions. So, in the empirical risk minimization, instead of taking the infimum over all the functions in L^1, we can only take the infimum over all the functions in $\mathcal{G}$. For this to be meaningful, $\mathcal{G}$ should have sufficient representation power. In particular, the result in Lemma 1 should hold approximately even if we replace the infimum over L^1 with infimum over $\mathcal{G}$. One can see that this is similar to the realizability requirement for the function class $\mathcal{F}$ as described in Section 3. We formalize the representation power of $\mathcal{G}$ in the following assumption.

Assumption 4 (Approximate dual realizability)**.** *For all $f \in \mathcal{F}$, there exists a uniform constant ε_{dual} such that* $\inf_{g\in\mathcal{G}} L_{\text{dual}}(g; f) - \inf_{g\in L^1} L_{\text{dual}}(g; f) \le \varepsilon_{dual}$

Using the above assumption, for any given $f \in \mathcal{F}$, we can find an approximately optimal dual function $\widehat{g}_f \in \mathcal{G}$ through the *empirical risk minimization* approach as $\widehat{g}_f = \arg\min_{g\in\mathcal{G}} \widehat{L}_{\text{dual}}(g; f)$.

In order to characterize the performance of this approach, consider the operator T_g for any $g \in \mathcal{G}$ as

$$(T_g f)(s,a) = r(s,a) - \gamma(\mathbb{E}_{s'\sim P^o_{s,a}}[(g(s,a) - \max_{a'} f(s',a'))_+] - (1-\rho)g(s,a)), \tag{9}$$

for all $f \in \mathcal{F}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$. We will show in Lemma 6 in Appendix C that the error $\sup_{f\in\mathcal{F}} \|Tf - T_{\widehat{g}_f} f\|_{1,\mu}$ is $\mathcal{O}(\log(|\mathcal{F}|/\delta)/\sqrt{N})$ with probability at least $1-\delta$.

4.3 Robust Fitted Q-iteration

The intuitive idea behind our robust fitted Q-iteration (RFQI) algorithm is to approximate the exact RQI update step $Q_{k+1} = TQ_k$ with function approximation using offline data. The exact RQI step requires updating each (s,a)-pair separately, which is not scalable to large state-action spaces. So, this is replaced by the function approximation as $Q_{k+1} = \arg\min_{f\in\mathcal{F}} \|TQ_k - f\|^2_{2,\nu}$. It is still infeasible to perform this update as it requires to exactly compute the expectation (w.r.t. P^o and ν) and to solve the dual problem accurately. We overcome these issues by replacing both these exact computations with empirical estimates using the offline data. We note that this intuitive idea is similar to that of the FQI algorithm in the non-robust case. However, RFQI has unique challenges due to the nature of the robust Bellman operator T and the presence of the dual optimization problem within T.

Given a dataset $\mathcal{D}$, we also follow the standard non-robust offline RL choice of least-squares residual minimization (Chen and Jiang, 2019; Xie et al., 2021; Wang et al., 2021). Define the empirical loss of f given f' (which represents the Q-function from the last iteration) and dual variable function g as

$$\widehat{L}_{\text{RFQI}}(f; f', g) = \frac{1}{N}\sum_{i=1}^{N}\Big(r(s,a) + \gamma\big(-(g(s_i,a_i) - \max_{a'} f'(s_i', a'))_+ + (1-\rho)g(s_i,a_i)\big) - f(s_i,a_i)\Big)^2. \tag{10}$$

The correct dual variable function to be used in (10) is the optimal dual variable $g^*_{f'} = \arg\min_{g\in\mathcal{G}} L_{\text{dual}}(g; f')$ corresponding to the last iterate f', which we will approximate it by $\widehat{g}_{f'} = \arg\min_{g\in\mathcal{G}} \widehat{L}_{\text{dual}}(g; f')$. The RFQI update is then obtained as $\arg\min_{f\in\mathcal{F}} \widehat{L}_{\text{RFQI}}(f; f', \widehat{g}_{f'})$.

Summarizing the individual steps described above, we formally give our RFQI algorithm below.

Algorithm 1 Robust Fitted Q-Iteration (RFQI) Algorithm

Input: Offline dataset $\mathcal{D} = (s_i, a_i, r_i, s_i')_{i=1}^N$, function classes $\mathcal{F}$ and $\mathcal{G}$.
Initialize: $Q_0 \equiv 0 \in \mathcal{F}$.
for $k = 0, \cdots, K-1$ **do**
  Dual variable function optimization: Compute the dual variable function corresponding to Q_k through empirical risk minimization as $g_k = \widehat{g}_{Q_k} = \arg\min_{g \in \mathcal{G}} \widehat{L}_{\text{dual}}(g; Q_k)$ (see (8)).
  Robust Q-update: Compute the next iterate Q_{k+1} through least-squares regression as $Q_{k+1} = \arg\min_{Q \in \mathcal{F}} \widehat{L}_{\text{RFQI}}(Q; Q_k, g_k)$ (see (10)).
end for
Output: $\pi_K = \arg\max_a Q_K(s, a)$

Now we state our main theoretical result on the performance of the RFQI algorithm.

Theorem 1. *Let Assumptions 1-4 hold. Let π_K be the output of the RFQI algorithm after K iterations. Denote $J^\pi = \mathbb{E}_{s \sim d_0}[V^\pi(s)]$ where d_0 is initial state distribution. Then, for any $\delta \in (0, 1)$, with probability at least $1 - 2\delta$, we have*

$$J^{\pi^*} - J^{\pi_K} \leq \frac{\gamma^K}{(1-\gamma)^2} + \frac{\sqrt{C}(\sqrt{6\varepsilon_c} + \gamma\varepsilon_{dual})}{(1-\gamma)^2} + \frac{16}{\rho(1-\gamma)^3}\sqrt{\frac{18C\log(2|\mathcal{F}||\mathcal{G}|/\delta)}{N}}.$$

Remark 1. Theorem 1 states that the RFQI algorithm can achieve approximate optimality. To see this, note that with $K \geq \mathcal{O}(\frac{1}{\log(1/\gamma)}\log(\frac{1}{\varepsilon(1-\gamma)}))$, and neglecting the second term corresponding to (inevitable) approximation errors ε_c and $\varepsilon_{\text{dual}}$, we get $J^{\pi^*} - J^{\pi_K} \leq \varepsilon/(1-\gamma)$ with probability greater than $1 - 2\delta$ for any $\varepsilon, \delta \in (0, 1)$, as long as the number of samples $N \geq \mathcal{O}(\frac{1}{(\rho\varepsilon)^2(1-\gamma)^4}\log\frac{|\mathcal{F}||\mathcal{G}|}{\delta})$. So, the above theorem can also be interpreted as a **sample complexity** result.

Remark 2. The known sample complexity of robust-RL in the tabular setting is $\widetilde{\mathcal{O}}(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(\rho\varepsilon)^2(1-\gamma)^4})$ (Yang et al., 2021; Panaganti and Kalathil, 2022). Considering $\widetilde{\mathcal{O}}(\log(|\mathcal{F}||\mathcal{G}|))$ to be $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|)$, we can recover the same bound as in the tabular setting (we save $|\mathcal{S}|$ due to the use of Bernstein inequality).

Remark 3. Under similar Bellman completeness and concentratability assumptions, RFQI sample complexity is comparable to that of a non-robust offline RL algorithm, i.e., $\mathcal{O}(\frac{1}{\varepsilon^2(1-\gamma)^4}\log\frac{|\mathcal{F}|}{\delta})$ (Chen and Jiang, 2019). As a consequence of robustness, we have ρ^{-2} and $\log(|\mathcal{G}|)$ factors in our bound.

4.4 Proof Sketch

Here we briefly explain the key ideas used in the analysis of RFQI for obtaining the optimality gap bound in Theorem 1. The complete proof is provided in Appendix C.

Step 1: To bound $J^{\pi^*} - J^{\pi_K}$, we connect it to the error $\|Q^{\pi^*} - Q_K\|_{1,\nu}$ for any state-action distribution ν. While the similar step follows almost immediately using the well-known performance lemma in the analysis of non-robust FQI, such a result is not known in the robust RL setting. So, we derive the basic inequalities to get a recursive form and to obtain the bound $J^{\pi^*} - J^{\pi_K} \leq 2\|Q^{\pi^*} - Q_K\|_{1,\nu}/(1-\gamma)$ (see (22) and the steps before in Appendix C).

Step 2: To bound $\|Q^{\pi^*} - Q_K\|_{1,\nu}$ for any state-action distribution ν such that $\|\nu/\mu\|_\infty \leq \sqrt{C}$, we decompose it to get a recursion, with approximation terms based on the least-squares regression and empirical risk minimization. Recall that $\widehat{g}_f$ is the dual variable function from the algorithm for state-action value function $f \in \mathcal{F}$. Denote $\widehat{f}_g$ as the least squares solution from the algorithm for the state-action value function $f \in \mathcal{F}$ and dual variable function $g \in \mathcal{G}$, i.e., $\widehat{f}_g = \arg\min_{Q \in \mathcal{F}} \widehat{L}_{\text{RFQI}}(Q; f, g)$. By recursive use of the obtained inequality (23) (see Appendix C) and using uniform bound, we get

$$\|Q^{\pi^*} - Q_K\|_{1,\nu} \leq \frac{\gamma^K}{1-\gamma} + \frac{\sqrt{C}}{1-\gamma}\sup_{f \in \mathcal{F}}\|Tf - T_{\widehat{g}_f} f\|_{1,\mu} + \frac{\sqrt{C}}{1-\gamma}\sup_{f \in \mathcal{F}}\sup_{g \in \mathcal{G}}\|T_g f - \widehat{f}_g\|_{2,\mu}.$$

Step 3: We recognize that $\sup_{f \in \mathcal{F}}\|Tf - T_{\widehat{g}_f} f\|_{1,\mu}$ is an empirical risk minimization error term. Using Rademacher complexity based bounds, we show in Lemma 6 that this error is $\mathcal{O}(\log(|\mathcal{F}|/\delta)/\sqrt{N})$

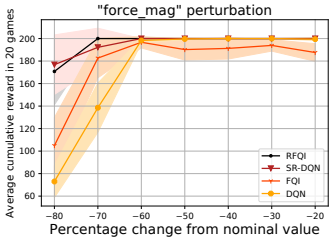

Figure 1: CartPole

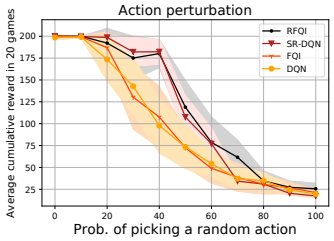

Figure 2: CartPole

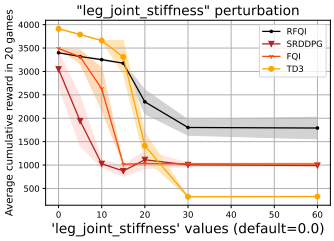

Figure 3: Hopper

with high probability.

Step 4: Similarly, we also recognize that $\sup_{f\in\mathcal{F}}\sup_{g\in\mathcal{G}}\|T_g f-\widehat{f}_g\|_{2,\mu}$ is a least-squares regression error term. We also show that this error is $\mathcal{O}(\log(|\mathcal{F}||\mathcal{G}|/\delta)/\sqrt{N})$ with high probability. We adapt the generalized least squares regression result to accommodate the modified target functions resulting from the robust Bellman operator to obtain this bound (see Lemma 7).

The proof is complete after combining steps 1-4 above.

5 Experiments

Here, we demonstrate the robust performance of our RFQI algorithm by evaluating it on *Cartpole* and *Hopper* environments in OpenAI Gym (Brockman et al., 2016). In all the figures shown, the quantity in the vertical axis is averaged over 20 different seeded runs depicted by the thick line and the band around it is the ± 0.5 standard deviation. *Due page limit, a more detailed description of the experiments, and results on additional experiments, are deferred to Appendix E.*

For the *Cartpole*, we compare RFQI algorithm against the non-robust RL algorithms FQI and DQN, and the soft-robust RL algorithm proposed in Derman et al. (2018). We test the robustness of the algorithms by changing the parameter *force_mag* (to model external force disturbance), and also by introducing action perturbations (to model actuator noise). Fig. 1 and Fig. 2 shows superior robust performance of RFQI compared to the non-robust FQI and DQN. The RFQI performance is similar to that of soft-robust DQN. We note that soft-robust RL algorithm (here soft-robust DQN) is an online deep RL algorithm (and not an offline RL algorithm) and has no provable performance guarantee. Moreover, soft-robust RL algorithm requires generating online data according a number of models in the uncertainty set, whereas RFQI only requires offline data according to a single nominal training model.

For the *Hopper*, we compare RFQI algorithm against the non-robust RL algorithms FQI and TD3 (Fujimoto et al., 2018), and the soft-robust RL (here soft-robust DDPG) algorithm proposed in Derman et al. (2018). We test the robustness of the algorithms by changing the parameter *leg_joint_stiffness*. Fig. 3 shows the superior performance of our RFQI algorithm against the non-robust algorithms and soft-robust DDPG algorithm. The average episodic reward of RFQI remains almost the same initially, and later decays much less and gracefully when compared to the non-robust FQI and TD3.

6 Conclusion

In this work, we presented a novel robust RL algorithm called Robust Fitted Q-Iteration algorithm with provably optimal performance for an RMDP with arbitrarily large state space, using only offline data with function approximation. We also demonstrated the superior performance of the proposed algorithm on standard benchmark problems.

One limitation of our present work is that, we considered only the uncertainty set defined with respect to the total variation distance. In future work, we will consider uncertainty sets defined with respect to other f-divergences such as KL-divergence and Chi-square divergence. Finding a lower bound for the sample complexity and relaxing the assumptions used are also important and challenging problems.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2019). Reinforcement learning: Theory and algorithms. CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep. 5, 20, 23

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In Conference on Learning Theory, pages 67–83. 24

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. Machine Learning, 71(1):89–129. 2, 24

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. Mach. Learn., 91(3):325–349. 24

Bertsekas, D. P. (2011). Approximate policy iteration: A survey and some new methods. Journal of Control Theory and Applications, 9(3):310–335. 2, 24

Borkar, V. S. (2002). Q-learning for risk-sensitive control. Mathematics of operations research, 27(2):294–311. 25

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. arXiv preprint arXiv:1606.01540. 9, 25, 27, 28, 29

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In International Conference on Machine Learning, pages 1042–1051. 2, 4, 5, 7, 8, 24

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. studia scientiarum Mathematicarum Hungarica, 2:229–318. 17

Derman, E., Mankowitz, D., Mann, T., and Mannor, S. (2020). A bayesian approach to robust reinforcement learning. In Uncertainty in Artificial Intelligence, pages 648–658. 2

Derman, E., Mankowitz, D. J., Mann, T. A., and Mannor, S. (2018). Soft-robust actor-critic policy-gradient. In AUAI press for Association for Uncertainty in Artificial Intelligence, pages 208–218. 2, 9, 24, 27

Duchi, J. and Namkoong, H. (2018). Learning models with uniform performance via distributionally robust optimization. arXiv preprint arXiv:1810.08750. 7, 17

Dullerud, G. E. and Paganini, F. (2013). A course in robust control theory: a convex approach, volume 36. Springer Science & Business Media. 25

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. Journal of Machine Learning Research, 6:503–556. 2, 24

Farahmand, A.-m., Szepesvári, C., and Munos, R. (2010). Error propagation for approximate policy and value iteration. Advances in Neural Information Processing Systems, 23. 2, 24

Fei, Y., Yang, Z., Chen, Y., and Wang, Z. (2021). Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. In Annual Conference on Neural Information Processing Systems 2021, pages 20436–20446. 25

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020). D4rl: Datasets for deep data-driven reinforcement learning. 27, 29, 30

Fujimoto, S., Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In International Conference on Machine Learning, pages 1582–1591. 9, 28

Fujimoto, S., Meger, D., and Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In International Conference on Machine Learning, pages 2052–2062. 2, 24, 25, 26, 27, 29

Gordon, G. J. (1995). Stable function approximation in dynamic programming. In Machine learning proceedings 1995, pages 261–268. 2, 24

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning, pages 1861–1870. 27

Haskell, W. B., Jain, R., and Kalathil, D. (2016). Empirical dynamic programming. Mathematics of Operations Research, 41(2):402–429. 24

Huang, P., Xu, M., Fang, F., and Zhao, D. (2022). Robust reinforcement learning as a stackelberg game via adaptively-regularized adversarial training. arXiv preprint arXiv:2202.09514. 25

Iyengar, G. N. (2005). Robust dynamic programming. Mathematics of Operations Research, 30(2):257–280. 1, 2, 3, 4, 22, 24

Kalathil, D., Borkar, V. S., and Jain, R. (2021). Empirical Q-Value Iteration. Stochastic Systems, 11(1):1–18. 24

Kaufman, D. L. and Schaefer, A. J. (2013). Robust modified policy iteration. INFORMS Journal on Computing, 25(3):396–410. 24

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 26

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. 25

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. In Advances in Neural Information Processing Systems, pages 11784–11794. 2, 24

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems, 33:1179–1191. 2, 24

Lange, S., Gabel, T., and Riedmiller, M. (2012). Batch reinforcement learning. In Reinforcement learning, pages 45–73. Springer. 2, 24

Lazaric, A., Ghavamzadeh, M., and Munos, R. (2012). Finite-sample analysis of least-squares policy iteration. Journal of Machine Learning Research, 13:3041–3074. 2, 24

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643. 2, 24, 27, 29, 30

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. In Advances in Neural Information Processing Systems, volume 33, pages 12861–12872. 24

Lim, S. H. and Autef, A. (2019). Kernel-based reinforcement learning in robust Markov decision processes. In International Conference on Machine Learning, pages 3973–3981. 25

Lim, S. H., Xu, H., and Mannor, S. (2013). Reinforcement learning in robust Markov decision processes. In Advances in Neural Information Processing Systems, pages 701–709. 2

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2020). Provably good batch off-policy reinforcement learning without great exploration. In Neural Information Processing Systems. 2, 4, 24, 25, 26, 27, 29, 30

Mankowitz, D. J., Levine, N., Jeong, R., Abdolmaleki, A., Springenberg, J. T., Shi, Y., Kay, J., Hester, T., Mann, T., and Riedmiller, M. (2020). Robust reinforcement learning for continuous control with model misspecification. In International Conference on Learning Representations. 2, 25

Mannor, S., Mebel, O., and Xu, H. (2016). Robust mdps with k-rectangular uncertainty. Mathematics of Operations Research, 41(4):1484–1509. 2

Moses, A. K. and Sundaresan, R. (2011). Further results on geometric properties of a family of relative entropies. In 2011 IEEE International Symposium on Information Theory Proceedings, pages 1940–1944. 17

Munos, R. (2003). Error bounds for approximate policy iteration. In ICML, volume 3, pages 560–567. 5

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. Journal of Machine Learning Research, 9(27):815–857. 2, 24

Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. Operations Research, 53(5):780–798. 1, 2, 4, 24

Panaganti, K. and Kalathil, D. (2021). Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In Proceedings of the 38th International Conference on Machine Learning, pages 511–520. 2, 25

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, pages 9582–9602. 2, 8, 25

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. (2018). Sim-to-real transfer of robotic control with dynamics randomization. In 2018 IEEE international conference on robotics and automation (ICRA), pages 3803–3810. IEEE. 1

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. (2017). Robust adversarial reinforcement learning. In International Conference on Machine Learning, pages 2817–2826. 2, 25

Prashanth, L. A. and Ghavamzadeh, M. (2016). Variance-constrained actor-critic algorithms for discounted and average reward mdps. Mach. Learn., 105(3):367–417. 25

Raffin, A. (2020). Rl baselines3 zoo. `https://github.com/DLR-RM/rl-baselines3-zoo`. 27

Rockafellar, R. T. and Wets, R. J.-B. (2009). Variational analysis, volume 317. Springer Science & Business Media. 7, 16, 17

Roy, A., Xu, H., and Pokutta, S. (2017). Reinforcement learning under model mismatch. In Advances in Neural Information Processing Systems, pages 3043–3052. 2, 24

Russel, R. H. and Petrik, M. (2019). Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps. Advances in Neural Information Processing Systems. 2

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347. 27

Shalev-Shwartz, S. and Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press. 16

Shapiro, A. (2017). Distributionally robust stochastic programming. SIAM Journal on Optimization, 27(4):2258–2275. 7, 17

Sidford, A., Wang, M., Wu, X., Yang, L. F., and Ye, Y. (2018). Near-optimal time and sample complexities for solving markov decision processes with a generative model. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 5192–5202. 24

Singh, S. P. and Yee, R. C. (1994). An upper bound on the loss from approximate optimal-value functions. Machine Learning, 16(3):227–233. 24

Sünderhauf, N., Brock, O., Scheirer, W., Hadsell, R., Fox, D., Leitner, J., Upcroft, B., Abbeel, P., Burgard, W., Milford, M., et al. (2018). The limits and potentials of deep learning for robotics. The International journal of robotics research, 37(4-5):405–420. 1

Szepesvári, C. and Munos, R. (2005). Finite time bounds for sampling based fitted value iteration. In Proceedings of the 22nd international conference on Machine learning, pages 880–887. 4

Tamar, A., Mannor, S., and Xu, H. (2014). Scaling up robust mdps using function approximation. In International Conference on Machine Learning, pages 181–189. 2, 24

Tessler, C., Efroni, Y., and Mannor, S. (2019). Action robust reinforcement learning and applications in continuous control. In International Conference on Machine Learning, pages 6215–6224. 24

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), pages 23–30. 1

Vershynin, R. (2018). High-Dimensional Probability: An Introduction with Applications in Data Science, volume 47. Cambridge University press. 16

Wang, R., Foster, D., and Kakade, S. M. (2021). What are the statistical limits of offline {rl} with linear function approximation? In International Conference on Learning Representations. 5, 7

Wang, Y. and Zou, S. (2021). Online robust reinforcement learning with model uncertainty. Advances in Neural Information Processing Systems, 34. 2

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust Markov decision processes. Mathematics of Operations Research, 38(1):153–183. 2, 24

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. Advances in neural information processing systems, 34. 2, 5, 7, 24

Xie, T. and Jiang, N. (2020). Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In Conference on Uncertainty in Artificial Intelligence, pages 550–559. 2, 24

Xu, H. and Mannor, S. (2010). Distributionally robust Markov decision processes. In Advances in Neural Information Processing Systems, pages 2505–2513. 2, 24

Yang, W., Zhang, L., and Zhang, Z. (2021). Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. arXiv preprint arXiv:2105.03863. 2, 8, 25

Yu, P. and Xu, H. (2015). Distributionally robust counterpart in Markov decision processes. IEEE Transactions on Automatic Control, 61(9):2538–2543. 2

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. (2020). Mopo: Model-based offline policy optimization. In Advances in Neural Information Processing Systems. 2, 24

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. (2020a). Robust deep reinforcement learning against adversarial perturbations on state observations. Advances in Neural Information Processing Systems, 33:21024–21037. 2

Zhang, K., Hu, B., and Basar, T. (2020b). Policy optimization for H_2 linear control with H_∞ robustness guarantee: Implicit regularization and global convergence. In Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control, volume 120, pages 179–190. 25

Zhang, S. and Jiang, N. (2021). Towards hyperparameter-free policy selection for offline reinforcement learning. In Advances in Neural Information Processing Systems, pages 12864–12875. 2, 24

Zhang, Y., Yang, Z., and Wang, Z. (2021). Provably efficient actor-critic for risk-sensitive and robust adversarial rl: A linear-quadratic case. In International Conference on Artificial Intelligence and Statistics, pages 2764–2772. 25

Zhou, K., Doyle, J. C., Glover, K., et al. (1996). Robust and optimal control, volume 40. Prentice hall New Jersey. 25

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In International Conference on Artificial Intelligence and Statistics, pages 3331–3339. 2, 25

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See **contributions** in the Introduction.
   (b) Did you describe the limitations of your work? [Yes] The discussions on the assumptions describes the limitations.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 3-4.3
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide proof sketch 4.4 in main paper and the complete proof in Appendix with self-contained material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Described in the Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Described in the main paper and the Appendix.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Mentioned in the Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Useful Technical Results

In this section, we state some existing results from concentration inequalities, generalization bounds, and optimization theory that we will use later in our analysis. We first state the Berstein's inequality that utilizes second-moment to get a tighter concentration inequality.

Lemma 2 (Bernstein's inequality (Vershynin, 2018, Theorem 2.8.4))**.** *Let $X_1, \cdots, X_T$ be independent random variables. Assume that $|X_t - \mathbb{E}[X_t]| \leq M$, for all t. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}(X_t - \mathbb{E}[X_t])\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{T^2\varepsilon^2}{2\sigma^2 + \frac{2MT\varepsilon}{3}}\right),$$

where $\sigma^2 = \sum_{t=1}^{T}\mathbb{E}[X_t^2]$. Furthermore, if $X_1, \cdots, X_T$ are independent and identically distributed random variables, then for any $\delta \in (0,1)$, we have

$$\left|\mathbb{E}[X_1] - \frac{1}{T}\sum_{t=1}^{T}X_t\right| \leq \sqrt{\frac{2\mathbb{E}[X_1^2]\log(2/\delta)}{T}} + \frac{M\log(2/\delta)}{3T},$$

with probability at least $1-\delta$.

We now state a result for the generalization bounds on empirical risk minimization (ERM) problems. This result is adapted from Shalev-Shwartz and Ben-David (2014, Theorem 26.5, Lemma 26.8, Lemma 26.9).

Lemma 3 (ERM generalization bound)**.** *Let P be the data generating distribution on the space $\mathcal{X}$ and let $\mathcal{H}$ be a given hypothesis class of functions. Assume that for all $x \in \mathcal{X}$ and $h \in \mathcal{H}$ we have that $|l(h,x)| \leq c_1$ for some positive constant $c_1 > 0$. Given a dataset $\mathcal{D} = \{X_i\}_{i=1}^N$, generated independently from P, denote $\hat{h}$ as the ERM solution, i.e. $\hat{h} = \arg\min_{h\in\mathcal{H}}(1/N)\sum_{i=1}^N l(h, X_i)$. For any fixed $\delta \in (0,1)$ and $h^* \in \arg\min_{h\in\mathcal{H}}\mathbb{E}_{X\sim P}[l(h,X)]$, we have*

$$\mathbb{E}_{X\sim P}[l(\hat{h},X)] - \mathbb{E}_{X\sim P}[l(h^*,X)] \leq 2R(l\circ\mathcal{H}\circ\mathcal{D}) + 5c_1\sqrt{\frac{2\log(8/\delta)}{N}}, \tag{11}$$

with probability at least $1-\delta$, where $R(\cdot)$ is the Rademacher complexity of $l\circ\mathcal{H}$ given by

$$R(l\circ\mathcal{H}\circ\mathcal{D}) = \frac{1}{N}\mathbb{E}_{\{\sigma_i\}_{i=1}^N}\left(\sup_{g\,\in\, l\circ\mathcal{H}}\ \sum_{i=1}^{N}\sigma_i g(X_i)\right),$$

in which σ_i's are independent from X_i's and are independently and identically distributed according to the Rademacher random variable σ, i.e. $\mathbb{P}(\sigma = 1) = 0.5 = \mathbb{P}(\sigma = -1)$.

Furthermore, if $\mathcal{H}$ is a finite hypothesis class, i.e. $|\mathcal{H}| < \infty$, with $|h\circ x| \leq c_2$ for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$, and $l(h,x)$ is c_3-Lipschitz in h, then we have

$$\mathbb{E}_{X\sim P}[l(\hat{h},X)] - \mathbb{E}_{X\sim P}[l(h^*,X)] \leq 2c_2c_3\sqrt{\frac{2\log(|\mathcal{H}|)}{N}} + 5c_1\sqrt{\frac{2\log(8/\delta)}{N}}, \tag{12}$$

with probability at least $1-\delta$.

We now mention two important concepts from variational analysis (Rockafellar and Wets, 2009) literature that is useful to relate minimization of integrals and the integrals of pointwise minimization under special class of functions.

Definition 1 (Decomposable spaces and Normal integrands (Rockafellar and Wets, 2009, Definition 14.59, Example 14.29))**.** *A space $\mathcal{X}$ of measurable functions is a decomposable space relative to an underlying measure space $(\Omega, \mathcal{A}, \mu)$, if for every function $x_0 \in \mathcal{X}$, every set $A \in \mathcal{A}$ with $\mu(A) < \infty$, and any bounded measurable function $x_1 : A \to \mathbb{R}$, the function $x(\omega) = x_0(\omega)\mathbb{1}(\omega \notin A) + x_1(\omega)\mathbb{1}(\omega \in A)$ belongs to $\mathcal{X}$. A function $f : \Omega\times\mathbb{R} \to \mathbb{R}$ (finite-valued) is a normal integrand, if and only if $f(\omega, x)$ is $\mathcal{A}$-measurable in ω for each x and is continuous in x for each ω.*

Remark 4. A few examples of decomposable spaces are $L^p(\mathcal{S}\times\mathcal{A},\Sigma(\mathcal{S}\times\mathcal{A}),\mu)$ for any $p\geq 1$ and $\mathcal{M}(\mathcal{S}\times\mathcal{A},\Sigma(\mathcal{S}\times\mathcal{A}))$, the space of all $\Sigma(\mathcal{S}\times\mathcal{A})$-measurable functions.

Lemma 4 (Rockafellar and Wets, 2009, Theorem 14.60)**.** *Let $\mathcal{X}$ be a space of measurable functions from Ω to $\mathbb{R}$ that is decomposable relative to a σ-finite measure μ on the σ-algebra $\mathcal{A}$. Let $f:\Omega\times\mathbb{R}\to\mathbb{R}$ (finite-valued) be a normal integrand. Then, we have*

$$\inf_{x\in\mathcal{X}}\int_{\omega\in\Omega} f(\omega,x(\omega))\mu(\mathrm{d}\,\omega)=\int_{\omega\in\Omega}\left(\inf_{x\in\mathbb{R}} f(\omega,x)\right)\mu(\mathrm{d}\,\omega).$$

Moreover, as long as the above infimum is not $-\infty$, we have that

$$x'\in\arg\min_{x\in\mathcal{X}}\int_{\omega\in\Omega} f(\omega,x(\omega))\mu(\mathrm{d}\,\omega),$$

if and only if $x'(\omega)\in\arg\min_{x\in\mathbb{R}} f(\omega,x)\cdot\mu$ almost surely.

We now give one result from distributioanlly robust optimization. The f-divergence between the distributions P and P^o is defined as

$$D_f(P\|P^o)=\int f(\frac{\mathrm{d}\,P}{\mathrm{d}\,P^o})\mathrm{d}\,P^o, \tag{13}$$

where f is a convex function (Csiszár, 1967; Moses and Sundaresan, 2011). We obtain different divergences for different forms of the function f, including some well-known divergences. For example, $f(t)=|t-1|/2$ gives Total Variation (TV), $f(t)=t\log t$ gives Kullback-Liebler (KL), $f(t)=(t-1)^2$ gives Chi-square, and $f(t)=(\sqrt{t}-1)^2$ gives squared Hellinger divergences.

Let P^o be a distribution on the space $\mathcal{X}$ and let $l:\mathcal{X}\to\mathbb{R}$ be a loss function. We have the following result from the *distributionally robust optimization* literature, see e.g., Duchi and Namkoong (2018, Proposition 1) and Shapiro (2017, Section 3.2).

Proposition 2. *Let D_f be the f-divergence as defined in (13). Then,*

$$\sup_{D_f(P\|P^o)\leq\rho}\mathbb{E}_P[l(X)]=\inf_{\lambda>0,\eta\in\mathbb{R}}\ \mathbb{E}_{P^o}\left[\lambda f^*\left(\frac{l(X)-\eta}{\lambda}\right)\right]+\lambda\rho+\eta, \tag{14}$$

where $f^*(s)=\sup_{t\geq 0}\{st-f(t)\}$ is the Fenchel conjugate.

Note that on the right hand side of (14), the expectation is taken only with respect to P^o. We will use the above result to derive the dual reformulation of the robust Bellman operator.

B Proof of the Proposition 1

As the first step, we adapt the result given in Proposition 2 in two ways: (i) Since Proposition 1 considers the TV uncertainty set, we will derive the specific form of this result for the TV uncertainty set, (ii) Since Proposition 1 considers the minimization problem instead of the maximization problem, unlike in Proposition 2, we will derive the specific form of this result for minimization.

Lemma 5. *Let D_f be as defined in (13) with $f(t)=|t-1|/2$ corresponding to the TV uncertainty set. Then,*

$$\inf_{D_f(P\|P^o)\leq\rho}\mathbb{E}_P[l(X)]=-\inf_{\eta\in\mathbb{R}}\ \mathbb{E}_{P^o}[(\eta-l(X))_+]+(\eta-\inf_{x\in\mathcal{X}} l(x))_+\times\rho-\eta,$$

Proof. First, we will compute the Fenchel conjugate of $f(t)=|t-1|/2$. We have

$$f^*(s)=\sup_{t\geq 0}\ \{st-\frac{1}{2}|t-1|\}=\max\Big\{\sup_{t\in[0,1]}\{(s+\frac{1}{2})t-\frac{1}{2}\}\ ,\ \sup_{t>1}\{(s-\frac{1}{2})t+\frac{1}{2}\}\Big\}.$$

It is easy to see that for $s>1/2$, we have $f^*(s)=+\infty$, and for $s\leq -1/2$, we have $f^*(s)=-1/2$. For $s\in[-1/2,1/2]$, we have

$$f^*(s)=\max\Big\{\sup_{t\in[0,1]}\{(s+\frac{1}{2})t-\frac{1}{2}\}\ ,\ \sup_{t>1}(\{(s-\frac{1}{2})t+\frac{1}{2}\}\Big\}$$

$$= \max\left\{((s+\frac{1}{2})\cdot 1 - \frac{1}{2}),\ ((s-\frac{1}{2})\cdot 1 + \frac{1}{2})\right\} = s.$$

Thus, we have

$$f^*(s) = \begin{cases} -\frac{1}{2} & s \le -\frac{1}{2}, \\ s & s \in [-\frac{1}{2}, \frac{1}{2}] \\ +\infty & s > \frac{1}{2}. \end{cases}.$$

From Proposition 2, we obtain

$$\begin{aligned}
\sup_{D_f(P\|P^o)\le\rho} \mathbb{E}_P[l(X)] &= \inf_{\lambda>0,\eta\in\mathbb{R}} \mathbb{E}_{P^o}[\lambda f^*(\frac{l(X)-\eta}{\lambda})] + \lambda\rho + \eta \\
&= \inf_{\lambda,\eta:\lambda>0,\eta\in\mathbb{R},\frac{\sup_{x\in\mathcal{X}} l(x)-\eta}{\lambda}\le\frac{1}{2}} \mathbb{E}_{P^o}[\lambda \max\{\frac{l(X)-\eta}{\lambda}, -\frac{1}{2}\}] + \lambda\rho + \eta \\
&= \inf_{\lambda,\eta:\lambda>0,\eta\in\mathbb{R},\frac{\sup_{x\in\mathcal{X}} l(x)-\eta}{\lambda}\le\frac{1}{2}} \mathbb{E}_{P^o}[\max\{l(X)-\eta, -\lambda/2\}] + \lambda\rho + \eta \\
&= \inf_{\lambda,\eta:\lambda>0,\eta\in\mathbb{R},\frac{\sup_{x\in\mathcal{X}} l(x)-\eta}{\lambda}\le\frac{1}{2}} \mathbb{E}_{P^o}[(l(X)-\eta+\lambda/2)_+] - \lambda/2 + \lambda\rho + \eta \\
&= \inf_{\lambda,\eta':\lambda>0,\eta'\in\mathbb{R},\frac{\sup_{x\in\mathcal{X}} l(x)-\eta'}{\lambda}\le 1} \mathbb{E}_{P^o}[(l(X)-\eta')_+] + \lambda\rho + \eta'.
\end{aligned}$$

The second equality follows since $f^*(\frac{l(X)-\eta}{\lambda}) = +\infty$ whenever $\frac{l(X)-\eta}{\lambda} > \frac{1}{2}$, which can be ignored as we are minimizing over λ and η. The fourth equality follows form the fact that $\max\{x, y\} = (x-y)_+ + y$ for any $x, y \in \mathbb{R}$. Finally, the last equality follows by making the substitution $\eta' = \eta - \lambda/2$. Taking the optimal value of λ, i.e., $\lambda = (\sup_{x\in\mathcal{X}} l(x) - \eta')_+$, we get

$$\sup_{D_f(P\|P^o)\le\rho} \mathbb{E}_P[l(X)] = \inf_{\eta\in\mathbb{R}} \mathbb{E}_{P^o}[(l(X)-\eta)_+] + (\sup_{x\in\mathcal{X}} l(x) - \eta)_+\rho + \eta.$$

Now,

$$\begin{aligned}
\inf_{D_f(P\|P^o)\le\rho} \mathbb{E}_P[l(X)] &= -\sup_{D_f(P\|P^o)\le\rho} \mathbb{E}_P[-l(X)] \\
&= -\inf_{\eta\in\mathbb{R}} \mathbb{E}_{P^o}[(-l(X)-\eta)_+] + (\sup_{x\in\mathcal{X}} -l(x) - \eta)_+\rho + \eta \\
&= -\inf_{\eta'\in\mathbb{R}} \mathbb{E}_{P^o}[(\eta' - l(X))_+] + (\eta' - \inf_{x\in\mathcal{X}} l(x))_+\rho - \eta',
\end{aligned}$$

which completes the proof. □

We are now ready to prove Proposition 1.

Proof of Proposition 1. For each (s, a), the optimization problem in (3) is given by $\min_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}[V(s')]$, and our focus is on the setting where $\mathcal{P}_{s,a}$ is given by the TV uncertainty set. So, $\mathcal{P}_{s,a}$ can be equivalently defined using the f-divergence with $f(t) = |t-1|/2$ as $\mathcal{P}_{s,a} = \{P_{s,a} : D_f(P_{s,a}||P^o_{s,a}) \le \rho\}$. We can now use the result of Lemma 5 to get

$$\inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}[V(s')] = -\inf_{\eta\in\mathbb{R}} \mathbb{E}_{s'\sim P^o_{s,a}}[(\eta - V(s'))_+] + (\eta - \inf_{s''\in\mathcal{S}} V(s''))_+\rho - \eta.$$

From Proposition 2, the function $h(\eta) = \mathbb{E}_{s'\sim P^o_{s,a}}[(\eta - V(s'))_+] + \rho(\eta - \inf_{s''} V(s''))_+ - \eta$ is convex in η. Since $V(s') \ge 0$, $h(\eta) = -\eta \ge 0$ when $\eta \le 0$. So, $\inf_{\eta\in(-\infty,0]} h(\eta)$, achieved at $\eta = 0$. Also, since $V(s) \le 1/(1-\gamma)$, we have

$$\begin{aligned}
h(\frac{2}{\rho(1-\gamma)}) &= \mathbb{E}_{s'\sim P^o_{s,a}}[\frac{2}{\rho(1-\gamma)} - V(s')] + \rho(\frac{2}{\rho(1-\gamma)} - \inf_{s''} V(s'')) - \frac{2}{\rho(1-\gamma)} \\
&\ge -\frac{1}{(1-\gamma)} + \rho(\frac{2}{\rho(1-\gamma)} - \frac{1}{(1-\gamma)}) = \frac{2}{(1-\gamma)} - \frac{(1+\rho)}{(1-\gamma)} \ge 0.
\end{aligned}$$

So, it is sufficient to consider $\eta \in [0, \frac{2}{\rho(1-\gamma)}]$ for the above optimization problem.

Using these, we get

$$(TQ)(s,a) = r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{P}_{s,a}} \mathbb{E}_{s' \sim P_{s,a}}[V(s')]$$
$$= r(s,a) + \gamma \cdot -1 \cdot \inf_{\eta \in \eta \in [0, \frac{2}{\rho(1-\gamma)}]} \mathbb{E}_{s' \sim P^o_{s,a}}[(\eta - V(s'))_+] + (\eta - \inf_{s'' \in \mathcal{S}} V(s''))_+ \rho - \eta.$$

This completes the proof of Proposition 1. □

C Proof of Theorem 1

We start by proving Lemma 1 which mainly follows from Lemma 4 in Appendix A.

Proof of Lemma 1. Let $h((s,a),\eta) = \mathbb{E}_{s' \sim P^o_{s,a}}((\eta - \max_{a'} f(s',a'))_+ - (1-\rho)\eta)$. We note that $h((s,a),\eta)$ is $\Sigma(\mathcal{S} \times \mathcal{A})$-measurable in $(s,a) \in \mathcal{S} \times \mathcal{A}$ for each $\eta \in [0, 1/(\rho(1-\gamma))]$ and is continuous in η for each $(s,a) \in \mathcal{S} \times \mathcal{A}$. Now it follows that $h((s,a),\eta)$ is a normal integrand (see Definition 1 in Appendix A). We now note that $L^1(\mathcal{S} \times \mathcal{A}, \Sigma(\mathcal{S} \times \mathcal{A}), \mu)$ is a decomposable space (Remark 4 in Appendix A). Thus, this lemma now directly follows from Lemma 4. □

Now we state a result and provide its proof for the empirical risk minimization on the dual parameter.

Lemma 6 (Dual Optimization Error Bound)**.** *Let $\widehat{g}_f$ be the dual optimization parameter from the algorithm (Step 4) for the state-action value function f and let T_g be as defined in (9). With probability at least $1-\delta$, we have*

$$\sup_{f \in \mathcal{F}} \|Tf - T_{\widehat{g}_f} f\|_{1,\mu} \leq \frac{4\gamma(2-\rho)}{\rho(1-\gamma)} \sqrt{\frac{2\log(|\mathcal{G}|)}{N}} + \frac{25\gamma}{\rho(1-\gamma)} \sqrt{\frac{2\log(8|\mathcal{F}|/\delta)}{N}} + \gamma \varepsilon_{dual}.$$

Proof. Fix an $f \in \mathcal{F}$. We will also invoke union bound for the supremum here. We recall from (8) that $\widehat{g}_f = \arg\min_{g \in \mathcal{G}} \widehat{L}_{\text{dual}}(g; f)$. From the robust Bellman equation, we directly obtain

$$\|T_{\widehat{g}_f} f - Tf\|_{1,\mu} = \gamma(\mathbb{E}_{s,a \sim \mu} | \mathbb{E}_{s' \sim P^o_{s,a}}((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)\widehat{g}_f(s,a))$$
$$- \inf_{\eta \in [0, 2/(\rho(1-\gamma))]} \mathbb{E}_{s' \sim P^o_{s,a}}((\eta - \max_{a'} f(s',a'))_+ - (1-\rho)\eta)|)$$
$$\overset{(a)}{=} \gamma(\mathbb{E}_{s,a \sim \mu} \mathbb{E}_{s' \sim P^o_{s,a}}((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)\widehat{g}_f(s,a))$$
$$- \mathbb{E}_{s,a \sim \mu}[\inf_{\eta \in [0, 2/(\rho(1-\gamma))]} \mathbb{E}_{s' \sim P^o_{s,a}}((\eta - \max_{a'} f(s',a'))_+ - (1-\rho)\eta)])$$
$$\overset{(b)}{=} \gamma(\mathbb{E}_{s,a \sim \mu, s' \sim P^o_{s,a}}((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)\widehat{g}_f(s,a))$$
$$- \inf_{g \in L^1} \mathbb{E}_{s,a \sim \mu, s' \sim P^o_{s,a}}((g(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)g(s,a)))$$
$$= \gamma(\mathbb{E}_{s,a \sim \mu, s' \sim P^o_{s,a}}((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)\widehat{g}_f(s,a))$$
$$- \inf_{g \in \mathcal{G}} \mathbb{E}_{s,a \sim \mu, s' \sim P^o_{s,a}}((g(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)g(s,a)))$$
$$+ \gamma(\inf_{g \in \mathcal{G}} \mathbb{E}_{s,a \sim \mu, s' \sim P^o_{s,a}}((g(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)g(s,a))$$
$$- \inf_{g \in L^1} \mathbb{E}_{s,a \sim \mu, s' \sim P^o_{s,a}}((g(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)g(s,a)))$$
$$\overset{(c)}{\leq} \gamma(\mathbb{E}_{s,a \sim \mu, s' \sim P^o_{s,a}}((\widehat{g}_f(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)\widehat{g}_f(s,a))$$
$$- \inf_{g \in \mathcal{G}} \mathbb{E}_{s,a \sim \mu, s' \sim P^o_{s,a}}((g(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)g(s,a))) + \gamma \varepsilon_{\text{dual}}$$
$$\overset{(d)}{\leq} 2\gamma R(l \circ \mathcal{G} \circ \mathcal{D}) + \frac{25\gamma}{\rho(1-\gamma)} \sqrt{\frac{2\log(8/\delta)}{N}} + \gamma \varepsilon_{\text{dual}}$$

$$\overset{(e)}{\leq} \frac{4\gamma(2-\rho)}{\rho(1-\gamma)}\sqrt{\frac{2\log(|\mathcal{G}|)}{N}} + \frac{25\gamma}{\rho(1-\gamma)}\sqrt{\frac{2\log(8/\delta)}{N}} + \gamma\varepsilon_{\text{dual}}.$$

(a) follows since $\inf_g h(g) \leq h(\widehat{g}_f)$. (b) follows from Lemma 1. (c) follows from the approximate dual realizability assumption (Assumption 4).

For (d), we consider the loss function $l(g,(s,a,s')) = (g(s,a) - \max_{a'} f(s',a'))_+ - (1-\rho)g(s,a)$ and dataset $\mathcal{D} = \{s_i, a_i, s_i'\}_{i=1}^N$. Note that $|l(g,(s,a,s'))| \leq 5/(\rho(1-\gamma))$ (since $f \in \mathcal{F}$ and $g \in \mathcal{G}$). Now, we can apply the empirical risk minimization result (11) in Lemma 3 to get (d), where $R(\cdot)$ is the Rademacher complexity.

Finally, (e) follows from (12) in Lemma 3 when combined with the facts that $l(g,(s,a,s'))$ is $(2-\rho)$-Lipschitz in g and $g(s,a) \leq 2/(\rho(1-\gamma))$, since $g \in \mathcal{G}$.

With union bound, with probability at least $1-\delta$, we finally get

$$\sup_{f\in\mathcal{F}} \|Tf - T_{\widehat{g}_f} f\|_{1,\mu} \leq \frac{4\gamma(2-\rho)}{\rho(1-\gamma)}\sqrt{\frac{2\log(|\mathcal{G}|)}{N}} + \frac{25\gamma}{\rho(1-\gamma)}\sqrt{\frac{2\log(8|\mathcal{F}|/\delta)}{N}} + \gamma\varepsilon_{\text{dual}},$$

which concludes the proof. □

We next prove the least-squares generalization bound for the RFQI algorithm.

Lemma 7 (Least squares generalization bound)**.** *Let $\widehat{f}_g$ be the least-squares solution from the algorithm (Step 5) for the state-action value function f and dual variable function g. Let T_g be as defined in (9). Then, with probability at least $1-\delta$, we have*

$$\sup_{f\in\mathcal{F}}\sup_{g\in\mathcal{G}} \|T_g f - \widehat{f}_g\|_{2,\mu} \leq \sqrt{6\varepsilon_c} + \frac{16}{\rho(1-\gamma)}\sqrt{\frac{18\log(2|\mathcal{F}||\mathcal{G}|/\delta)}{N}}.$$

Proof. We adapt the least-squares generalization bound given in Agarwal et al. (2019, Lemma A.11) to our setting. We recall from (10) that $\widehat{f}_g = \arg\min_{Q\in\mathcal{F}} \widehat{L}_{\text{RFQI}}(Q; f, g)$. We first fix functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$. For any function $f' \in \mathcal{F}$, we define random variables $z_i^{f'}$ as

$$z_i^{f'} = (f'(s_i,a_i) - y_i)^2 - ((T_g f)(s_i,a_i) - y_i)^2,$$

where $y_i = r_i - \gamma(g(s_i,a_i) - \max_{a'} f(s_i',a'))_+ + \gamma(1-\rho)g(s_i,a_i)$, and $(s_i,a_i,s_i') \in \mathcal{D}$ with $(s_i,a_i) \sim \mu, s_i' \sim P^o_{s_i,a_i}$. It is straightforward to note that for a given (s_i,a_i), we have $\mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[y_i] = (T_g f)(s_i,a_i)$.

Also, since $g(s_i,a_i) \leq 2/(\rho(1-\gamma))$ (because $g \in \mathcal{G}$) and $f(s_i,a_i), f'(s_i,a_i) \leq 1/(1-\gamma)$ (because $f, f' \in \mathcal{F}$), we have $(T_g f)(s_i,a_i) \leq 5/(\rho(1-\gamma))$. This also gives us that $y_i \leq 5/(\rho(1-\gamma))$.

Using this, we obtain the first moment and an upper-bound for the second moment of $z_i^{f'}$ as follows:

$$\begin{aligned}
\mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[z_i^{f'}] &= \mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[(f'(s_i,a_i) - (T_g f)(s_i,a_i)) \cdot (f'(s_i,a_i) + (T_g f)(s_i,a_i) - 2y_i)] \\
&= (f'(s_i,a_i) - (T_g f)(s_i,a_i))^2, \\
\mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[(z_i^{f'})^2] &= \mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[(f'(s_i,a_i) - (T_g f)(s_i,a_i))^2 \cdot (f'(s_i,a_i) + (T_g f)(s_i,a_i) - 2y_i)^2] \\
&= (f'(s_i,a_i) - (T_g f)(s_i,a_i))^2 \cdot \mathbb{E}_{s_i'\sim P^o_{s_i,a_i}}[(f'(s_i,a_i) + (T_g f)(s_i,a_i) - 2y_i)^2] \\
&\leq C_1 (f'(s_i,a_i) - (T_g f)(s_i,a_i))^2,
\end{aligned}$$

where $C_1 = 16^2/(\rho^2(1-\gamma)^2)$. This immediately implies that

$$\begin{aligned}
\mathbb{E}_{s_i,a_i\sim\mu, s_i'\sim P^o_{s_i,a_i}}[z_i^{f'}] &= \|T_g f - f'\|_{2,\mu}^2, \\
\mathbb{E}_{s_i,a_i\sim\mu, s_i'\sim P^o_{s_i,a_i}}[(z_i^{f'})^2] &\leq C_1 \|T_g f - f'\|_{2,\mu}^2.
\end{aligned}$$

From these calculations, it is also straightforward to see that $|z_i^{f'} - \mathbb{E}_{s_i,a_i\sim\mu, s_i'\sim P^o_{s_i,a_i}}[z_i^{f'}]| \leq 2C_1$ almost surely.

Now, using the Bernstein's inequality (Lemma 2), together with a union bound over all $f' \in \mathcal{F}$, with probability at least $1-\delta$, we have

$$|\|T_g f - f'\|_{2,\mu}^2 - \frac{1}{N}\sum_{i=1}^N z_i^{f'}| \leq \sqrt{\frac{2C_1\|T_g f - f'\|_{2,\mu}^2 \log(2|\mathcal{F}|/\delta)}{N}} + \frac{2C_1\log(2|\mathcal{F}|/\delta)}{3N}, \quad (15)$$

for all $f' \in \mathcal{F}$. Setting $f' = \widehat{f}_g$, with probability at least $1-\delta/2$, we have

$$\|T_g f - \widehat{f}_g\|_{2,\mu}^2 \leq \frac{1}{N}\sum_{i=1}^N z_i^{\widehat{f}_g} + \sqrt{\frac{2C_1\|T_g f - \widehat{f}_g\|_{2,\mu}^2 \log(4|\mathcal{F}|/\delta)}{N}} + \frac{2C_1\log(4|\mathcal{F}|/\delta)}{3N}. \quad (16)$$

Now we upper-bound $(1/N)\sum_{i=1}^N z_i^{\widehat{f}_g}$ in the following. Consider a function $\widetilde{f} \in \arg\min_{h\in\mathcal{F}} \|h - T_g f\|_{2,\mu}^2$. Note that $\widetilde{f}$ is independent of the dataset. We note that our earlier first and second moment calculations hold true for $\widetilde{f}$, replacing f', as well. Now, from (15) setting $f' = \widetilde{f}$, with probability at least $1-\delta/2$ we have

$$\frac{1}{N}\sum_{i=1}^N z_i^{\widetilde{f}} - \|T_g f - \widetilde{f}\|_{2,\mu}^2 \leq \sqrt{\frac{2C_1\|T_g f - \widetilde{f}\|_{2,\mu}^2 \log(4|\mathcal{F}|/\delta)}{N}} + \frac{2C_1\log(4|\mathcal{F}|/\delta)}{3N}. \quad (17)$$

Suppose $(1/N)\sum_{i=1}^N z_i^{\widetilde{f}} \geq 2C_1\log(4|\mathcal{F}|/\delta)/N$ holds, then from (17) we get

$$\frac{1}{N}\sum_{i=1}^N z_i^{\widetilde{f}} - \|T_g f - \widetilde{f}\|_{2,\mu}^2 \leq \sqrt{\|T_g f - \widetilde{f}\|_{2,\mu}^2 \cdot \frac{1}{N}\sum_{i=1}^N z_i^{\widetilde{f}}} + \frac{2C_1\log(4|\mathcal{F}|/\delta)}{N}. \quad (18)$$

We note the following algebra fact: Suppose $x^2 - ax + b \leq 0$ with $b > 0$ and $a^2 \geq 4b$, then we have $x \leq a$. Taking $x = (1/N)\sum_{i=1}^N z_i^{\widetilde{f}}$ in this fact, from (18) we get

$$\frac{1}{N}\sum_{i=1}^N z_i^{\widetilde{f}} \leq 3\|T_g f - \widetilde{f}\|_{2,\mu}^2 + \frac{4C_1\log(4|\mathcal{F}|/\delta)}{3N} \leq 3\|T_g f - \widetilde{f}\|_{2,\mu}^2 + \frac{2C_1\log(4|\mathcal{F}|/\delta)}{N}. \quad (19)$$

Now suppose $(1/N)\sum_{i=1}^N z_i^{\widetilde{f}} \leq 2C_1\log(4|\mathcal{F}|/\delta)/N$, then (19) holds immediately. Thus, (19) always holds with probability at least $1-\delta/2$. Furthermore, recall $\widetilde{f} \in \arg\min_{h\in\mathcal{F}} \|h - T_g f\|_{2,\mu}^2$, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N z_i^{\widetilde{f}} &\leq 3\|T_g f - \widetilde{f}\|_{2,\mu}^2 + \frac{2C_1\log(4|\mathcal{F}|/\delta)}{N} \\ &= 3\min_{h\in\mathcal{F}}\|h - T_g f\|_{2,\mu}^2 + \frac{2C_1\log(4|\mathcal{F}|/\delta)}{N} \leq 3\varepsilon_{\mathrm{c}} + \frac{2C_1\log(4|\mathcal{F}|/\delta)}{N}, \end{aligned} \quad (20)$$

where the last inequality follows from the approximate robust Bellman completion assumption (Assumption 2).

We note that since $\widehat{f}_g$ is the least-squares regression solution, we know that $(1/N)\sum_{i=1}^N z_i^{\widehat{f}_g} \leq (1/N)\sum_{i=1}^N z_i^{\widetilde{f}}$. With this note in (20), from (16), with probability at least $1-\delta$, we have

$$\begin{aligned}\|T_g f - \widehat{f}_g\|_{2,\mu}^2 &\leq 3\varepsilon_{\mathrm{c}} + \frac{2C_1\log(4|\mathcal{F}|/\delta)}{N} \\ &\quad + \sqrt{\frac{2C_1\|T_g f - \widehat{f}_g\|_{2,\mu}^2 \log(4|\mathcal{F}|/\delta)}{N}} + \frac{2C_1\log(4|\mathcal{F}|/\delta)}{3N} \\ &\leq 3\varepsilon_{\mathrm{c}} + \frac{3C_1\log(4|\mathcal{F}|/\delta)}{N} + \sqrt{\frac{3C_1\|T_g f - \widehat{f}_g\|_{2,\mu}^2 \log(4|\mathcal{F}|/\delta)}{N}}. \end{aligned}$$

From the earlier algebra fact, taking $x = \|T_g f - \widehat{f}_g\|_{2,\mu}^2$, with probability at least $1-\delta$, we have

$$\|T_g f - \widehat{f}_g\|_{2,\mu}^2 \leq 6\varepsilon_{\mathrm{c}} + \frac{9C_1 \log(4|\mathcal{F}|/\delta)}{N}.$$

From the fact $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, with probability at least $1-\delta$, we get

$$\|T_g f - \widehat{f}_g\|_{2,\mu} \leq \sqrt{6\varepsilon_{\mathrm{c}}} + \sqrt{\frac{9C_1 \log(4|\mathcal{F}|/\delta)}{N}}.$$

Using union bound for $f \in \mathcal{F}$ and $g \in \mathcal{G}$, with probability at least $1-\delta$, we finally obtain

$$\sup_{f\in\mathcal{F}} \sup_{g\in\mathcal{G}} \|T_g f - \widehat{f}_g\|_{2,\mu} \leq \sqrt{6\varepsilon_{\mathrm{c}}} + \sqrt{\frac{18C_1 \log(2|\mathcal{F}||\mathcal{G}|/\delta)}{N}},$$

which completes the least-squares generalization bound analysis. □

We are now ready to prove the main theorem.

Proof of Theorem 1. We let $V_k(s) = Q_k(s, \pi_k(s))$ for every $s \in \mathcal{S}$. Since π_k is the greedy policy w.r.t Q_k, we also have $V_k(s) = Q_k(s, \pi_k(s)) = \max_a Q_k(s,a)$. We recall that $V^* = V^{\pi^*}$ and $Q^* = Q^{\pi^*}$. We also recall from Section 2 that Q^{π^*} is a fixed-point of the robust Bellman operator T defined in (3). We also note that the same holds true for any stationary deterministic policy π from Iyengar (2005) that Q^π satisfies $Q^\pi(s,a) = r(s,a) + \gamma \min_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}[V^\pi(s')]$. We can now further use the dual form (5) under Assumption 3. We first characterize the performance decomposition between V^{π^*} and V^{π_K}. For a given $s_0 \in \mathcal{S}$, we observe that

$$\begin{aligned}
&V^{\pi^*}(s_0) - V^{\pi_K}(s_0) = (V^{\pi^*}(s_0) - V_K(s_0)) - (V^{\pi_K}(s_0) - V_K(s_0)) \\
&= (Q^{\pi^*}(s_0, \pi^*(s_0)) - Q_K(s_0, \pi_K(s_0))) - (Q^{\pi_K}(s_0, \pi_K(s_0)) - Q_K(s_0, \pi_K(s_0))) \\
&\overset{(a)}{\leq} Q^{\pi^*}(s_0, \pi^*(s_0)) - Q_K(s_0, \pi^*(s_0)) + Q_K(s_0, \pi_K(s_0)) - Q^{\pi_K}(s_0, \pi_K(s_0)) \\
&= Q^{\pi^*}(s_0, \pi^*(s_0)) - Q_K(s_0, \pi^*(s_0)) + Q_K(s_0, \pi_K(s_0)) - Q^{\pi^*}(s_0, \pi_K(s_0)) \\
&\qquad + Q^{\pi^*}(s_0, \pi_K(s_0)) - Q^{\pi_K}(s_0, \pi_K(s_0)) \\
&\overset{(b)}{\leq} Q^{\pi^*}(s_0, \pi^*(s_0)) - Q_K(s_0, \pi^*(s_0)) + Q_K(s_0, \pi_K(s_0)) - Q^{\pi^*}(s_0, \pi_K(s_0)) \\
&\qquad + \gamma \sup_\eta (\mathbb{E}_{s_1\sim P^o_{s_0,\pi_K(s_0)}}((\eta - V^{\pi_K}(s_1))_+ - (\eta - V^{\pi^*}(s_1))_+)) \\
&\overset{(c)}{\leq} |Q^{\pi^*}(s_0, \pi^*(s_0)) - Q_K(s_0, \pi^*(s_0))| + |Q^{\pi^*}(s_0, \pi_K(s_0)) - Q_K(s_0, \pi_K(s_0))| \\
&\qquad + \gamma \mathbb{E}_{s_1\sim P^o_{s_0,\pi_K(s_0)}}(|V^{\pi^*}(s_1) - V^{\pi_K}(s_1)|).
\end{aligned}$$

(a) follows from the fact that π_K is the greedy policy with respect to Q_K. (b) follows from the Bellman optimality equations and the fact $|\sup_x f(x) - \sup_x g(x)| \leq \sup_x |f(x) - g(x)|$. Finally, (c) follows from the facts $(x)_+ - (y)_+ \leq (x-y)_+$ and $(x)_+ \leq |x|$ for any $x, y \in \mathbb{R}$.

We now recall the initial state distribution d_0. Thus, we have

$$\begin{aligned}
&\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}] - \mathbb{E}_{s_0\sim d_0}[V^{\pi_K}] \leq \\
&\quad \mathbb{E}_{s_0\sim d_0}\Big[|Q^{\pi^*}(s_0, \pi^*(s_0)) - Q_K(s_0, \pi^*(s_0))| + |Q^{\pi^*}(s_0, \pi_K(s_0)) - Q_K(s_0, \pi_K(s_0))| \\
&\qquad + \gamma \mathbb{E}_{s_1\sim P^o_{s_0,\pi_K(s_0)}}(|V^{\pi^*}(s_1) - V^{\pi_K}(s_1)|)\Big].
\end{aligned}$$

Since $V^{\pi^*}(s) \geq V^{\pi_K}(s)$ for any $s \in \mathcal{S}$, by telescoping we get

$$\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}] - \mathbb{E}_{s_0\sim d_0}[V^{\pi_K}] \leq \sum_{h=0}^{\infty} \gamma^h \times$$

$$\left(\mathbb{E}_{s\sim d_{h,\pi_K}}[|Q^{\pi^*}(s,\pi^*(s)) - Q_K(s,\pi^*(s))| + |Q^{\pi^*}(s,\pi_K(s)) - Q_K(s,\pi_K(s))|]\right), \quad (21)$$

where $d_{h,\pi_K} \in \Delta(\mathcal{S})$ for all natural numbers $h \geq 0$ is defined as

$$d_{h,\pi_K} = \begin{cases} d_0 & \text{if } h = 0, \\ P^o_{s',\pi_K(s')} & \text{otherwise, with } s' \sim d_{h-1,\pi_K}. \end{cases}$$

We emphasize that the state distribution d_{h,π_K}'s are different from the discounted state-action occupancy distributions. We note that a similar state distribution proof idea is used in Agarwal et al. (2019).

Recall $\|f\|^2_{p,\nu} = (\mathbb{E}_{s,a\sim\nu}|f(s,a)|^p)^{1/p}$, where $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$. With this we have

$$\mathbb{E}_{s_0\sim d_0}[V^{\pi^*}] - \mathbb{E}_{s_0\sim d_0}[V^{\pi_K}] \leq \sum_{h=0}^{\infty}\gamma^h\left(\|Q^{\pi^*} - Q_K\|_{1,d_{h,\pi_K}\circ\pi^*} + \|Q^{\pi^*} - Q_K\|_{1,d_{h,\pi_K}\circ\pi_K}\right), \quad (22)$$

where the state-action distributions $d_{h,\pi_K}\circ\pi^*(s,a) \propto d_{h,\pi_K}(s)\mathbb{1}\{a = \pi^*(s)\}$ and $d_{h,\pi_K}\circ\pi_K(s,a) \propto d_{h,\pi_K}(s)\mathbb{1}\{a = \pi_K(s)\}$ directly follows by comparing with (21).

We now bound one of the RHS terms above by bounding for any state-action distribution ν satisfying Assumption 1 (in particular the following bound is true for $d_{h,\pi_K}\circ\pi^*$ or $d_{h,\pi_K}\circ\pi_K$ in (21)):

$$\begin{aligned}
&\|Q^{\pi^*} - Q_K\|_{1,\nu} \leq \|Q^{\pi^*} - TQ_{K-1}\|_{1,\nu} + \|TQ_{K-1} - Q_K\|_{1,\nu} \\
&\overset{(a)}{\leq} \|Q^{\pi^*} - TQ_{K-1}\|_{1,\nu} + \sqrt{C}\|TQ_{K-1} - Q_K\|_{1,\mu} \\
&= (\mathbb{E}_{s,a\sim\nu}|Q^{\pi^*}(s,a) - TQ_{K-1}(s,a)|) + \sqrt{C}\|TQ_{K-1} - Q_K\|_{1,\mu} \\
&\overset{(b)}{\leq} (\mathbb{E}_{s,a\sim\nu}\gamma\sup_{\eta}|\mathbb{E}_{s'\sim P^o_{s,a}}((\eta - \max_{a'}Q_{K-1}(s',a'))_+ - (\eta - \max_{a'}Q^{\pi^*}(s',a'))_+)|) \\
&\qquad\qquad + \sqrt{C}\|TQ_{K-1} - Q_K\|_{1,\mu} \\
&\overset{(c)}{\leq} (\mathbb{E}_{s,a\sim\nu}|\mathbb{E}_{s'\sim P^o_{s,a}}(\max_{a'}Q^{\pi^*}(s',a') - \max_{a'}Q_{K-1}(s',a'))_+|) + \sqrt{C}\|TQ_{K-1} - Q_K\|_{1,\mu} \\
&\overset{(d)}{\leq} \gamma(\mathbb{E}_{s,a\sim\nu}\mathbb{E}_{s'\sim P^o_{s,a}}\max_{a'}|Q^{\pi^*}(s',a') - Q_{K-1}(s',a')|) + \sqrt{C}\|TQ_{K-1} - Q_K\|_{1,\mu} \\
&\overset{(e)}{\leq} \gamma\|Q^{\pi^*} - Q_{K-1}\|_{1,\nu'} + \sqrt{C}\|TQ_{K-1} - Q_K\|_{1,\mu} \\
&\overset{(f)}{\leq} \gamma\|Q^{\pi^*} - Q_{K-1}\|_{1,\nu'} + \sqrt{C}\|T_{g_{K-1}}Q_{K-1} - Q_K\|_{2,\mu} + \sqrt{C}\|TQ_{K-1} - T_{g_{K-1}}Q_{K-1}\|_{1,\mu},
\end{aligned} \quad (23)$$

where (a) follows by the concentratability assumption (Assumption 1), (b) from Bellman equation, operator T, and the fact $|\sup_x p(x) - \sup_x q(x)| \leq \sup_x |p(x) - q(x)|$, (c) from the fact $|(x)_+ - (y)_+| \leq |(x-y)_+|$ for any $x, y \in \mathbb{R}$, (d) follows by Jensen's inequality and by the facts $|\sup_x p(x) - \sup_x q(x)| \leq \sup_x |p(x)-q(x)|$ and $(x)_+ \leq |x|$ for any $x, y \in \mathbb{R}$, and (e) by defining the distribution ν' as $\nu'(s',a') = \sum_{s,a}\nu(s,a)P^o_{s,a}(s')\mathbb{1}\{a' = \arg\max_b |Q^{\pi^*}(s',b) - Q_{K-1}(s',b)|\}$, and (f) using the fact that $\|\cdot\|_{1,\mu} \leq \|\cdot\|_{2,\mu}$.

Now, by recursion until iteration 0, we get

$$\begin{aligned}
\|Q^{\pi^*} - Q_K\|_{1,\nu} &\leq \gamma^K\sup_{\bar{\nu}}\|Q^{\pi^*} - Q_0\|_{1,\bar{\nu}} + \sqrt{C}\sum_{t=0}^{K-1}\gamma^t\|TQ_{K-1-t} - T_{g_{K-1-t}}Q_{K-1-t}\|_{1,\mu} \\
&\qquad + \sqrt{C}\sum_{t=0}^{K-1}\gamma^t\|T_{g_{K-1-t}}Q_{K-1-t} - Q_{K-t}\|_{2,\mu} \\
&\overset{(a)}{\leq} \frac{\gamma^K}{1-\gamma} + \sqrt{C}\sum_{t=0}^{K-1}\gamma^t\|TQ_{K-1-t} - T_{g_{K-1-t}}Q_{K-1-t}\|_{1,\mu}
\end{aligned}$$

$$+\sqrt{C}\sum_{t=0}^{K-1}\gamma^t\|T_{g_{K-1-t}}Q_{K-1-t}-Q_{K-t}\|_{2,\mu}$$

$$\overset{(b)}{\leq}\frac{\gamma^K}{1-\gamma}+\frac{\sqrt{C}}{1-\gamma}\sup_{f\in\mathcal{F}}\|Tf-T_{\widehat{g}_f}f\|_{1,\mu}+\frac{\sqrt{C}}{1-\gamma}\sup_{f\in\mathcal{F}}\|T_{\widehat{g}_f}f-\widehat{f}_{\widehat{g}_f}\|_{2,\mu}$$

$$\leq\frac{\gamma^K}{1-\gamma}+\frac{\sqrt{C}}{1-\gamma}\sup_{f\in\mathcal{F}}\|Tf-T_{\widehat{g}_f}f\|_{1,\mu}+\frac{\sqrt{C}}{1-\gamma}\sup_{f\in\mathcal{F}}\sup_{g\in\mathcal{G}}\|T_gf-\widehat{f}_g\|_{2,\mu}. \tag{24}$$

where (a) follows since $|Q^{\pi^*}(s,a)|\leq 1/(1-\gamma), Q_0(s,a)=0$, and (b) follows since $\widehat{g}_f$ is the dual variable function from the algorithm for the state-action value function f and $\widehat{f}_g$ as the least squares solution from the algorithm for the state-action value function f and dual variable function g pair.

The proof is now complete combining (22) and (24) with Lemma 6 and Lemma 7. □

D Related Works

Here we provide a more detailed description of the related work to complement what we listed in the introduction (Section 1).

Offline RL: The problem of learning the optimal policy only using an offline dataset is first addressed under the generative model assumption (Singh and Yee, 1994; Azar et al., 2013; Haskell et al., 2016; Sidford et al., 2018; Agarwal et al., 2020; Li et al., 2020; Kalathil et al., 2021). This assumption requires generating the same uniform number of next-state samples for each and every state-action pairs. To account for large state spaces, there are number of works (Antos et al., 2008; Bertsekas, 2011; Lange et al., 2012; Chen and Jiang, 2019; Xie and Jiang, 2020; Levine et al., 2020; Xie et al., 2021) that utilize function approximation under similar assumption, concentratability assumption (Chen and Jiang, 2019) in which the data distribution μ sufficiently covers the discounted state-action occupancy. There is rich literature (Munos and Szepesvári, 2008; Farahmand et al., 2010; Lazaric et al., 2012; Chen and Jiang, 2019; Liu et al., 2020; Xie et al., 2021) in the conquest of identifying and improving these necessary and sufficient assumptions for offline RL that use variations of Fitted Q-Iteration (FQI) algorithm (Gordon, 1995; Ernst et al., 2005). There is also rich literature (Fujimoto et al., 2019; Kumar et al., 2019, 2020; Yu et al., 2020; Zhang and Jiang, 2021) that develop offline deep RL algorithms focusing on the algorithmic and empirical aspects and propose multitude heuristic approaches to advance the field. All these results assume that the offline data is generated according to a single model and the goal is to find the optimal policy for the MDP with the same model. In particular, none of these works consider the *offline robust RL problem* where the offline data is generated according to a (training) model which can be different from the one in testing, and the goal is to learn a policy that is robust w.r.t. an uncertainty set.

Robust RL: To address the parameter uncertainty problem, Iyengar (2005) and Nilim and El Ghaoui (2005) introduced the RMDP framework. Iyengar (2005) showed that the optimal robust value function and policy can be computed using the robust counterparts of the standard value iteration and policy iteration algorithms. To tackle the parameter uncertainty problem, other works considered distributionally robust setting (Xu and Mannor, 2010), modified policy iteration (Kaufman and Schaefer, 2013), and more general uncertainty set (Wiesemann et al., 2013). These initial works mainly focused on the planning problem (known transition probability dynamics) in the tabular setting. Tamar et al. (2014) proposed linear function approximation method to solve large RMDPs. Though this work suggests a sampling based approach, a general model-free learning algorithm and analysis was not included. Roy et al. (2017) proposed the robust versions of the classical model-free reinforcement learning algorithms, such as Q-learning, SARSA, and TD-learning in the tabular setting. They also proposed function approximation based algorithms for the policy evaluation. However, this work does not have a policy iteration algorithm with provable guarantees for learning the optimal robust policy. Derman et al. (2018) introduced soft-robust actor-critic algorithms using neural networks, but does not provide any global convergence guarantees for the learned policy. Tessler et al. (2019) proposed a min-max game framework to address the robust learning problem focusing on the

tabular setting. Lim and Autef (2019) proposed a kernel-based RL algorithm for finding the robust value function in a batch learning setting. Mankowitz et al. (2020) employed an entropy-regularized policy optimization algorithm for continuous control using neural network, but does not provide any provable guarantees for the learned policy. Panaganti and Kalathil (2021) proposed least-squares policy iteration method to handle large state-action space in robust RL, but only provide asymptotic policy evaluation convergence guarantees whereas Panaganti and Kalathil (2021) provide finite time convergence for the policy iteration to optimal robust value.

Other robust RL related works: Robust control is a well-studied area in the classical control theory (Zhou et al., 1996; Dullerud and Paganini, 2013). Recently, there are some interesting works that address the robust RL problem using this framework, especially focusing on the linear quadratic regulator setting (Zhang et al., 2020b). Risk sensitive RL algorithms (Borkar, 2002; Prashanth and Ghavamzadeh, 2016; Fei et al., 2021) and adversarial RL algorithms (Pinto et al., 2017; Zhang et al., 2021; Huang et al., 2022) also address the robustness problem implicitly under different frameworks which are independent from RMDPs. Our framework and approach of robust MDP is significantly different from these line of works.

The works that are closest to ours are by Zhou et al. (2021); Yang et al. (2021); Panaganti and Kalathil (2022) that address the robust RL problem in a tabular setting under the generative model assumption. Due to the generative model assumption, the offline data has the same uniform number of samples corresponding to each and every state-action pair, and tabular setting allows the estimation of the uncertainty set followed by solving the planning problem. Our work is significantly different from these in the following way: (i) we consider a robust RL problem with arbitrary large state space, instead of the small tabular setting, (ii) we consider a true offline RL setting where the state-action pairs are sampled according to an arbitrary distribution, instead of using the generative model assumption, (iii) we focus on a function approximation approach where the goal is to directly learn optimal robust value/policy using function approximation techniques, instead of solving the tabular planning problem with the estimated model. *To the best of our knowledge, this is the first work that addresses the offline robust RL problem with arbitrary large state space using function approximation, with provable guarantees on the performance of the learned policy.*

E Experiment Details

We provide more detailed and practical version of our RFQI algorithm (Algorithm 1) in this section. We also provide more experimental results evaluated on *Cartpole*, *Hopper*, and *Half-Cheetah* OpenAI Gym Mujoco (Brockman et al., 2016) environments.

We provide our code in an **anonymous github webpage** `https://github.com/curious-beaver/RFQI` containing instructions to reproduce all results in this paper. We implemented our RFQI algorithm based on the architecture of Batch Constrained deep Q-learning (BCQ) algorithm (Fujimoto et al., 2019) [2] and Pessimistic Q-learning (PQL) algorithm (Liu et al., 2020) [3]. We note that PQL algorithm (with $b = 0$ filtration thresholding (Liu et al., 2020)) and BCQ algorithm are the practical versions of FQI algorithm with neural network architecture.

E.1 RFQI Practical Algorithm

We provide the practical version of our RFQI algorithm in Algorithm 2 and highlight the difference with BCQ and PQL algorithms in blue (steps 8 and 9).

RFQI algorithm implementation details: The Variational Auto-Encoder (VAE) G^a_ω (Kingma and Welling, 2013) is defined by two networks, an encoder $E_{\omega_1}(s, a)$ and decoder $D_{\omega_2}(s, z)$, where $\omega = \{\omega_1, \omega_2\}$. The encoder outputs mean and standard deviation, $(\mu, \sigma) = E_{\omega_1}(s, a)$, of a normal distribution. A latent vector z is sampled from the standard normal distribution and for a state s, the decoder maps them to an action $D_{\omega_2} : (s, z) \mapsto \tilde{a}$. Then the evidence lower bound $(ELBO)$ of

[2] Available at `https://github.com/sfujim/BCQ`

[3] Available at `https://github.com/yaoliucs/PQL`

Algorithm 2 RFQI Practical Algorithm

1: **Input:** Offline dataset $\mathcal{D}$, radius of robustness ρ, maximum perturbation Φ, target update rate τ, mini-batch size N, maximum number of iterations K, number of actions u.
2: **Initialize:** Two state-action neural networks Q_{θ_1} and Q_{θ_2}, one dual neural network g_{θ_3} policy (perturbation) model: $\xi_\varphi \in [-\Phi, \Phi]$), and action VAE G^a_ω, with random parameters θ_1, θ_2, φ, ω, and target networks $Q_{\theta'_1}, Q_{\theta'_2}, \xi_{\varphi'}$ with $\theta'_1 \leftarrow \theta_1, \theta'_2 \leftarrow \theta_2, \varphi' \leftarrow \varphi$.
3: **for** $k = 1, \cdots, K$ **do**
4: Sample a minibatch B with N samples from $\mathcal{D}$.
5: Train $\omega \leftarrow \arg\min_\omega ELBO(B; G^a_\omega)$. Sample u actions a'_i from $G^a_\omega(s')$ for each s'.
6: Perturb u actions $a'_i = a'_i + \xi_\varphi(s', a'_i)$.
7: Compute next-state value target for each s' in B:

$$V_t = \max_{a'_i}(0.75 \cdot \min\{Q_{\theta'_1}, Q_{\theta'_2}\} + 0.25 \cdot \max\{Q_{\theta'_1}, Q_{\theta'_2}\}).$$

8: $\theta_3 \leftarrow \arg\min_\theta \sum[\max\{g_\theta(s,a) - V_t(s'), 0\} - (1-\rho)g_\theta(s,a)]$.
9: Compute next-state Q target for each (s, a, r, s') pair in B:

$$Q_t(s,a) = r - \gamma \cdot \max\{g_{\theta_3}(s,a) - V_t(s'), 0\} + \gamma(1-\rho)g_{\theta_3}(s,a).$$

10: $\theta \leftarrow \arg\min_\theta \sum(Q_t(s,a) - Q_\theta(s,a))^2$.
11: Sample u actions a_i from $G^a_\omega(s)$ for each s.
12: $\varphi \leftarrow \arg\max_\varphi \sum \max_{a_i} Q_{\theta_1}(s, a_i + \xi_\varphi(s, a_i))$.
13: Update target network: $\theta' = (1-\tau)\theta' + \tau\theta, \varphi' = (1-\tau)\varphi' + \tau\varphi$.
14: **end for**
15: **Output policy:** Given s, sample u actions a_i from $G^a_\omega(s)$. Select action $a = \arg\max_{a_i} Q_{\theta_1}(s, a_i + \xi_\varphi(s, a_i))$.

VAE is given by $ELBO(B; G^a_\omega) = \sum_B (a - \tilde{a})^2 + D_{\text{KL}}(\mathcal{N}(\mu, \sigma), \mathcal{N}(0, 1))$, where $\mathcal{N}$ is the normal distribution with mean and standard deviation parameters. We refer to (Fujimoto et al., 2019) for more details on VAE. We also use the default VAE architecture from BCQ algorithm (Fujimoto et al., 2019) and PQL algorithm (Liu et al., 2020) in our RFQI algorithm.

We now focus on the additions described in blue (steps 8 and 9) in Algorithm 2. For all the other networks we use default architecture from BCQ algorithm (Fujimoto et al., 2019) and PQL algorithm (Liu et al., 2020) in our RFQI algorithm.
(1) In each iteration k, we solve the dual variable function g_θ optimization problem (step 4 in Algorithm 1, step 8 in Algorithm 2) implemented by ADAM (Kingma and Ba, 2014) on the minibatch B with the learning rate l_1 mentioned in Table 1.
(2) Our state-action value target function corresponds to the robust state-action value target function described in (10). This is reflected in step 9 of Algorithm 2. The state-action value function Q_θ optimization problem (step 5 in Algorithm 1, step 9 in Algorithm 2) is implemented by ADAM (Kingma and Ba, 2014) on the minibatch B with the learning rate l_2 mentioned in Table 1.

Environment	Discount γ	Learning rates $[l_1, l_2]$	Q Neural nets $\theta_1 = \theta_2 = [h_1, h_2]$	Dual Neural nets $\theta_3 = [h_1, h_2]$
CartPole	0.99	$[10^{-3}, 10^{-3}]$	$[400, 300]$	$[64, 64]$
Hopper	0.99	$[10^{-3}, 8 \times 10^{-4}]$ $[3 \times 10^{-4}, 6 \times 10^{-4}]$	$[400, 300]$	$[64, 64]$
Half-Cheetah	0.99	$[10^{-3}, 8 \times 10^{-4}]$ $[3 \times 10^{-4}, 6 \times 10^{-4}]$	$[400, 300]$	$[64, 64]$

Table 1: Details of hyper-parameters in FQI and RFQI algorithms experiments.

Hyper-parameters details: We now give the description of hyper-parameters used in our codebase in Table 1. We use same hyper-parameters across different algorithms. Across all learning algorithms we use $\tau = 0.005$ for the target network update, $K = 5 \times 10^5$ for the maximum iterations,

$|\mathcal{D}| = 10^6$ for the offline dataset, $|B| = 1000$ for the minibatch size. We used grid-search for ρ in $\{0.2, 0.3, \cdots, 0.6\}$. We also picked best of the two sets of learning rates mentioned in Table 1. For all the other hyper-parameters we use default values from BCQ algorithm (Fujimoto et al., 2019) and PQL algorithm (Liu et al., 2020) in our RFQI algorithm that can be found in our code.

Offline datasets: Now we discuss the offline dataset used in the our training of FQI and RFQI algorithms. For the fair comparison in every plot, we train both FQI and RFQI algorithms on same offline datasets.

Cartpole dataset $\mathcal{D}_{\text{c}}$: We first train proximal policy optimization (PPO) (Schulman et al., 2017) algorithm, under default RL baseline zoo (Raffin, 2020) parameters. We then generate the Cartpole dataset $\mathcal{D}_{\text{c}}$ with 10^5 samples using an ε-greedy ($\varepsilon = 0.3$) version of this PPO trained policy. We note that this offline dataset contains non-expert behavior meeting the richness of the data-generating distribution assumption in practice.

Mixed dataset $\mathcal{D}_{\text{m}}$: For the MuJoCo environments, *Hopper* and *Half-Cheetah*, we increase the richness of the dataset since these are high dimensional problems. We first train soft actor-critic (SAC) (Haarnoja et al., 2018) algorithm, under default RL baseline zoo (Raffin, 2020) parameters, with replay buffer updated by a fixed ε-greedy ($\varepsilon = 0.1$) policy with the model parameter *actuator_ctrlrange* set to $[-0.85, 0.85]$. We then generate the mixed dataset $\mathcal{D}_{\text{m}}$ with 10^6 samples from this ε-greedy ($\varepsilon = 0.3$) SAC trained policy. We note that such a dataset generation gives more diverse set of observations than the process of $\mathcal{D}_{\text{c}}$ generation for fair comparison between FQI and RFQI algorithms.

D4RL dataset $\mathcal{D}_{\text{d}}$: We consider the *hopper-medium* and *halfcheetah-medium* offline datasets in (Fu et al., 2020) which are benchmark datasets in offline RL literature (Fu et al., 2020; Levine et al., 2020; Liu et al., 2020). These 'medium' datasets are generated by first training a policy online using Soft Actor-Critic (Haarnoja et al., 2018), early-stopping the training, and collecting 10^6 samples from this partially-trained policy. We refer to (Fu et al., 2020) for more details.

We end this section by mentioning the software and hardware configurations used. The training and evaluation is done using three computers with the following configuration. Operating system is Ubuntu 18.04 and Lambda Stack; main softwares are PyTorch, Caffe, CUDA, cuDNN, Numpy, Matplotlib; processor is AMD Threadripper 3960X (24 Cores, 3.80 GHz); GPUs are 2x RTX 2080 Ti; memory is 128GB RAM; Operating System Drive is 1 TB SSD (NVMe); and Data Drive is 4TB HDD.

E.2 More Experimental Results

Here we provide more experimental results and details in addition to Fig. 1-3 in Section 5.

For the *Cartpole*, we compare RFQI algorithm against the non-robust RL algorithms FQI and DQN, and the soft-robust RL algorithm proposed in Derman et al. (2018). We trained FQI and RFQI algorithms on the dataset $\mathcal{D}_{\text{c}}$ (a detailed description of data set is provided in Appendix E.1). We test the robustness of the algorithms by changing the parameters *force_mag* (to model external force disturbance), *length* (to model change in pole length), and also by introducing action perturbations (to model actuator noise). The nominal value of *force_mag* and *length* parameters are 10 and 0.5 respectively. Fig. 4 shows superior robust performance of RFQI compared to the non-robust FQI and DQN. For example, consider the action perturbation performance plot in Fig. 4 where RFQI algorithm improves by 75% compared to FQI algorithm in average cumulative reward for a 40% chance of action perturbation. We note that we found $\rho = 0.5$ is the best from grid-search for RFQI algorithm. The RFQI performance is similar to that of soft-robust DQN. We note that soft-robust DQN algorithm is an online deep RL algorithm (and not an offline RL algorithm) and has no provable performance guarantee. Moreover, soft-robust DQN algorithm requires generating online data according a number of models in the uncertainty set, whereas RFQI only requires offline data according to a single nominal training model.

Before we proceed to describe our results on the OpenAI Gym MuJoCo (Brockman et al., 2016) environments *Hopper* and *Half-Cheetah*, we first mention their model parameters and its corresponding

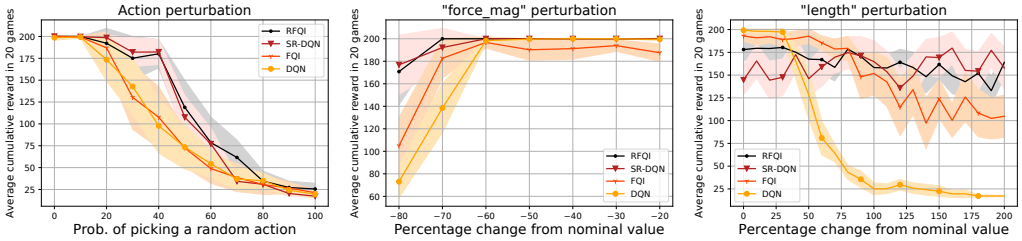

Figure 4: *Cartpole simulation results on offline dataset* $\mathcal{D}_c$. Average cumulative reward in 20 episodes versus different model parameter perturbations mentioned in the respective titles.

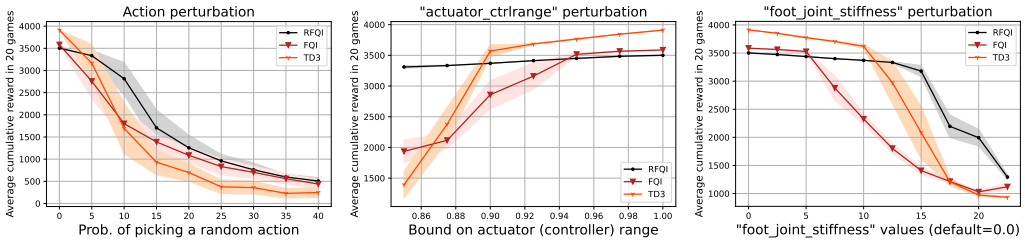

Figure 5: *Hopper simulation results on offline dataset* $\mathcal{D}_m$. Average cumulative reward in 20 episodes versus different model parameter perturbations mentioned in the respective titles.

nominal values in Table 2. The model parameter names are self-explanatory, for example, stiffness control on the leg joint is the *leg_joint_stiffness*, range of actuator values is the *actuator_ctrlrange*. The front and back parameters in Half-Cheetah are for the front and back legs. We refer to the perturbed environments provided in our code and the *hopper.xml, halfcheetah.xml* files in the environment assets of OpenAI Gym MuJoCo (Brockman et al., 2016) for more information regarding these model parameters.

Environment	Model parameter	Nominal range/value
Hopper	*actuator_ctrlrange*	$[-1, 1]$
	foot_joint_stiffness	0
	leg_joint_stiffness	0
	thigh_joint_stiffness	0
	joint_damping	1
	joint_frictionloss	0
Half-Cheetah	*joint_frictionloss*	0
	front *actuator_ctrlrange*	$[-1, 1]$
	back *actuator_ctrlrange*	$[-1, 1]$
	front *joint_stiffness* = (*thigh_joint_stiffness*, *shin_joint_stiffness*, *foot_joint_stiffness*)	$(180, 120, 60)$
	back *joint_stiffness* = (*thigh_joint_stiffness*, *shin_joint_stiffness*, *foot_joint_stiffness*)	$(240, 180, 120)$
	front *joint_damping* = (*thigh_joint_damping*, *shin_joint_damping*, *foot_joint_damping*)	$(4.5, 3.0, 1.5)$
	back *joint_damping* = (*thigh_joint_damping*, *shin_joint_damping*, *foot_joint_damping*)	$(6.0, 4.5, 3.0)$

Table 2: Details of model parameters for *Hopper* and *Half-Cheetah* environments.

For the *Hopper*, we compare RFQI algorithm against the non-robust RL algorithms FQI and TD3 (Fujimoto et al., 2018). We trained FQI and RFQI algorithms on the mixed dataset $\mathcal{D}_{\text{m}}$ (a detailed description of dataset provided in Appendix E.1). We note that we do not compare with soft robust RL algorithms because of its poor performance on MuJoCo environments in the rest of our figures. We test the robustness of the algorithm by introducing action perturbations, and by changing the model parameters *actuator_ctrlrange*, *foot_joint_stiffness*, and *leg_joint_stiffness*. Fig. 3 and Fig. 5 shows RFQI algorithm is consistently robust compared to the non-robust algorithms. We note that

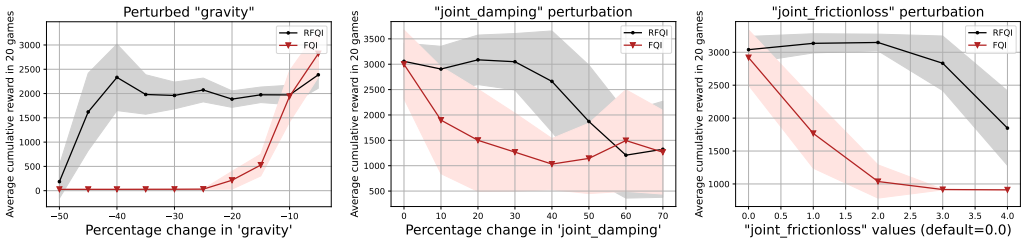

Figure 6: *Hopper evaluation simulation results on offline dataset $\mathcal{D}_d$.* Average cumulative reward in 20 episodes versus different model parameter perturbations mentioned in the respective titles.

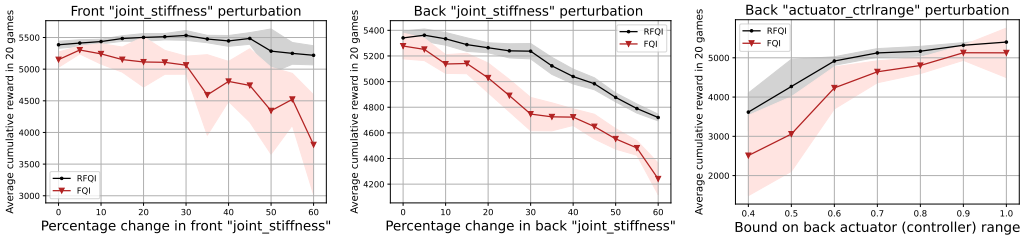

Figure 7: *Half-Cheetah evaluation simulation results on offline dataset $\mathcal{D}_d$.* Average cumulative reward in 20 episodes versus different model parameter perturbations mentioned in the respective titles.

we found $\rho = 0.5$ is the best from grid-search for RFQI algorithm. The average episodic reward of RFQI remains almost the same initially, and later decays much less and gracefully when compared to FQI and TD3 algorithms. For example, in plot 3 in Fig. 5, at the *foot_joint_stiffness* parameter value 15, the episodic reward of FQI is only around 1400 whereas RFQI achieves an episodic reward of 3200. Similar robust performance of RFQI can be seen in other plots as well. We also note that TD3 (Fujimoto et al., 2019) is a powerful off-policy policy gradient algorithm that relies on large 10^6 replay buffer of online data collection, unsurprisingly performs well initially with less perturbation near the nominal models.

In order to verify the effectiveness and consistency of our algorithm across different offline dataset, we repeat the above experiments, on additional OpenAI Gym MuJoCo (Brockman et al., 2016) environment *Half-Cheetah*, using D4RL dataset $\mathcal{D}_{\text{d}}$ (a detailed description of dataset provided in Appendix E.1) which are benchmark in offline RL literature (Fu et al., 2020; Levine et al., 2020; Liu et al., 2020) than our mixed dataset $\mathcal{D}_{\text{m}}$. Since D4RL dataset is a benchmark dataset for offline RL algorithms, here we focus only on the comparison between the two offline RL algorithms we consider, our RFQI algorithm and its non-robust counterpart FQI algorithm. We now showcase the results on *Hopper* and *Half-Cheetah* for this setting.

For the *Hopper*, we test the robustness by changing the model parameters *gravity*, *joint_damping*, and *joint_frictionloss*. Fig. 6 shows RFQI algorithm is consistently robust compared to the non-robust FQI algorithm. We note that we found $\rho = 0.5$ is the best from grid-search for RFQI algorithm. The average episodic reward of RFQI remains almost the same initially, and later decays much less and gracefully when compared to FQI algorithm. For example, in plot 2 in Fig. 6, for the 30% change in *joint_damping* parameter, the episodic reward of FQI is only around 1400 whereas RFQI achieves an episodic reward of 3000 which is almost the same as for unperturbed model. Similar robust performance of RFQI can be seen in other plots as well.

For the *Half-Cheetah*, we test the robustness by changing the model parameters *joint_stiffness* of front and back joints, and *actuator_ctrlrange* of back joint. Fig. 7 shows RFQI algorithm is consistently robust compared to the non-robust FQI algorithm. We note that we found $\rho = 0.3$ is the best from grid-search for RFQI algorithm. For example, in plot 1 in Fig. 7, RFQI episodic reward stays at around 5500 whereas FQI drops faster to 4300 for more than 50% change in the nominal value. Similar robust performance of RFQI can be seen in other plots as well.

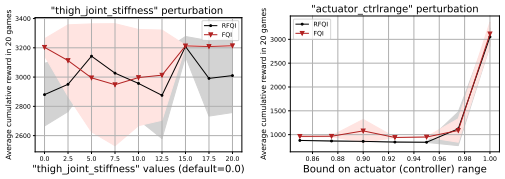

Figure 8: Similar performance of RFQI and FQI in Hopper on dataset $\mathcal{D}_{\text{d}}$ w.r.t. parameters *actuator_ctrlrange* and *thigh_joint_stiffness*.

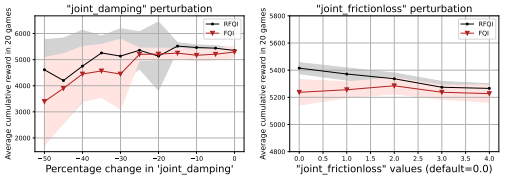

Figure 9: Similar performance of RFQI and FQI in Half-Cheetah on dataset $\mathcal{D}_{\text{d}}$ w.r.t. parameters *joint_damping* and *joint_frictionloss*.

As part of discussing the limitations of our work, we also provide two instances where RFQI and FQI algorithm behave similarly. RFQI and FQI algorithms trained on the D4RL dataset $\mathcal{D}_{\text{d}}$ perform similarly under the perturbations of the *Hopper* model parameters *actuator_ctrlrange* and *thigh_joint_stiffness* as shown in Fig. 8. We also make similar observations under the perturbations of the *Half-Cheetah* model parameters *joint_damping* (both front *joint_damping* and back *joint_damping*) and *joint_frictionloss* as shown in Fig. 9. We observed that the robustness performance can depend on the offline data available, which was also observed for non-robust offline RL algorithms (Liu et al., 2020; Fu et al., 2020; Levine et al., 2020). Also, perturbing some parameters may make the problem really hard especially if the data is not representative with respect to that parameter. We believe that this is the reason for the similar performance of RFQI and FQI w.r.t. some parameters. We believe that this opens up an exciting area of research on developing online policy gradient algorithms for robust RL, which may be able to overcome the restriction and challenges due to offline data. We plan to pursue this goal in our future work.