SOFT ROBUST MDPS AND RISK-SENSITIVE MDPS: EQUIVALENCE, POLICY GRADIENT, AND SAMPLE COMPLEXITY

Runyu (Cathy) Zhang
Harvard University
runyuzhang@fas.harvard.edu

Yang Hu
Harvard University
yanghu@g.harvard.edu

Na Li
Harvard University
nali@seas.harvard.edu

ABSTRACT

Robust Markov Decision Processes (MDPs) and risk-sensitive MDPs are both powerful tools for making decisions in the presence of uncertainties. Previous efforts have aimed to establish their connections, revealing equivalences in specific formulations. This paper introduces a new formulation for risk-sensitive MDPs, which assesses risk in a slightly different manner compared to the classical Markov risk measure [71], and establishes its equivalence with a class of soft robust MDP (RMDP) problems, including the standard RMDP as a special case. Leveraging this equivalence, we further derive the policy gradient theorem for both problems, proving gradient domination and global convergence of the exact policy gradient method under the tabular setting with direct parameterization. This forms a sharp contrast to the Markov risk measure, known to be potentially non-gradient-dominant [39]. We also propose a sample-based offline learning algorithm, namely the robust fitted-Z iteration (RFZI), for a specific soft RMDP problem with a KL-divergence regularization term (or equivalently the risk-sensitive MDP with an entropy risk measure). We showcase its streamlined design and less stringent assumptions due to the equivalence and analyze its sample complexity.

1 INTRODUCTION

Making decisions amidst uncertainty presents a fundamental challenge cutting across diverse domains, including finance [32, 80], engineering [45, 74], and robotics [88] etc. Within these realms, decisions carry consequences that depend not only on expected rewards but also on the level of uncertainty and associated risks. Addressing this challenge necessitates approaches such as robust, and risk-sensitive decision-making. These approaches explicitly incorporate uncertainty and aim to find policies that perform well across a spectrum of scenarios and adeptly strike a balance between expected gains and potential risks.

For robust decision-making in a dynamic environment, the robust Markov Decision Process (RMDP) is a popular framework. RMDPs model the environment as a Markov decision process, seeking policies that excel across various potential models. This involves solving a max-min problem, optimizing an objective function that considers the policy's worst-case performance across all models within a defined uncertainty set. The RMDP framework was introduced by [41, 57], spurring research into efficient planning algorithms when the model is given [34, 96, 93, 100, 56].There are also works focusing on the computational facets for these problems [37, 6, 35, 21] which leverage convex formulation and regularization techniques to tackle robustness. In cases of unknown models [1], recent efforts have designed reinforcement learning (RL) algorithms with guarantees, but most are model-based for tabular cases, i.e., requiring an empirical estimation of the probability transition model [51, 63, 105, 99, 98, 77], thereby impeding their applicability to large state spaces. Some works focus on the model-free setting and employ linear function approximation for handling large state spaces [85, 70, 4]. However, these approaches provide only asymptotic guarantees and rely on approximated robust dynamic programming, which inherently is computationally more expensive than standard dynamic programming. A recent contribution by [64] offers non-asymptotic sample

[1]By 'unknown model' we refer to the setting where the nominal probability transition model is unknown. Both model-based and model-free methods belong to this setting, where model-based methods keep an empirical estimate of the nominal model whereas model-free algorithms don't require this empirical estimation step.

complexity guarantees in the context of model-free robust RL. This achievement, however, introduces additional dual variables, thus adding additional computational complexity and imposing more stringent assumptions.

An alternative approach for handling uncertainty is risk-sensitive decision-making, which intriguingly shares an elegant equivalence with robust decision-making. The concept of coherent risk measures was initially introduced and explored in [2, 18, 69], where the uncertainty is represented by a static random variable. The connection to robustness was established by characterizing risk measures as the infimum of expected shortfall across a set of probability measures, known as the risk envelope. The risk notion is further extended to convex risk measures which capture a broader class of risk evaluation functions [30, 73, 31]. Subsequently, conditional and dynamic risk measures were introduced to generalize risk assessment from static random variables to stochastic processes [3, 14, 29, 22, 68, 72, 65]. In particular, [71] introduces the *Markov risk measure* in the context of Markov Decision Processes (MDPs). However, the equivalence between the Markov risk measure and robust MDPs is not as straightforward as in static settings. Notably, [71, 75, 11, 5, 62] established the equivalence between optimizing the Markov risk measure and solving a modified RMDP problem, where the uncertainty set dynamically changes with the implemented policy. This differs from the standard RMDPs, where the uncertainty sets are typically unrelated to the policy. Though [62] attains stronger equivalence results with RMDPs, it is only applicable to specific risk measures, such as Conditional Value at Risk (CVaR). Similar to RMDPs, optimizing Markov risk measures also faces many challenges. Firstly, building upon the equivalence with the modified RMDP with policy-dependent uncertainty set, Huang et al. [39] highlights that, even in a tabular setting with direct parameterization, Markov risk measures may lack gradient-dominance – a stark contrast to the gradient domination observed in standard MDPs [1]. This implies that policy gradient algorithms may not ensure global optima, even in a straightforward, full-information environment. Further, the sample complexity is also harder to obtain. While there is a series of efforts dedicated to optimizing the Markov risk measure within the realm of RL [12, 76, 46], these works primarily provide asymptotic convergence results.

The challenges outlined above motivate us to investigate the potential of introducing an alternative risk formulation. This new formulation seeks to capture risk in a way similar to Markov risk measures while achieving a stronger and broader equivalence with RMDPs. Moreover, we aim to enhance convergence properties, including the crucial aspect of gradient domination. These improvements are poised to support the development of learning algorithms for both RMDPs and risk-sensitive MDPs while maintaining provable guarantees.

Our Contributions: In this paper, we propose a new formulation for risk-sensitive MDP, whose definition incorporates the general concepts of convex risk measures. We first establish the equivalence of risk-sensitive MDP with a class of soft RMDP problems, which includes the standard RMDP as a special case. Leveraging this equivalence, we proceed to derive the policy gradient theorem for both the aforementioned class of soft RMDPs and risk-sensitive MDPs (Theorem 3) and prove the global convergence of the exact policy gradient method under the tabular setting with direct parameterization. Our result, to the best of our knowledge, presents the first global convergence analysis with iteration complexity for a general class of risk-sensitive MDPs.

Based on the policy gradient theorem, we also highlight the difficulty of gradient estimation using samples compared with the standard MDP setting, motivating us to seek other types of sample-based learning methods. In the last part of this paper, we mainly focus on the setting of offline learning with nonlinear function approximation which is a relatively less-studied scenario, and propose a sample-based offline learning algorithm, namely the robust fitted-Z iteration (RFZI), that resembles policy iteration rather than policy gradient.Specifically, we focus on a setting where the regularization term for the RMDP is a KL-divergence term, which is equivalent to the risk-sensitive MDP with the entropy risk measure. The algorithm utilizes the equivalence between the two problems, which enables simpler algorithm design. Notably, our algorithm is model-free and does not rely on an empirically estimated probability transition model. The sample complexity for RFZI is also provided. Compared with [64] which considers offline robust RL with sample-complexity guarantees, our work considers a different uncertainty set, requires less computational and implementation complexity, and less stringent assumptions.

Due to space limit, we defer a detailed literature review and numerical simulations to the appendix.

2 Problem Settings and Preliminaries

Markov Decision Processes (MDPs). A finite Markov decision process (MDP) is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho)$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions available to the agent, and P is the transition probability function such that $P(s'|s,a)$ describes the probability of transitioning from one state s to another s' given a particular action a. For the sake of notation simplicity, we use $P_{s,a}$ to denote the probability distribution $P(\cdot|s,a)$ over the state space $\mathcal{S}$. $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is a reward function, $\gamma \in [0,1)$ is a discounting factor, and ρ specifies the initial probability distribution over the state space $\mathcal{S}$.

A stochastic policy $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$ specifies a strategy where the agent chooses its action based on the current state in a stochastic fashion; more specifically, the probability of choosing action a at state s is given by $\Pr(a|s) = \pi(a|s)$. A deterministic policy is a special case of the stochastic policy where for every state s there is an action a_s such that $\pi(a_s|s) = 1$. For notation simplicity, we slightly overload the notation and use $\pi(s)$ to denote the action a_s for deterministic policies. For a given stationary policy π and a set of transition probability distributions $\{P_{s,a}\}_{s\in\mathcal{S},a\in\mathcal{A}}$, we denote the discounted state visitation distribution by

$$d^{\pi,P}(s) := (1-\gamma)\sum_{t=0}^{+\infty}\gamma^t \Pr^{\pi,P}(s_t = s \mid s_0 \sim \rho).$$

Robust MDPs (RMDPs) and Soft Robust MDPs. Unlike the standard MDP which considers a fixed transition model $\{P_{s,a}\}$, the robust MDP considers a set $\mathcal{P}$ of transition probability distributions and aims to solve the sup-inf problem [41]

$$\sup_\pi \inf_{\{\widehat{P}_t\in\mathcal{P}\}_{t\ge 0}} \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P},s_0\sim\rho}\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t)\right)^{2} \tag{1}$$

where the objective is to find the best action sequence that maximizes a worst-case objective over all possible models in the uncertainty set $\mathcal{P}$. Many papers [41, 57, 99, 63, 4] consider the uncertainty set under the (s,a)-rectangularity condition $\mathcal{P} = \otimes_{s\in\mathcal{S},a\in\mathcal{A}}\mathcal{P}_{s,a}$, where $\mathcal{P}_{s,a} = \{\widehat{P}_{s,a} : \ell(\widehat{P}_{s,a}, P_{s,a}) \le \epsilon\}$, and ℓ is a penalty function that captures the deviation of $\widehat{P}_{s,a}$ from a nominal model $P_{s,a}$. Some popular penalty functions are KL divergence, total variation distance, etc.

In this paper, we generalize the above robust MDP problem to a wider range of problems which we call the *soft robust MDP*[3]. The objective of the soft robust MDP solves the following sup-inf problem:

$$\sup_\pi \inf_{\{\widehat{P}_t\}_{t\ge 0}} \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P},s_0\sim\rho}\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right). \tag{2}$$

Note that here $\inf_{\{\widehat{P}_t\}_{t\ge 0}}$ is with respect to all the possible state-transition probability distributions. When the penalty function D is chosen as the indicator function

$$D(\widehat{P}_{s,a}, P_{s,a}) = \begin{cases} 0 & \ell(\widehat{P}_{s,a}, P_{s,a}) \le \epsilon \\ +\infty & \text{otherwise} \end{cases},$$

it recovers the robust MDP problem (1). When D is set as non-indicator functions, for example, $D(\widehat{P}_{s,a}, P_{s,a}) = \mathrm{KL}(\widehat{P}_{s,a}||P_{s,a})$, Problem (2) is a robust MDP with a soft penalty term D on the deviation of $\widehat{P}_{s,a}$ from $P_{s,a}$ rather than a hard constraint on $\widehat{P}_{s,a}$.

Similar to the robust MDP problem, we can define the optimal value function as

$$\overline{V}^\star(s) := \sup_\pi \inf_{\{\widehat{P}_t\}_{t\ge 0}} \mathbb{E}_{s_t\sim\widehat{P}}\left[\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)\Big| s_0 = s\right]. \tag{3}$$

Additionally, given a stationary policy π, the value function $\overline{V}^\pi$ under policy π is defined as follows:

$$\overline{V}^\pi(s) := \inf_{\{\widehat{P}_t\}_{t\ge 0}} \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)\Big| s_0 = s\right]. \tag{4}$$

We also define the corresponding Q-functions as

$$\overline{Q}^\star(s,a) := \sup_{\{a_t\}_{t\ge 1}} \inf_{\{\widehat{P}_t\}_{t\ge 0}} \mathbb{E}_{s_t\sim\widehat{P}}\left[\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)\Big| s_0 = s, a_0 = a\right]$$

$$\overline{Q}^\pi(s,a) := \inf_{\{\widehat{P}_t\}_{t\ge 0}} \mathbb{E}_{s_t,a_t\sim\pi,t\ge 1,\widehat{P}}\left[\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)\Big| s_0 = s, a_0 = a\right].$$

[2]For the sake of generality, we allow the transition probability to be non-stationary and the policy to be non-Markovian and stochastic. However, in later sections we will show that the sup-inf solution can be obtained by a stationary deterministic Markov policy and a stationary transition probability (Theorem 2).

[3]We adopt the term from robust optimization literature, the concept of regularizing the adversaries actions is referred as soft-robustness [9] (or comprehensive robustness [8] and globalized robustness [10]).

Remark 1 (Soft Robust MDP.). *The soft robust MDP problem is useful, especially when the uncertainty set is not explicitly given. In this case, it is more desirable to consider all possible probability transition models $\{\widehat{P}_t\}_{t\geq 0}$ while treating the deviation from the nominal model as a soft penalty term D rather than constraining it to be within a specified uncertainty set.*

In this paper, we establish a connection between the soft robust MDP and another class of MDPs, namely risk-sensitive MDPs. To define the risk-sensitive MDP, we will first introduce the notation of convex risk measures.

Convex Risk Measures [30]. Consider a finite set $\mathcal{S}$, let $\mathbb{R}^{|\mathcal{S}|}$ denote the set of real-valued functions over $\mathcal{S}$. A convex risk measure $\sigma : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ is a function that satisfies the following properties:

1. Monotonicity: for any $V', V \in \mathbb{R}^{|\mathcal{S}|}$, if $V' \leq V$, then $\sigma(V) \leq \sigma(V')$.
2. Translation invariance: for any $V \in \mathbb{R}^{|\mathcal{S}|}, m \in \mathbb{R}$, $\sigma(V + m) = \sigma(V) - m$.
3. Convexity: for any $V', V \in \mathbb{R}^{|\mathcal{S}|}, \lambda \in [0, 1]$, $\sigma(\lambda V + (1-\lambda)V') \leq \lambda\sigma(V) + (1-\lambda)\sigma(V')$.

Using standard duality theory, it is shown in classical results [30] that convex risk measures satisfy the following dual representation theorem:

Theorem 1 (Dual Representation Theorem [30]). *The function $\sigma : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ is a convex risk measure if and only if there exists a "penalty function" $D(\cdot) : \Delta^{|\mathcal{S}|} \to \mathbb{R}$ such that*

$$\sigma(V) = \sup_{\widehat{\mu}\in\Delta^{|\mathcal{S}|}} \left(-\mathbb{E}_{\widehat{\mu}} V - D(\widehat{\mu})\right). \tag{5}$$

Further, the penalty function D can be chosen to satisfy the condition $D(\widehat{\mu}) \geq -\sigma(0)$ for any $\widehat{\mu} \in \Delta^{|\mathcal{S}|}$ and it can be taken to be convex and lower-semicontinuous. In specific, it can be written in the following form:

$$D(\widehat{\mu}) = \sup_V \left(-\sigma(V) - \mathbb{E}_{s\sim\widehat{\mu}} V(s)\right) \tag{6}$$

Note that σ and D serve as the Fenchel conjugate of each other. In most cases, the convex risk measure $\sigma(V)$ can be interpreted as the risk associated with a random variable that takes on values $V(s)$ where s is drawn from some distribution $s \sim \mu$. Consequently, most commonly used risk measures are typically associated with an underlying probability distribution $\mu \in \Delta^{|\mathcal{S}|}$ (e.g., Examples 1). This paper focuses on this type of risk measures and thus we use $\sigma(\mu, \cdot)$ to denote the risk measure, where the additional variable μ indicates the associated probability distribution. Correspondingly, we denote the penalty term $D(\hat{\mu})$ of $\sigma(\mu, \cdot)$ in the dual representation theorem as $D(\hat{\mu}, \mu)$.[4]

Here we provide an example of convex risk measure and its dual form.

Example 1 (Entropy risk measure [30]). *For a given $\beta > 0$, the entropy risk measure takes the form:*

$$\sigma(\mu, V) = \beta^{-1} \log \mathbb{E}_{s\sim\mu} e^{-\beta V(s)}.$$

Its corresponding penalty function D in the dual representation theorem is the KL divergence

$$D(\widehat{\mu}, \mu) = \beta^{-1}\mathrm{KL}(\widehat{\mu}||\mu) = \beta^{-1} \sum_{s\in\mathcal{S}} \widehat{\mu}(s) \log\left(\widehat{\mu}(s)/\mu(s)\right).$$

Risk-Sensitive MDPs. Convex risk measures capture the risk associated with random variables. It would be desirable if the notion could be adapted to the MDP to capture the risk of a given policy under the Markov process. Given an MDP $\mathcal{M}$, a class of convex risk measures $\{\sigma(P_{s,a}, \cdot)\}_{s\in\mathcal{S}, a\in\mathcal{A}}$, and a policy $\pi(\cdot|s)$, the risk-sensitive value function $\widetilde{V}^\pi$ for the infinite discounted MDP is given as

$$\widetilde{V}^\pi(s) = \sum_a \pi(a|s) \left(r(s, a) - \gamma\sigma(P_{s,a}, \widetilde{V}^\pi)\right), \forall s \in \mathcal{S}. \tag{7}$$

With the definition of risk-sensitive $\widetilde{V}^\pi$, the risk-sensitive MDP problem is to find the policy that maximizes $\max_\pi \widetilde{V}^\pi$. We denote the optimal value by $\widetilde{V}^\star$, which is the fix-point solution of the following equation,

[4]Please note that the symbol D serves a dual purpose, representing both the regularization term in (2) and the penalty function for a risk measure in (5) and (6). This intentional notation overlap will become clear in the following sections, which reveal the connection between these two terms.

$$\widetilde{V}^\star(s) := \max_a \left(r(s,a) - \gamma\sigma(P_{s,a}, \widetilde{V}^\star)\right), \forall s \in \mathcal{S}. \tag{8}$$

It is worth noting that the fixed point operators for (7),(8) are contractive (proof deferred to Appendix D), which immediately implies the following lemma which verifies that the fixed point equations for $\widetilde{V}^\pi$ (7) and $\widetilde{V}^\star$ (8) are well-defined .

Lemma 1. *The solution to* (7) *exists and is unique. Same argument holds for* (8)*.*

Remark 2. *We would like to emphasize that when the policy π is stochastic, our definition of the value functions $\widetilde{V}^\pi$ are different from the Markov risk measures defined in [71, 39, 86, 87]* [5]*. However, the two quantities are equivalent when π is deterministic. Additionally, when further assuming that the risk measure σ is mixture quasiconcave (c.f. [17]), the optimal policy for the Markov risk measure is also deterministic and thus the risk-sensitive MDP and the Markov risk measure obtain the same optimal value $\widetilde{V}^\star$* [6] *(see Appendix C for more details).*

We also define the Q-function of the risk sensitive MDP as:

$$\widetilde{Q}^\star(s,a) := r(s,a) - \gamma\sigma(P_{s,a}, \widetilde{V}^\star), \quad \widetilde{Q}^\pi(s,a) := r(s,a) - \gamma\sigma(P_{s,a}, \widetilde{V}^\pi).$$

Other notations: For any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, state-action distribution $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ the μ-weighted 2-norm of f is defined as $\|f\|_{2,\mu} = \left(\mathbb{E}_{s,a\sim\mu} f(s,a)^2\right)^{1/2}$.

3 Equivalence of Soft RMDPs and Risk-Sensitive MDPs

Theorem 2 (Equivalence of Soft RMDPs and Risk-Sensitive MDPs)**.** *For a given MDP $\mathcal{M}$, a penalty function D, a class of convex risk measures $\{\sigma(P_{s,a}, \cdot)\}$, and a stationary policy π, if the penalty function D satisfies*

$$D(\widehat{P}_{s,a}, P_{s,a}) = \sup_V \left(-\sigma(P_{s,a}, V) - \mathbb{E}_{s'\sim\widehat{P}_{s,a}} V(s')\right), \tag{9}$$

then the value functions and Q-functions of the soft RMDP and the risk-sensitive MDP are always the same. That is, $\overline{V}^\star = \widetilde{V}^\star =: V^\star$, $\overline{V}^\pi = \widetilde{V}^\pi =: V^\pi$, $\overline{Q}^\star = \widetilde{Q}^\star =: Q^\star$, $\overline{Q}^\pi = \widetilde{Q}^\pi =: Q^\pi$. [7]

Further, for every initial state s_0, the sup-inf solution of the policy and transition probabilities for $V^\star(s_0)$ defined in (3) *is given by:*

$$\begin{aligned} &\pi^\star(s) = \operatorname{argmax}_a \left(r(s,a) - \gamma\sigma(P_{s,a}, V^\star)\right), \\ &\widehat{P}^\star_{t;s,a} = \widehat{P}^\star_{s,a} = \operatorname{argmin}_{\widehat{P}} D(\widehat{P}, P_{s,a}) + \mathbb{E}_{s'\sim\widehat{P}} V^\star(s'). \end{aligned} \tag{10}$$

where (10) *means that the optimal action sequence $\{a_t\}_{t\ge1}$ can be achieved by implementing the deterministic policy $a_t = \pi^\star(s_t)$.*

Similarly, for any initial state s_0, the minimum solution of the transition probabilities for $V^\pi(s_0)$ defined in (4) *is given by*

$$\widehat{P}^\pi_{t;s,a} = \widehat{P}^\pi_{s,a} = \operatorname{argmin}_{\widehat{P}} D(\widehat{P}, P_{s,a}) + \mathbb{E}_{s'\sim\widehat{P}} V^\pi(s'). \tag{11}$$

Since Theorem 2 has established the equivalence of risk-sensitive MDPs and soft RMDPs, *from now on we use $V^\star, V^\pi, Q^\star, Q^\pi$ to denote the value functions and Q-functions for both settings and assume by default that the penalty function D and the risk measure σ satisfy relationship* (9).

Remark 3. *As a comparison to the equivalence result for the Markov risk measures [72, 71, 86], their uncertainty set for the robust problem generally depends on the policy π (see e.g. Assumption 2.2 in [86]), while in our setting, the penalization function D is independent of the policy and matches with the most standard formulation of RMDPs.*

---

[5] Due to this difference, the value function V^π can no longer be written as $\rho(\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t))$ where ρ is a time-consistent dynamic risk measure. This makes our definition different from the usual interpretation of the dynamic risk measures.

[6] We would like to note that the equivalence of optimal value might fail if σ is not mixture semiconcave (e.g. mean (semi)-deviation, mean (semi)-moment measures [17]) or if policy regularization is added into the value function because the optimal policy might no longer be deterministic.

[7] The equivalence $\overline{V}^\pi = \widetilde{V}^\pi, \overline{Q}^\pi = \widetilde{Q}^\pi$ easily extends to the setting with policy regularization, since adding regularization only requires changing the reward function $r(s,a)$ to be $r^\pi(s,a) = r(s,a) + \mathcal{R}(\pi(\cdot|s))$, where $\mathcal{R}$ is the policy regularizer, in which case the proof of Theorem 2 can still carry through naturally.

4 POLICY GRADIENT FOR SOFT RMDPS

In this section, we present the policy gradient theorem for a differentiable policy π_θ parameterized by θ, which provides an analytical method for computing the gradient in soft RMDPs. Additionally, we prove the global convergence of the exact policy gradient ascent algorithm for the direct parameterization case. For simplicity, in this section we use the abbreviations $V^\theta, Q^\theta, \widehat{P}^\theta, V^{(t)}, Q^{(t)}, \widehat{P}^{(t)}$ to denote $V^{\pi_\theta}, Q^{\pi_\theta}, \widehat{P}^{\pi_\theta}, V^{\pi_{\theta^{(t)}}}, Q^{\pi_{\theta^{(t)}}}, \widehat{P}^{\pi_{\theta^{(t)}}}$, respectively.

Theorem 3 (Policy gradient theorem)**.** *Suppose that π_θ is differentiable with respect to θ and that $\sigma(P_{s,a},\cdot):\mathbb{R}^{|\mathcal{S}|}\to\mathbb{R}$ is a differentiable function, then $V^\theta(s)$ is also a differentiable function with respect to θ and the gradient is given by*

$$\nabla_\theta V^\theta(s) = \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t), s_{t+1}\sim\widehat{P}^\theta_{s_t,a_t}}\left[\sum_{t=0}^{+\infty}\gamma^t Q^\theta(s_t,a_t)\nabla_\theta\log\pi_\theta(a_t|s_t)\Big|s_0=s\right],$$

where $\widehat{P}^\theta$ is defined in (11)*.*

We leave the discussion of this result to the end of this section in Remark 4. Theorem 3 immediately implies the following corollary on the policy gradient under direct parameterization (c.f. [1, 94]), where the parameter $\theta_{s,a}$ directly represents the probability of choosing action a at state s, i.e., $\theta_{s,a}=\pi_\theta(a|s)$.

Corollary 1 (Policy gradient for direct parameterization)**.** *Under direct parameterization,*

$$\frac{\partial\mathbb{E}_{s_0\sim\rho}V^\theta(s_0)}{\partial\theta_{s,a}} = \frac{1}{1-\gamma}d^{\pi_\theta,\widehat{P}^\theta}(s)Q^\theta(s,a). \tag{12}$$

Note that the policy gradient theorem only holds for the case where $\sigma(P_{s,a},\cdot)$ is differentiable; nevertheless, we can generalize (12) to the non-differentiable case by defining the variable $G(\theta)\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as follows:

$$[G(\theta)]_{s,a} := \frac{1}{1-\gamma}d^{\pi_\theta,\widehat{P}^\theta}(s)Q^\theta(s,a).$$

For both differentiable and non-differentiable cases, we could perform the following ('quasi'-)gradient ascent algorithm:

$$\theta^{(t+1)} = \mathrm{Proj}_{\mathcal{X}}(\theta^{(t)}+\eta G(\theta^{(t)})), \tag{13}$$

where $\mathcal{X}=\otimes_{s\in\mathcal{S}}\Delta^{|\mathcal{A}|}$ denotes the feasible region of θ. For the standard MDP case, it is known that the value function satisfies the gradient domination property under direct parameterization [1], which enables global convergence of the policy gradient algorithm. Similar properties also hold for the soft RMDP/risk-sensitive MDP setting which is shown in the following lemma:

Lemma 2 (Gradient domination under direct parameterization)**.**

$$\mathbb{E}_{s_0\sim\rho}V^\star(s_0) - V^\theta(s) \le \left\|\frac{d^{\pi^\star,\widehat{P}^\theta}}{d^{\pi_\theta,\widehat{P}^\theta}}\right\|_\infty \max_{\overline{\pi}}\left\langle\overline{\pi}-\pi_\theta, G(\theta)\right\rangle,$$

where $\left\|\frac{d^{\pi^\star,\widehat{P}^\theta}}{d^{\pi_\theta,\widehat{P}^\theta}}\right\|_\infty := \max_s\frac{d^{\pi^\star,\widehat{P}^\theta}(s)}{d^{\pi_\theta,\widehat{P}^\theta}(s)}$.

The gradient domination property suggests that as long as the term $\left\|\frac{d^{\pi^\star,\widehat{P}^\theta}}{d^{\pi_\theta,\widehat{P}^\theta}}\right\|_\infty$ is not infinite, all the first order stationary points are global optimal solutions. Based on this observation, we further derive the convergence rate for the policy gradient algorithm. Before that, we introduce the following sufficient exploration assumption:

Assumption 1 (Sufficient Exploration)**.** *For any policy π, it holds that $d^{\pi,\widehat{P}^\pi}(s)>0$, where $\widehat{P}^\pi$ is defined as in* (11)*. We define the distributional shift factor M to be a constant that satisfies $M\ge\frac{1}{d^{\pi,\widehat{P}^\pi}(s)}$ for all state s and policy π.*

Note that when we start with a initial distribution where $\rho(s)>0$ for every state s, the term M can be upper bounded by $\frac{1}{(1-\gamma)\min_s\rho(s)}$. If Assumption 1 is satisfied, it can be concluded that $\left\|\frac{d^{\pi^\star,\widehat{P}^\theta}}{d^{\pi_\theta,\widehat{P}^\theta}}\right\|_\infty\le M$. Thus we could use gradient domination to derive the global convergence rate.

Theorem 4 (Convergence rate for exact policy gradient under direct parameterization)**.** *Under Assumption 1, by setting $\eta = \frac{(1-\gamma)^3}{2|\mathcal{A}|M}$, running* (13) *guarantees that*

$$\sum_{k=1}^{K} \left(\mathbb{E}_{s_0\sim\rho} V^\star(s_0) - V^{(k)}(s_0)\right)^2 \leq \frac{16|\mathcal{A}|M^4}{(1-\gamma)^4}.$$

Therefore, by setting $K \geq \frac{16|\mathcal{A}|M^4}{(1-\gamma)^4\epsilon^2}$, it is guaranteed that $\min_{1\leq k\leq K} \mathbb{E}_{s_0\sim\rho}(V^\star(s_0) - V^{(k)}(s_0)) \leq \epsilon$.

If we apply the same proof technique to standard MDPs, the convergence rate is $O\left(\frac{|\mathcal{A}|M^2}{(1-\gamma)^4}\right)$. The dependency on the distributional shift factor M is worse for soft RMDPs, which is caused by the choice of a smaller stepsize η (see Remark 7 in the Appendix for more details). It is an interesting open question whether this worse dependency is fundamental or just a proof artifact.

Remark 4 (Difficulties of Sample-based Gradient Estimation)**.** *Though Theorem 4 establishes the global convergence of exact policy gradient, it is hard to generalize the result to sample-based settings. Note that the policy gradient in Theorem 3 takes a similar form as compared to standard MDPs [83], however, there's a primary distinction that the expectation is taken over trajectories sampled from the probability transition model $\widehat{P}^\theta$ instead of the nominal model P. Consequently, when confined to samples exclusively from the nominal model, estimating this expectation becomes exceptionally challenging, particularly in the context of non-generative models.*

5 Offline Reinforcement Learning of the KL-soft RMDP

Since the previous section considers learning with full information and studies iteration complexity, the major motivation for this section is to examine sample-based learning for risk sensitivity MDPs and soft robust MDPs. As discussed in Remark 4, developing sample-based policy gradient learning methods might be difficult, therefore, we seek an alternative sample-based method that resembles policy iteration rather than policy gradient. Specifically, we mainly focus on the setting of offline learning with nonlinear function approximation which is a relatively less-studied scenario. Moreover, due to the challenge in developing a method for soft MDPs with general D functions (or equivalently for risk sensitive MDPs with general risk functions σ), in this section, we look into a particular and important case of soft RMDP where the regularization term is the KL-divergence, i.e.,

$$\max_\pi \min_{\widehat{P}_t} \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}_t,s_0\sim\rho} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t) + \gamma\beta^{-1}\mathrm{KL}(\widehat{P}_{t;s_t,a_t}||P_{s_t,a_t})\right). \tag{14}$$

The hyperparameter β represents the penalty strength of the deviation of $\widehat{P}$ from P, the smaller β is, the larger the penalty strength. From Example 1 and Theorem 2, the KL-soft RMDP is equivalent to the risk-sensitive MDP problem with the risk measures $\sigma(P_{s,a},\cdot)$ chosen as the entropy risk measure

$$\sigma(P_{s,a}, V) = \beta^{-1}\log \mathbb{E}_{s'\sim P_{s,a}} e^{-\beta V(s')}.$$

In this case, the Bellman equations for the value functions $V^\pi, V^\star, Q^\pi, Q^\star$ are given by:

$$V^\pi(s) = \textstyle\sum_a \pi(a|s)Q(s,a), \quad Q^\pi(s,a) = r(s,a) - \gamma\beta^{-1}\log \mathbb{E}_{s'\sim P_{s,a}} e^{-\beta V^\pi(s')},$$
$$V^\star(s) = \max_a Q(s,a), \qquad Q^\star(s,a) = r(s,a) - \gamma\beta^{-1}\log \mathbb{E}_{s'\sim P_{s,a}} e^{-\beta V^*(s')}.$$

For notational simplicity, we define the Bellman operator on the Q-functions $\mathcal{T}_Q : \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as:

$$[\mathcal{T}_Q Q](s,a) := r(s,a) - \gamma\beta^{-1}\log \mathbb{E}_{s'\sim P(\cdot|s,a)} e^{-\beta \max_{a'} Q(s',a')}. \tag{15}$$

It is not hard to verify from the above arguments that the optimal Q function $Q^\star$ satisfies

$$Q^\star = \mathcal{T}_Q Q^\star.$$

Offline robust reinforcement learning. The remainder of the paper focuses on finding the optimal robust policy $\pi^\star$ for the soft robust MDP problem (14). Specifically, we explore offline robust reinforcement learning algorithms which use a pre-collected dataset $\mathcal{D}$ to learn $\pi^\star$. The dataset is typically generated under the nominal model $\{P_{s,a}\}_{s\in\mathcal{S},a\in\mathcal{A}}$, such that $\mathcal{D} = \{s_i, a_i, r_i, s_i'\}_{i=1}^N$, where the state-action pairs $(s_i, a_i) \sim \mu$ are drawn from a specific data-generating distribution μ.

Definition 1 (Robustly Admissible Distributions)**.** *A distribution $\nu \in \Delta^{|\mathcal{S}|\times|\mathcal{A}|}$ is robustly admissible if there exists $h \geq 0$ and a policy π and transition probability $\widehat{P} \in \{P' : \mathrm{KL}(P'_{t;s,a}||P_{s,a}) \leq \beta\}$ (both can be non-stationary) such that $\nu(s,a) = \Pr(s_h, a_h | s_0 \sim \rho, \pi, \widehat{P})$.*

Assumption 2 (Concentrability)**.** *The data-generating distribution μ satisfies concentrability if there exists a constant C such that for any ν that is robustly admissible, $\max_{s,a} \frac{\nu(s,a)}{\mu(s,a)} \leq C$.*

Remark 5. *The notion of robustly admissible distribution and concentrability are adapted from the the corresponding notions defined for the standard MDP setting [13], where they also demonstrate the necessity of this assumption for standard RL with function approximation. It would be an interesting open question whether Assumption 2 is also necessary for robust RL settings. Recent works for standard offline RL also show that by considering variations of the RL algorithms (e.g. exploring pessimism [95] or the primal-dual formulation [102]), the concentrability assumption can be weakened to single-policy concentrability. Another interesting future direction is to study whether applying similar approaches for the soft RMDP would result in the same improvement.*

5.1 Robust Fitted-Z Iteration (RFZI)

The offline robust MDP learning method we propose is Robust fitted-Z iteration (RFZI). The main idea is to utilize the fix point equation $Q^\star = \mathcal{T}_Q Q^\star$ with the Bellman operator (15) from the corresponding equivalent risk-sensitive MDP. However, $\mathcal{T}_Q$ involves a term $\log \mathbb{E}_{s'\sim P_{s,a}}$ which is hard to approximate with empirical estimation. Thus, instead of directly solving Q^* using $Q^\star = \mathcal{T}_Q Q^\star$, we introduce an auxiliary variable, Z-function and solve a fix point equation for Z, which play an important role in our algorithm design and theoretical analysis.

The Z-functions. For a given Q-function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define its corresponding Z-function as below:

$$Z(s,a) := \mathbb{E}_{s'\sim P_{s,a}} e^{-\beta \max_{a'} Q(s',a')}.$$

One can establish the relationship between the Z-function and the Q-function by

$$[\mathcal{T}_Q Q](s,a) = r(s,a) - \gamma\beta^{-1} \log Z(s,a).$$

Further, we also define the Z-Bellman operator on Z-functions as:

$$[\mathcal{T}_Z Z](s,a) := \mathbb{E}_{s'\sim P_{s,a}} e^{-\beta \max_{a'}(r(s',a') - \gamma\beta^{-1}\log Z(s',a'))}.$$

Then $\mathcal{T}_Q[\mathcal{T}_Q Q]](s,a) = r(s,a) - \gamma\beta^{-1}\log[\mathcal{T}_Z Z](s,a)$. Thus, instead of solving $Q^\star = \widetilde{\mathcal{T}}_Q Q^\star$, an alternative approach is to solve $Z^\star = \widetilde{\mathcal{T}}_Z Z^\star$ and recover $Q^\star$ by $Q^\star = r - \gamma\beta^{-1}\log Z^\star$. This is the key intuition of our RFZI algorithm. Note that compare with $\mathcal{T}_Q$, $\mathcal{T}_Z$ eliminates the log dependency on the expectation term $\mathbb{E}_{s'\sim P_{s,a}}$, which makes it easier for empirical estimation.

Function approximation and projected Z-Bellman operator. Given that $\widetilde{\mathcal{T}}_Z$ is a contraction mapping, the solution $Z^\star$ can be obtained by running $Z_{k+1} = \widetilde{\mathcal{T}}_Z Z_k$, $\lim_{k\to+\infty} Z_k = Z^\star$. However, when the problem considered is of large state space, it is computationally very expensive to compute the Bellman operator $\widetilde{\mathcal{T}}_Z$ exactly. Thus function approximation might be needed to solve the problem approximately. Given a function class $\mathcal{F}$, we define the projected Z-Bellman operator as:

$$[\mathcal{T}_{Z,\mathcal{F}} Z](s,a) := \mathrm{argmin}_{Z'\in\mathcal{F}} \|Z' - \mathcal{T}_Z Z\|^2_{2,\mu}.$$

One can verify that $\mathcal{T}_{Z,\mathcal{F}} Z$ is also the minimizer of the following loss function $\mathcal{L}$,

$$\mathcal{L}(Z', Z) := \mathbb{E}_{s,a\sim\mu}\mathbb{E}_{s'\sim P_{s,a}} \left(Z'(s,a) - \exp\left(-\beta \max_{a'}(r(s',a') - \gamma\beta^{-1}\log Z(s',a'))\right)\right)^2,$$

i.e., $\mathcal{T}_{Z,\mathcal{F}} Z = \mathrm{argmin}_{Z'\in\mathcal{F}} \mathcal{L}(Z', Z)$.

We make the following assumptions on the expressive power of the function class $\mathcal{F}$

Assumption 3 (Approximate Completeness)**.** $\sup_{Z\in\mathcal{F}} \inf_{Z'\in\mathcal{F}} \|Z' - \mathcal{T}_Z Z\|_{2,\mu} \leq \epsilon_c$.

Assumption 4 (Positivity)**.** $e^{-\frac{\beta}{1-\gamma}} \leq Z \leq 1, \;\; \forall Z \in \mathcal{F}$.

Approximate the projected Z-Bellman operator with empirical loss minimization. The computation of the loss function $\mathcal{L}$ requires knowledge of the empirical model $P_{s,a}$ which the algorithm doesn't have access to. Thus, we introduce the following empirical loss to further approximate the loss function $\mathcal{L}$. Given an offline data set $\{(s_i, a_i, s_i')\}_{i=1}^N$ generated from the distribution $(s_i, a_i) \sim \mu,\ s_i' \sim P_{s_i,a_i}$, we can define the empirical loss $\widehat{\mathcal{L}}$ as:

$$\widehat{\mathcal{L}}(Z', Z) := \frac{1}{N}\sum_{i=1}^N \left(Z'(s_i, a_i) - \exp\left(-\beta \max_{a'}(r(s_i', a') - \gamma\beta^{-1}\log Z(s_i', a'))\right)\right)^2$$

Given the empirical loss $\widehat{\mathcal{L}}$ the empirical projected Bellman operator is defined as $\widehat{\mathcal{T}}_{Z,\mathcal{F}} Z := \operatorname{argmin}_{Z' \in \mathcal{F}} \widehat{\mathcal{L}}(Z', Z)$. Our robust fitted Z iteration (RFZI) is essentially updating the Z-functions iteratively by $Z_{k+1} = \widehat{\mathcal{T}}_{Z,\mathcal{F}} Z_k$. The detailed algorithm is displayed in Algorithm 1.

Algorithm 1 Robust Fitted Z Iteration (RFZI)

1: **Input:** Offline dataset $\mathcal{D} = (s_i, a_i, r_i, s_i')_{i=1}^N$, function class $\mathcal{F}$.
2: **Initialize:** $Z_0 = 1 \in \mathcal{F}$
3: **for** $k = 0, \ldots, K-1$ **do**
4: Update $Z_{k+1} = \operatorname{argmin}_{Z \in \mathcal{F}} \widehat{\mathcal{L}}(Z, Z_k)$.
5: **end for**
6: **Output:** $\pi_K = \operatorname{argmax}_a r(s,a) - \gamma\beta^{-1}\log Z_K(s,a)$

5.2 Sample complexity

This section provides the theoretical guarantee for the convergence of the RFZI algorithm. Due to space limit, we defer the proof sketches as well as detailed proofs to Appendix H.

Theorem 5 (Sample complexity for RFZI). *Suppose Assumption 2,3 and 4 hold, then for any $\delta \in (0,1)$, with probability at least $1-\delta$, the policy π_K obtained from RFQI algorithm (Algorithm 1) satisfies:*

$$\mathbb{E}_{s_0\sim\rho} V^\star(s_0) - V^{\pi_K}(s_0) \le \frac{2\gamma^K}{(1-\gamma)^2} + \gamma\beta^{-1} e^{\frac{\beta}{1-\gamma}} \frac{2C}{(1-\gamma)^2}\left(4\sqrt{\frac{2\log(|\mathcal{F}|)}{N}} + 5\sqrt{\frac{2\log(8/\delta)}{N}} + \epsilon_c\right).$$

The performance gap in Theorem 5 consists of three parts. The first part $\frac{2\gamma^K}{(1-\gamma)^2}$ captures the effect of γ-contraction of the Bellman operators. The second term, which is the term with $\gamma\beta^{-1}e^{\frac{\beta}{1-\gamma}}\frac{2C}{(1-\gamma)^2}\epsilon_c$, is related to the approximation error caused by using function approximation. The third term $\gamma\beta^{-1}e^{\frac{\beta}{1-\gamma}}\frac{2C}{(1-\gamma)^2}\left(4\sqrt{\frac{2\log(|\mathcal{F}|)}{N}} + 5\sqrt{\frac{2\log(8/\delta)}{N}}\right)$ is caused by the error of replacing the projected Z-Bellman operator with its empirical version.

Remark 6 (Comparison and Discussions). *Under similar Bellman completeness and concentrability assumptions, the sample complexity for risk-neutral offline RL [13] is $\widetilde{O}\left(C\frac{\log|\mathcal{F}|}{(1-\gamma)^4\epsilon^2}\right)$, while our result gives $O\left(C^2\left(\beta^{-1}e^{\frac{\beta}{1-\gamma}}\right)^2\frac{\log|\mathcal{F}|}{(1-\gamma)^4\epsilon^2}\right)$ (assuming $\epsilon_c = 0$). As a consequence of robustness, our bound has a worse dependency on the concentrability factor C and an additional factor $\left(\beta^{-1}e^{\frac{\beta}{1-\gamma}}\right)^2$. Note that the term $\beta^{-1}e^{\frac{\beta}{1-\gamma}}$ first decreases and then increases with β as it goes from 0 to $+\infty$, suggesting that the hyperparameter β also affects the learning difficulty of the problem. The choice of β should not be either too large or too small, ideally on the same scale with $1-\gamma$. It is still unclear to us whether the exponential dependency $e^{\frac{\beta}{1-\gamma}}$ is a proof artifact or intrinsic in our setting, however, there are results under similar settings that suggest this exponential dependency on parameter β and the effective length $\frac{1}{1-\gamma}$ is fundamental (e.g. Theorem 3 in [26]).*

We also compare our performance bound with the RFQI algorithm [64] which considers a similar offline learning setting and obtains sample complexity $O\left(\frac{\log(|\mathcal{F}||\mathcal{G}|)}{(\beta\epsilon)^2(1-\gamma)^6}\right)$, where β in their setting is the radius of the uncertainty set. Note that both results share the same dependency on ϵ and the concentrability constant C. However, the bound in [64] includes an additional term on the size of the dual variable space $\log|\mathcal{G}|$, whereas we have the exponential dependence term $e^{\frac{2\beta}{1-\gamma}}$.

6 Conclusions and Discussions

This paper proposes a new formulation of risk-sensitive MDP and establishes its equivalence with the soft robust MDP. This equivalence enables us to develop the policy gradient theorem and prove the global convergence of the exact policy gradient method under direct parameterization. Additionally, for the KL-soft robust MDP (or equivalently the risk-sensitive MDP with entropy risk measure) scenario, we propose a sample-based offline learning algorithm, namely the robust fitted-Z iteration (RFZI), and analyze its sample complexity.

Our work admittedly has its limitations. Currently, our policy gradient result is limited to the exact gradient case, and further research is needed to extend it to approximate gradients. The RFZI algorithm is specifically designed for KL-soft problems and may be more suitable for small action spaces. Our future work will focus on developing practical algorithms that can handle large or even continuous state and action spaces, as well as generalizing the approach to accommodate different penalty functions.

Acknowledgement

This work was funded by NSF AI institute: 2112085, NSF ECCS: 2328241.

References

[1] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift, 2020.

[2] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[3] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, David Heath, and Hyejin Ku. Coherent multiperiod risk adjusted values and Bellman's principle. *Annals of Operations Research*, 152: 5–22, 2007.

[4] Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR, 2021.

[5] Nicole Bäuerle and Alexander Glauner. Distributionally robust Markov decision processes and their connection to risk measures. *Mathematics of Operations Research*, 47(3):1757–1780, 2022.

[6] Bahram Behzadian, Marek Petrik, and Chin Pang Ho. Fast algorithms for l_∞-constrained s-rectangular robust MDPs. *Advances in Neural Information Processing Systems*, 34:25982–25992, 2021.

[7] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

[8] Aharon Ben-Tal, Stephen Boyd, and Arkadi Nemirovski. Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems. *Mathematical Programming*, 107(1-2):63–89, 2006.

[9] Aharon Ben-Tal, Dimitris Bertsimas, and David B Brown. A soft robust model for optimization under ambiguity. *Operations research*, 58(4-part-2):1220–1234, 2010.

[10] Aharon Ben-Tal, Ruud Brekelmans, Dick Den Hertog, and Jean-Philippe Vial. Globalized robust optimization for nonlinear uncertain inequalities. *INFORMS Journal on Computing*, 29 (2):350–366, 2017.

[11] Nicole Bäuerle and Alexander Glauner. Markov decision processes with recursive risk measures. *European Journal of Operational Research*, 296(3):953–966, 2022. ISSN 0377-2217.

[12] Özlem Çavuş and Andrzej Ruszczyński. Computational methods for risk-averse undiscounted transient Markov models. *Operations Research*, 62(2):401–417, 2014.

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

[14] Patrick Cheridito, Freddy Delbaen, and Michael Kupper. Dynamic monetary risk measures for bounded discrete-time processes. *Electronic Journal of Probability*, 2006.

[15] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[16] Anthony Coache and Sebastian Jaimungal. Reinforcement learning with dynamic convex risk measures. *Mathematical Finance*, 2023.

[17] Erick Delage, Daniel Kuhn, and Wolfram Wiesemann. "dice"-sion-making under uncertainty: When can a random decision reduce risk? *Manag. Sci.*, 65:3282–3301, 2016. URL `https://api.semanticscholar.org/CorpusID:6113617`.

[18] Freddy Delbaen. Coherent risk measures on general probability spaces. *Advances in finance and stochastics: essays in honour of Dieter Sondermann*, pages 1–37, 2002.

[19] Esther Derman and Shie Mannor. Distributional robustness and regularization in reinforcement learning. *ICML Workshop on Theoretical Foundations of Reinforcement Learning*, 2020.

[20] Esther Derman, Daniel Jaymin Mankowitz, Timothy A. Mann, and Shie Mannor. Soft-robust actor-critic policy-gradient. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

[21] Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized MDPs and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34:22274–22287, 2021.

[22] Kai Detlefsen and Giacomo Scandolo. Conditional and dynamic convex risk measures. *Finance and stochastics*, 9:539–561, 2005.

[23] Wenhao Ding, Laixi Shi, Yuejie Chi, and Ding Zhao. Seeing is not believing: Robust reinforcement learning against spurious correlation. *Advances in Neural Information Processing Systems*, 2023.

[24] Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

[25] Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. In *International Conference on Learning Representations*, 2022.

[26] Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

[27] Yingjie Fei, Zhuoran Yang, Yudong Chen, and Zhaoran Wang. Exponential Bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20436–20446, 2021.

[28] Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, pages 3198–3207. PMLR, 2021.

[29] Hans Föllmer and Irina Penner. Convex risk measures and the dynamics of their penalty functions. *Statistics & Risk Modeling*, 24(1):61–96, 2006.

[30] Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and stochastics*, 6:429–447, 2002.

[31] Hans Föllmer and Alexander Schied. Robust preferences and convex measures of risk. *Advances in Finance and Stochastics: Essays in Honour of Dieter Sondermann*, pages 39–56, 2002.

[32] Hans Föllmer and Alexander Schied. Stochastic finance. In *Stochastic Finance*. de Gruyter, 2016.

[33] Peter Geibel and Fritz Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

[34] JI González-Trejo, Onésimo Hernández-Lerma, and Luis F Hoyos-Reyes. Minimax control of discrete-time stochastic systems. *SIAM Journal on Control and Optimization*, 41(5): 1626–1659, 2002.

[35] Julien Grand-Clément and Marek Petrik. On the convex formulations of robust Markov decision processes. *arXiv preprint arXiv:2209.10187*, 2022.

[36] Aria HasanzadeZonuzy, Archana Bura, Dileep Kalathil, and Srinivas Shakkottai. Learning with safety constraints: Sample complexity of reinforcement learning for constrained MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[37] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Robust ϕ-divergence MDPs. *Advances in Neural Information Processing Systems*, 35:32680–32693, 2022.

[38] Ronald A Howard and James E Matheson. Risk-sensitive Markov decision processes. *Management science*, 18(7):356–369, 1972.

[39] Audrey Huang, Liu Leqi, Zachary C Lipton, and Kamyar Azizzadenesheli. On the convergence and optimality of policy gradient for Markov coherent risk. *arXiv preprint arXiv:2103.02827*, 2021.

[40] Vasile I Istratescu. *Fixed point theory: an introduction*. Springer, 1981.

[41] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

[42] Sebastian Jaimungal, Silvana M Pesenti, Ye Sheng Wang, and Hariom Tatsat. Robust risk-aware reinforcement learning. *SIAM Journal on Financial Mathematics*, 13(1):213–226, 2022.

[43] Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In Claude Sammut and Achim G. Hoffmann, editors, *Machine Learning, Proceedings of the Nineteenth International Conference*, pages 267–274, 2002.

[44] IS Khalil, JC Doyle, and K Glover. *Robust and optimal control*. Prentice hall, 1996.

[45] Faisal Khan, Seyed Javad Hashemi, Nicola Paltrinieri, Paul Amyotte, Valerio Cozzani, and Genserik Reniers. Dynamic risk management: a contemporary approach to process safety management. *Current opinion in chemical engineering*, 14:9–17, 2016.

[46] Ümit Köse and Andrzej Ruszczyński. Risk-averse learning by temporal difference methods with Markov risk measures. *The Journal of Machine Learning Research*, 22(1):1800–1833, 2021.

[47] Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Yehuda Levy, and Shie Mannor. Policy gradient for rectangular robust markov decision processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[48] Prashanth La and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. *Advances in neural information processing systems*, 26, 2013.

[49] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

[50] Mengmeng Li, Tobias Sutter, and Daniel Kuhn. Policy gradient algorithms for robust MDPs with non-rectangular uncertainty sets. *arXiv preprint arXiv:2305.19004*, 2023.

[51] Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust Markov decision processes. *Advances in Neural Information Processing Systems*, 26, 2013.

[52] Lynn Harold Loomis and Shlomo Sternberg. *Advanced calculus*. World Scientific, 1968.

[53] Xiaoteng Ma, Li Xia, Zhengyuan Zhou, Jun Yang, and Qianchuan Zhao. Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning. *arXiv preprint arXiv:2004.14547*, 2020.

[54] Xiaoteng Ma, Zhipeng Liang, Li Xia, Jiheng Zhang, Jose Blanchet, Mingwen Liu, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

[55] Daniel J. Mankowitz, Nir Levine, Rae Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Yuanyuan Shi, Jackie Kay, Todd Hester, Timothy Mann, and Martin Riedmiller. Robust reinforcement learning for continuous control with model misspecification. In *International Conference on Learning Representations*, 2020.

[56] Shie Mannor, Ofir Mebel, and Huan Xu. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

[57] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[58] Erfaun Noorani and John S Baras. Risk-sensitive reinforce: A monte carlo policy gradient algorithm for exponential performance criteria. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 1522–1527. IEEE, 2021.

[59] Erfaun Noorani and John S Baras. Risk-sensitive reinforcement learning and robust learning for control. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 2976–2981. IEEE, 2021.

[60] Erfaun Noorani and John S Baras. From regularization to risk-sensitivity–and back again. *IFAC-PapersOnLine*, 55(15):33–38, 2022.

[61] Erfaun Noorani, Christos Mavridis, and John Baras. Risk-sensitive reinforcement learning with exponential criteria. *arXiv preprint arXiv:2212.09010*, 2022.

[62] Takayuki Osogami. Robustness and risk-sensitivity in Markov decision processes. *Advances in Neural Information Processing Systems*, 25, 2012.

[63] Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR, 2022.

[64] Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[65] Alois Pichler, Rui Peng Liu, and Alexander Shapiro. Risk-averse stochastic programming: Time consistency and optimal stopping. *Operations Research*, 70(4):2439–2455, 2022.

[66] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017.

[67] Tongzheng Ren, Zhaolin Ren, Na Li, and Bo Dai. Stochastic nonlinear control via finite-dimensional spectral dynamic embedding. *arXiv preprint arXiv:2304.03907*, 2023.

[68] Frank Riedel. Dynamic coherent risk measures. *Stochastic processes and their applications*, 112(2):185–200, 2004.

[69] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[70] Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

[71] Andrzej Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical programming*, 125:235–261, 2010.

[72] Andrzej Ruszczyński and Alexander Shapiro. Conditional risk mappings. *Mathematics of operations research*, 31(3):544–561, 2006.

[73] Andrzej Ruszczyński and Alexander Shapiro. Optimization of convex risk functions. *Mathematics of operations research*, 31(3):433–452, 2006.

[74] Mohammad Shahidehpour, Hatim Yamin, and Zuyi Li. *Market operations in electric power systems: forecasting, scheduling, and risk management*. John Wiley & Sons, 2003.

[75] Yun Shen, Wilhelm Stannat, and Klaus Obermayer. Risk-sensitive Markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

[76] Yun Shen, Michael J Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328, 2014.

[77] Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

[78] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 2023.

[79] Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Improving robustness via risk averse distributional reinforcement learning. In *Learning for Dynamics and Control*, pages 958–968. PMLR, 2020.

[80] Songsak Sriboonchita, Wing-Keung Wong, Sompong Dhompongsa, and Hung T Nguyen. *Stochastic dominance and applications to finance, risk and economics*. CRC Press, 2009.

[81] Silvestr Stanko and Karel Macek. Risk-averse distributional reinforcement learning: A cvar optimization approach. In *IJCCI*, pages 412–423, 2019.

[82] Ke Sun, Yi Liu, Yingnan Zhao, Hengshuai Yao, Shangling Jui, and Linglong Kong. Exploring the training robustness of distributional reinforcement learning against noisy state observations. In *ECML/PKDD*, 2021.

[83] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[84] Aviv Tamar, Dotan Di Castro, and Shie Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*, pages 387–396, 2012.

[85] Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust MDPs using function approximation. In *International conference on machine learning*, pages 181–189. PMLR, 2014.

[86] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28, 2015.

[87] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338, 2017. doi: 10.1109/TAC.2016.2644871.

[88] Sebastian Thrun. Probabilistic algorithms in robotics. *Ai Magazine*, 21(4):93–93, 2000.

[89] Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee. In *International Conference on Machine Learning*, 2022.

[90] Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

[91] Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *International Conference on Machine Learning*, pages 23484–23526. PMLR, 2022.

[92] Yue Wang, Fei Miao, and Shaofeng Zou. Robust constrained reinforcement learning. *arXiv preprint arXiv:2209.06866*, 2022.

[93] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[94] Lin Xiao. On the convergence rates of policy gradient methods. *The Journal of Machine Learning Research*, 23(1):12887–12922, 2022.

[95] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

[96] Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

[97] Mengdi Xu, Zuxin Liu, Peide Huang, Wenhao Ding, Zhepeng Cen, Bo Li, and Ding Zhao. Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability. *arXiv preprint arXiv:2209.08025*, 2022.

[98] Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 9728–9754. PMLR, 25–27 Apr 2023.

[99] Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50 (6):3223–3248, 2022.

[100] Pengqian Yu and Huan Xu. Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543, 2015.

[101] Xian Yu and Lei Ying. On the global convergence of risk-averse policy gradient methods with expected conditional risk measures. In *Proceedings of the 40th International Conference on Machine Learning*, pages 40425–40451, 2023.

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

[103] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020.

[104] Kaiqing Zhang, Bin Hu, and Tamer Basar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. In *Learning for Dynamics and Control*, pages 179–190. PMLR, 2020.

[105] Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR, 2021.

A More Related Works

Relationship between risk and robustness: Apart from the works mentioned in the Introduction, there are also other works that discuss the relationship between risk and robustness under different settings or utility functions [59, 60, 25, 19]. In particular, Eysenbach and Levine [25] do not establish exact equivalence, but rather an equivalence of a lower-bound of certain robust objective; Noorani and Baras [59, 60] considers exponential utility which is a different risk measure with the risk-sensitive MDP considered in the paper as well as the Markov risk measure.

Policy Gradient for Risk-Sensitive/Robust Learning: There are many previous works focusing on applying policy gradient-based algorithm for risk-sensitive/robust learning [84, 86, 87, 39, 101, 42, 16, 89, 91, 47, 50]. However, most of the works lack theoretical guarantees on the global convergence of these gradient-based algorithms [84, 86, 87, 42, 16]. There are some recent studies focusing on the global convergence guarantees for the policy gradient theorem. In particular, the works by Huang et al. [39] and Yu and Ying [101] are most related to our work. Huang et al. [39] give negative results showing that Markov risk measures are not gradient-dominant. Yu and Ying [101] prove global convergence for one specific type of risk measure - expected conditional risk measures (ECRMs). There are also works that considers policy gradient for RMDP with different types of uncertainty sets, for example, Wang and Zou [91] considers RMDPs with a particular R-contamination ambiguity set Kumar et al. [47], Li et al. [50], Wang et al. [89] considers policy gradient for RMDP with different uncertainty sets (e.g. s-rectangularity and non-rectangularity uncertainty set), which is out of the scope of discussion of this paper.

Offline learning for robust RL: More and more attention is drawn to the offline learning of robust MDP. In the tabular setting, Zhou et al. [105] examined the uncertainty set defined by the KL divergence for offline data with uniformly lower bounded data visitation distribution. Shi and Chi [77] and Li et al. [49] provide near-optimal sample complexity bound for offline RL with weaker data coverage assumptions. There are also works focusing on offline learning with large state space, for example, Ma et al. [54] considers offline RL with linear approximation to deal with large state space. Our offline algorithm is most related to the work by Panaganti et al. [64], where they propose the fitted-Q iteration where the Q function as well as certain dual variables are approximated by possibly nonlinear functions such as neural networks.

Sample complexity for robust RL: Apart from the offline setting, there are a number of works focusing on finite-sample performance guarantees of robust RL algorithms under different data-generating mechanisms. For example, Panaganti and Kalathil [63], Yang et al. [99], Shi et al. [78] developed sample complexities for a model-based robust RL algorithm with a variety of uncertainty sets where the data are collected using a generative model. In the online learning setting, Wang and Zou [90] proposed a robust Q-learning algorithm with an R-contamination uncertain set which achieves a similar bound as its non-robust counterpart. Badrinath and Kalathil [4] proposed a model-free algorithm with linear function approximation to cope with large state spaces.

Other related works In addition to advances in the learning for RMDPs, there are also works focusing on the planning and computational facets for these problems [37, 6, 35]. Furthermore, research efforts have extended beyond theoretical considerations (e.g. [48, 66, 20, 55, 97, 103, 23]) where robust RL algorithms are proposed to handle more complicated practical problems.

A variety of different approaches have been proposed to model and address risk and robustness. Notably, in addition to Markov risk measures and robust MDPs, researchers have explored alternative methodologies, such as the use of exponential utility [38, 26, 27, 28, 58, 61], constraint MDPs [92, 15, 36, 33], distributional RL[7, 79, 53, 81, 82], and robust control [44, 104] etc.

We would also like to note that the risk-sensitive MDP with entropy risk measure (which we have proved to be equivalent to the KL-soft RMDP) is related but not identical to the exponential utility [38, 26, 27, 28, 58, 61]. There are works that discuss the equivalence of optimizing the exponential utility and solving the KL-soft RMDP in the finite-horizon undiscounted-sum setting [62]. However, the result cannot be generalized to the infinite-horizon-discounted-sum setting which is considered in this paper. Under this setting, it is unclear whether exponential utility still obtains a similar interpretation in terms of robustness.

B Numerical Simulations

In this section, we present simulation results that evaluate the exact policy gradient algorithm (see Section 3) and the RFZI algorithm (see Section 5.1) in the following environment.

Environment setups: traveling on a cycle graph. This is a environment with finite state and action space consisting of n states $\mathcal{S} = \mathbb{Z}_n$, which can be conceptually regarded as arranged on a cycle. The initial distribution is a uniformly random distribution over $\mathcal{S}$. At each state the agent is allowed to select one action from $\mathcal{A} = \{-1, 0, 1\}$ (corresponding to `left`, `stay` and `right`), and receives a reward $r(s)$ that only depends on s ($r(\cdot)$ is called the hitting reward function). As the aliases suggest, an action a at state s is supposed to move the agent to $(s + a) \mod n$. The uncertainty in the environment appears in the form of stochastic transitions, and is characterized by the probability $\alpha \in [0, \frac{1}{2}]$ of missing the expected destination by one step; i.e., transition probability is

$$P(s' \mid s, a) = \begin{cases} \alpha & s' = (s + a \pm 1) \mod n \\ 1 - 2\alpha & s' = (s + a) \mod n \\ 0 & \text{otherwise} \end{cases}.$$

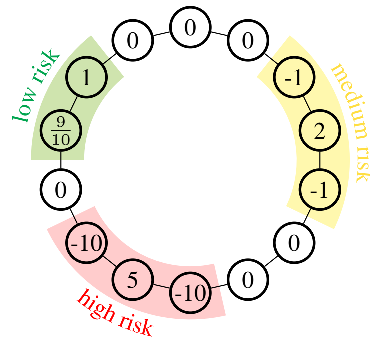


Figure 1: An exemplary 14-state environment with high-, medium-, and low-risk zones.

In this section, we focus on solving the KL-soft robust MDP problem (14) in this environment with varying penalty magnitude β.

Metrics. The performance of the algorithms may be evaluated by the following metrics:

- *Optimality gap* $\mathbb{E}_{s_0 \sim \rho}[V^\star(s_0) - V^\pi(s_0)]$, where π is the policy generated by the algorithm.
- *Average test reward* over a few test episodes (20 by default).
- *Robustness value* $\widehat{V}_\pi(\delta) := \inf_{P \in \mathcal{P}_\delta} \mathbb{E}^{\pi, P} \left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t) \,\middle|\, s_0 \sim \rho \right]$, where the model uncertainty set is selected as $\mathcal{P}_\delta := \{\widetilde{P} \mid \mathrm{KL}(\widetilde{P}_{s,a} || P_{s,a}) \le \delta, \ \forall s, a\}$.

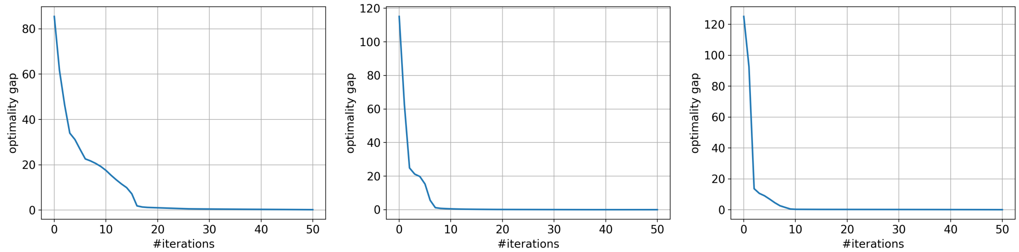
Note that the optimality gap and average test reward are usually plotted along the training trajectory, while the robustness value is usually plotted against the perturbation magnitude δ.

Exact policy gradient. We first examine the performance of the exact policy gradient algorithm under direct parameterization. Here we consider a 14-state environment as illustrated in Figure 1, where the rewards are marked in the nodes representing states. For this specific example, we can roughly classify the states into high-, medium-, and low-risk zones, as suggested in the figure.

For clarity of exposition, for now we focus on three specific settings, i.e. $(\alpha, \beta) = (0.01, 0.1)$, $(0.01, 1.0)$, $(0.15, 1.0)$ (more results can be found in Section B.1). The optimality gap curves under these settings are shown in Figure 2 below. It can be observed that the optimality gap decays to exactly 0 in all these settings, which justifies Theorem 4 that guarantees convergence of the exact policy gradient algorithm under direct parameterization. We also point out that the loss curve is a little crooked because we use projected gradient ascent, so that the optimization dynamics is not smooth when the policy at each state is pushed to the boundary.

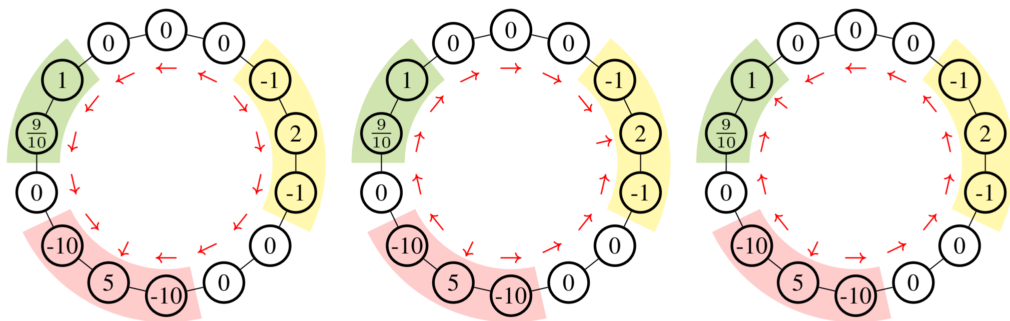
Further, we take a closer look at the learned policies in these settings. The policies are illustrated in Figure 3. To understand the role of β, we compare Figure 3a with 3b — when the uncertainty in the model (represented by α) is fixed and mild, the optimal policy for $\beta = 0.1$ is to greedily pursue the largest possible reward (i.e., staying at the state in high-risk zone with hitting reward 5); however, an agent with higher risk-sensitivity level $\beta = 1$ cares more about the potential losses caused by model uncertainty, and thus its optimal policy shifts to seeking safer options (i.e., moving to the medium-risk zone and staying there).

Similarly, to understand the role of α, we compare Figure 3b with 3c — for an agent with fixed moderate penalty magnitude β, when the noise in the model is small ($\alpha = 0.01$), it is still optimal



(a) $\alpha = 0.01, \beta = 0.1$. (b) $\alpha = 0.01, \beta = 1.0$. (c) $\alpha = 0.15, \beta = 1.0$.

Figure 2: Optimality gap curves for the exact policy gradient algorithm in different settings.



(a) $\alpha = 0.01, \beta = 0.1$ (b) $\alpha = 0.01, \beta = 1.0$ (c) $\alpha = 0.15, \beta = 1.0$

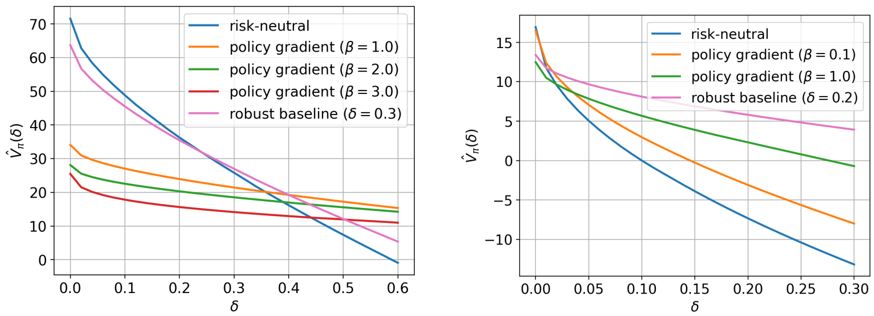
Figure 3: Illustrated policies learned by the exact policy gradient algorithm in different settings.

for the agent to stay in the medium-risk zone for a balance of risks and rewards; nevertheless, if the agent is put in a noisier environment ($\alpha = 0.15$), then its optimal policy would be directly moving to the low-risk zone to avoid potential risks.

To further examine the robustness of the risk-sensitive policies generated by the exact policy gradient algorithm, we calculate their robustness values with respect to different perturbation magnitudes δ, and plot them for comparison in Figure 4. Here the *risk-neutral* policy refers to the optimal policy of the standard risk-neutral MDP, while the *robust baseline* policy refers to the optimal policy of the RMDP with KL-rectangular ambiguity set $\mathcal{P}_\delta$ (as defined above). It can be observed that, when δ is large, the risk-sensitive policies outperform the risk-neutral policy in both settings. Meanwhile, our algorithm generally exhibits a comparable level of robustness as compared to the robust baseline that directly optimizes over RMDPs; sometimes it is even more robust than the baseline, especially when the actual ambiguity set is significantly larger than the one assumed in training (see Figure 4a for the curve where $\delta \gg 0.3$). Moreover, policies generated from higher penalty magnitude β tend to have lower robustness values when δ is small, but gradually become more robust as δ increases.

The above discussion reveals that risk-sensitive agents in face of the transition uncertainty do learn to avoid those states that could bring small instant rewards at the risk of potential future losses, and their risk-averse tendency increases when β is set larger. These numerical evidence, in turn, further motivates and justifies our focus on KL-soft RMDPs. On the one hand, learning in the context of KL-soft RMDPs are generally more computationally tractable by relatively straightforward algorithm designs, without introducing complicated optimization techniques or additional assumptions. On the other hand, the optimal policy learned in the KL-soft RMDP context does exhibit robustness in the presence of model uncertainty. These two observations, in combination, show that the research into KL-soft RMDPs may offer an alternative, analytically tractable, and potentially more accessible approach to robust reinforcement learning while maintaining comparable robust behavior.

RFZI. Now we proceed to examine the performance of the RFZI algorithm. The algorithm is tested in an 100-state environment. The hitting reward design of this environment is conceptually similar to



(a) $\alpha = 0.01$.

(b) $\alpha = 0.15$.

Figure 4: Robustness values of the generated policies with respect to different δ.
†*Risk-neutral* policy refers to the optimal policy of the risk-neutral MDP.
Robust baseline refers to the optimal policy of the RMDP with KL-rectangular ambiguity set $\mathcal{P}_\delta$.

the exemplary 14-state environment (see our code for more details). For practical implementation, the function family is selected as a 3-layer neural network, which may take any proper state-action representation as input (details deferred to Section B.1). Note that it is crucial to select good representations for efficient reinforcement learning, as the state itself might not provide sufficient information for learning (see e.g. [24] for more discussions). Further, since the minimizer in each iteration cannot be exactly calculated, instead we simply perform a batch of stochastic gradient descent updates for approximation, where for each update we sample a subset from the offline dataset. The practical algorithm is shown in Algorithm 2 below.

Algorithm 2 The practical RFZI algorithm

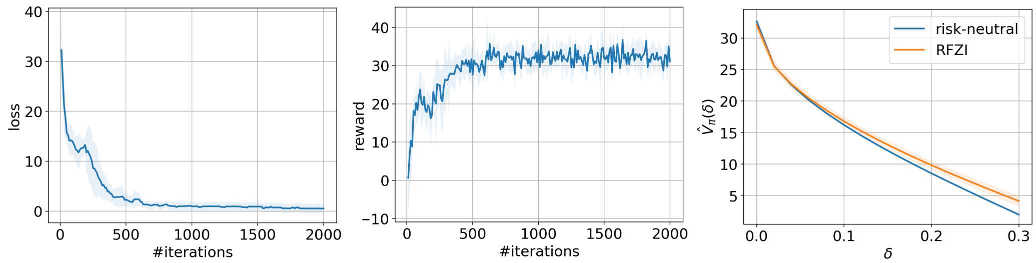
Input: Offline dataset $\mathcal{D} = (s_i, a_i, r_i, s_i')_{i=1}^N$, function family $\mathcal{F} = \{Z(\cdot;\theta) \mid \theta\}$, learning rate η, update rate τ, number of batches T_{batch}, batch size N_{batch}.
Initialize: $\theta_{\text{current}}, \theta_{\text{target}}$.
for $k = 0, \ldots, K-1$ **do**
    for $t = 1, 2, \ldots, T_{\text{batch}}$ **do**
        Sample a batch of transitions $\{(s_i, a_i, r_i, s_i') \mid i \in [N_{\text{batch}}]\}$ from the dataset $\mathcal{D}$.
        Perform gradient descent $\theta_{\text{current}} \leftarrow \theta_{\text{current}} - \eta \nabla \widehat{\mathcal{L}}(\theta_{\text{current}})$, where

$$\widehat{\mathcal{L}}(\theta_{\text{current}}) := \widehat{\mathcal{L}}(Z(\cdot;\theta_{\text{current}}), Z(\cdot;\theta_{\text{target}})).$$

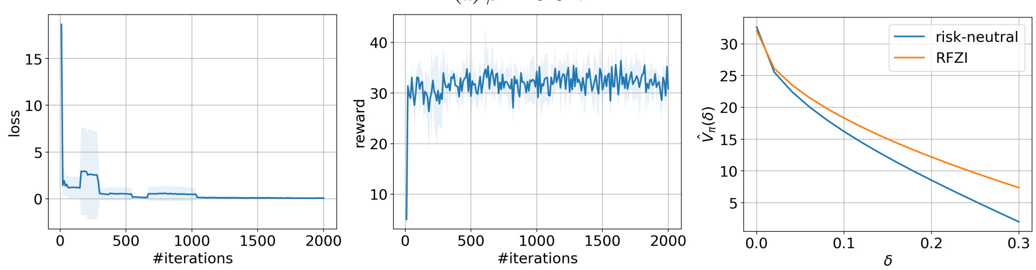
    end for
    Update: $\theta_{\text{target}} \leftarrow (1-\tau)\theta_{\text{target}} + \tau\theta_{\text{current}}$, $\theta_{\text{current}} \leftarrow \theta_{\text{target}}$.
end for
Output: $\pi_K = \text{argmax}_a[r(s,a) - \gamma\beta^{-1}\log Z(s,a;\theta_{\text{target}})]$.

Simulation results for different penalty magnitudes are shown below in Figure 5. It can be observed that in both cases the optimality gap decays to close to 0 over time, and the average test reward also converges to oscillating around a stable value. To further examine the robustness of the policies, we compare their robustness values against the risk-neutral policy (i.e., the optimal policy of the standard risk-neutral MDP). The robustness value curve suggests that the performance of our RFZI policy is more robust than the risk-neutral policy in face of model uncertainty, and the advantage increases with larger penalty magnitude β.

However, we would also like to point out some limitations of the practical RFZI algorithm. Firstly, the training dynamics becomes unstable with larger β, reflected by slower convergence of the gradient descent updates in each iteration. Additionally, the training of the network is sensitive to other hyperparameters including learning rate, number of batches and batch size, which have to be carefully tuned for satisfactory performance. It remains future work to design algorithms that are more robust and more stable with regard to the choice of the hyperparameters.



(a) $\beta = 0.01$.



(b) $\beta = 0.1$.

Figure 5: Simulation results for practical RFZI in the 100-state environment. (**Left:** optimality gap; **Middle:** average test reward; **Right:** robustness value.)

B.1 More Numerical Details

Code for reproducing the simulation results can be found at `https://github.com/huyangsh/risk-sensitive-RL_ICRL-2024`.

Exact policy gradient. Here we show the optimality gap (Figure 6) and the policies (Table 1) of a full range of experiments as specified therein. It can be verified that all converged policies are exactly the optimal policies obtained by solving the Bellman optimality equation for $Q^\star$. A closer look at the policies reveals a pattern that is similar to what we have observed before — agents learn to move to lower-risk zones when the uncertainty in the environment (i.e., α) is higher or when its penalty magnitude (i.e., β) is higher.

Table 1: Policies found by the exact policy gradient algorithm.

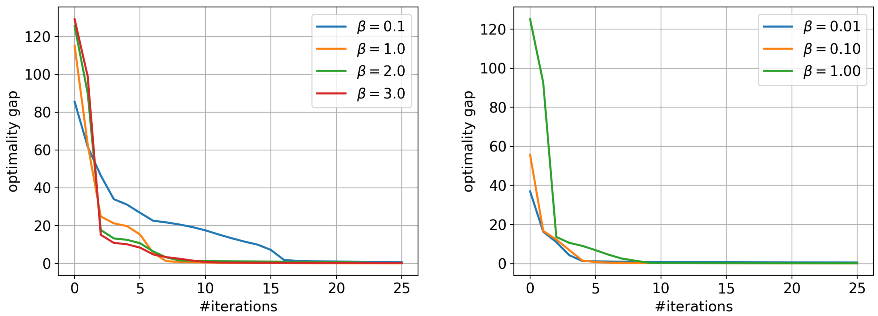
$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	**#steps ($\boldsymbol{\eta = 0.1}$)**	**policy**[†]
0.01	0.1	226	[-1, -1, 1, 1, 1, 1, 1, 1, 0, -1, -1, -1, -1, -1]
	1.0	67	[1, 1, 1, 0, -1, -1, -1, -1, 0, 1, 1, 1, 1, 1]
	2.0	76	[-1, -1, 1, 0, -1, -1, -1, -1, 0, 1, 1, 1, 0, -1]
	3.0	192	[-1, -1, -1, 0, -1, -1, -1, -1, 0, 1, 1, 1, 0, -1]
0.15	0.01	114	[1, 1, 1, 0, -1, -1, -1, -1, 0, 1, 1, 1, 0, 1]
	0.1	109	[1, 1, 1, 0, -1, -1, -1, -1, 0, 1, 1, 1, 0, -1]
	1.0	234	[-1, -1, -1, -1, -1, -1, -1, -1, 0, 1, 1, 1, 0, -1]

[†] Deterministic policies are represented by a vector in $\mathcal{A}^n$, where an entry of the vector represents the action taken at the corresponding state.

RFZI. In the implementation, we use sinusoidal embedding of states, i.e.

$$\phi(s) = \left[\sin\frac{2\pi}{N}, \sin\frac{4\pi}{N}, \ldots, \sin\frac{2N\pi}{N}, \cos\frac{2\pi}{N}, \cos\frac{4\pi}{N}, \ldots, \cos\frac{2N\pi}{N}\right],$$

which is similar to the embedding used in [67]. The Z-functions are approximated by a 3-layer network with a 256-dimensional first hidden layer and a 32-dimensional second hidden layer (both fully-connected and activated by ReLU). The output of the network is normalized by a sigmoid function to clamp the output in $(0, 1)$ (in accordance with Assumption 4).



(a) $\alpha = 0.01$. (b) $\alpha = 0.15$.

Figure 6: Optimality gap curves for the exact policy gradient algorithm in different settings.

Training details. Training is performed on a workstation equipped with a 32-core CPU (Intel® Xeon Platinum 8358, 2.60 GHz) and a NVIDIA® A100 GPU. The average training time for a typical RFZI training scheme of 2000 iterations and 500 batches is about 5 hours.

C RELATIONSHIP AND DIFFERENCE WITH MARKOV RISK MEASURES

In this section, we compare our definition of risk-sensitive MDP with the Markov risk measure. Intuitively speaking, the Markov risk measure also takes the risk generated by the randomness of the policy into account whereas our definition treats the randomness of the policy in a risk-neutral manner and only considers risk from the uncertainty of the transition probability. This intuitively explains why the two notions are equivalent for deterministic policies but not for stochastic policies. For clearness, we compare with the definition considered in [39, 86, 87], where the reward $r(s)$ is only dependent on the state s but not on action a. The Markov risk measure for policy π is defined as :

$$V^\pi_{\text{MRM}}(s_0) = r(s_0) - \gamma\sigma\left(P^\pi_{s_0}, r(s_1) - \gamma\sigma\left(P^\pi_{s_1}, r(s_2) - \gamma\sigma\left(P^\pi_{s_2}, r(s_3) - \cdots\right)\right)\right),$$

where P^π_s is the transition probability defined by $P^\pi_s(s') = \sum_a \pi(a|s)P_{s,a}(s')$. Thus, if we define the Markov-risk-measure-Bellman operator $\widetilde{\mathcal{T}}^\pi_{\text{MRM}} : \mathbb{R}^{|\mathcal{S}|\to\mathbb{R}^{|\mathcal{S}|}}$ as:

$$[\widetilde{\mathcal{T}}^\pi_{\text{MRM}} V](s) := r(s) - \gamma\sigma(P^\pi_s, V) = r(s) - \gamma\sigma(\sum_a \pi(a|s)P_{s,a}, V)$$

then the Markov risk measure is the fixed point of the Bellman operator, i.e. $V^\pi_{\text{MRM}} = \widetilde{\mathcal{T}}^\pi_{\text{MRM}} V^\pi_{\text{MRM}}$.

In contrast, the value function $\widetilde{V}^\pi$ of the risk-sensitive MDP (7) is the fixed point of the following risk-sensitive Bellman operator:

$$[\widetilde{\mathcal{T}}^\pi V](s) := r(s) - \gamma \sum_a \pi(a|s)\sigma(P_{s,a}, V)$$

Note that the risk-sensitive Bellman operator $\widetilde{\mathcal{T}}^\pi$ is linear with respect to the policy π, whereas $\widetilde{\mathcal{T}}_{\text{MRM}}$ can be potentially nonlinear w.r.t. π as σ is generally a nonlinear function. Thus for stochastic policies, V^π_{MRM} and $\widetilde{V}^\pi$ are not equivalent. However, it is not hard to verify that when π is a deterministic policy, $\widetilde{\mathcal{T}}^\pi$ and $\widetilde{\mathcal{T}}^\pi_{\text{MRM}}$ are the same

$$[\widetilde{\mathcal{T}}^\pi_{\text{MRM}} V](s) = r(s) - \gamma\sigma(P_{s,\pi(s)}, V) = [\widetilde{\mathcal{T}}^\pi V](s).$$

Thus, for deterministic policies, the value function $\widetilde{V}^\pi$ and the Markov risk measure V^π_{MRM} are the same. Additionally, when the risk-measure σ is mixture quasiconcave (c.f. [17]), it can be shown that the optimal policy π for the Markov risk measure can also be chosen as a deterministic policy; thus under this case the Markov risk measure and the risk-sensitive MDP obtain the same optimal value, i.e. $V^\star_{\text{MRM}} = \widetilde{V}^\star$. However, we would also like to emphasize that when adding policy regularization or that the risk measure σ is not mixture quasiconcave (e.g. mean semi-deviation), the optimal policy might no longer be a deterministic policy. In this setting, the optimal policy and optimal value of the Markov risk measure and the risk-sensitive MDP might not be the same.

D Proof of Lemma 1

Given a Markovian policy π and a function on the state space $\mathcal{S}$, define the Bellman operator as:

$$[\widetilde{\mathcal{T}}^\pi V](s) := \sum_a \pi(a|s)\left(r(s,a) - \gamma\sigma(P_{s,a}, V)\right), \tag{16}$$

Further, define the optimal Bellman operator $\widetilde{\mathcal{T}}^\star$ as:

$$[\widetilde{\mathcal{T}}^\star V](s) := \max_a \left(r(s,a) - \gamma\sigma(P_{s,a}, V)\right). \tag{17}$$

Lemma 3.

$$\|\widetilde{\mathcal{T}}^\pi(V' - V)\|_\infty \le \gamma\|V' - V\|_\infty, \;\; \|\widetilde{\mathcal{T}}^\star(V' - V)\|_\infty \le \gamma\|V' - V\|_\infty.$$

Proof.

$$\begin{aligned}
[\widetilde{\mathcal{T}}^\pi(V' - V)](s) &= \sum_a \pi(a|s)\left(r(s,a) - \gamma\sigma(P_{s,a}, V')\right) - \sum_a \pi(a|s)\left(r(s,a) - \gamma\sigma(P_{s,a}, V)\right) \\
&= \gamma\left(\sigma(P_{s,a}, V) - \sigma(P_{s,a}, V')\right) \\
&\le \gamma\left(\sigma_{P_{s,a}}(V' - \|V' - V\|_\infty) - \sigma(P_{s,a}, V')\right) \quad \text{(monotonicity)} \\
&= \gamma\|V' - V\|_\infty \qquad \text{(translation invariance)}.
\end{aligned}$$

Using the same analysis we can also get,

$$\begin{aligned}
&[\widetilde{\mathcal{T}}^\pi(V - V')](s) \le \gamma\|V' - V\|_\infty \\
\Longrightarrow \;\; &\|\widetilde{\mathcal{T}}^\pi(V' - V)\|_\infty \le \gamma\|V' - V\|_\infty.
\end{aligned}$$

Similarly, for $\widetilde{\mathcal{T}}^\star$,

$$\begin{aligned}
[\widetilde{\mathcal{T}}^\star(V' - V)](s) &= \max_a \left(r(s,a) - \gamma\sigma(P_{s,a}, V')\right) - \max_a \left(r(s,a) - \gamma\sigma(P_{s,a}, V)\right) \\
&= \gamma\max_a \left(\sigma(P_{s,a}, V) - \sigma(P_{s,a}, V')\right) \quad (\max_x f(x) - \max g(x) \le \max_x (f(x) - g(x))) \\
&\le \gamma\max_a \left(\sigma_{P_{s,a}}(V' - \|V' - V\|_\infty) - \sigma(P_{s,a}, V')\right) \quad \text{(monotonicity)} \\
&= \gamma\|V' - V\|_\infty \qquad \text{(translation invariance)}.
\end{aligned}$$

$$\Longrightarrow \;\; [\widetilde{\mathcal{T}}^\star(V - V')](s) \le \gamma\|V' - V\|_\infty.$$

$$\Longrightarrow \;\; \|\widetilde{\mathcal{T}}^\star(V' - V)\|_\infty \le \gamma\|V' - V\|_\infty.$$

□

Proof of Lemma 1. Lemma 1 is an immediate corollary of Lemma 3. Note that $\widetilde{V}^\pi$ and $\widetilde{V}^\star$ in (7) and (8) is the fixed point solution of

$$\widetilde{V}^\pi = \widetilde{\mathcal{T}}^\pi \widetilde{V}^\pi, \quad \widetilde{V}^\star = \widetilde{\mathcal{T}}^\star \widetilde{V}^\star.$$

From Lemma 9 and the contraction mapping theorem [40], the fixed point solution exists and is unique, which completes the proof. □

E Proof of Theorem 2

E.1 Finite horizon discounting case

We first define the value functions and Bellman operators for the finite horizon case. For any policy π (doesn't necessarily need to be stationary or Markovian), define the value function as:

$$\overline{V}^\pi_{0:h}(s) := \min_{\{\widehat{P}_t\}_{t=0}^h} \mathbb{E}_{s_t, a_t \sim \pi, \widehat{P}} \sum_{t=0}^h \gamma^t \left(r(s_t, a_t) + \gamma D(\widehat{P}_{t; s_t, a_t}, P_{s_t, a_t})\right). \tag{18}$$

Define $\overline{V}^\star_{0:h}(s)$ as

$$\overline{V}^\star_{0:h}(s) := \max_\pi \min_{\{\widehat{P}_t\}_{t=0}^h} \mathbb{E}_{s_t, a_t \sim \pi, \widehat{P}} \sum_{t=0}^h \gamma^t \left(r(s_t, a_t) + \gamma D(\widehat{P}_{t; s_t, a_t}, P_{s_t, a_t})\right). \tag{19}$$

Lemma 4. $\overline{V}^\star_{0:h}$ *is given by:*

$$\overline{V}^\star_{0:h+1} = \widetilde{\mathcal{T}}^\star \overline{V}^\star_{0:h}, \quad (\overline{V}^\star_{0:-1} := 0)$$

where $\widetilde{\mathcal{T}}^\star$ *is defined as in* (17)*. Further, for all state* s*, the max-min solution of* (19) *is given by a same set of policies and probability transitions:*

$$\pi^\star_{t|h}(s) = \operatorname*{argmax}_a \left(r(s,a) - \gamma\sigma(P_{s,a}, \overline{V}^\star_{0:h-t-1}) \right),$$
$$\widehat{P}^\star_{t|h;s,a} = \operatorname*{argmin}_{\widehat{P}} D(\widehat{P}, P_{s,a}) + \mathbb{E}_{s'\sim\widehat{P}} \overline{V}^\star_{0:h-t-1}(s').$$

Proof. We prove by induction. The statements are trivial for $h = 0$. Assume that the statements are true for $0 \le t \le h$, then for $h+1$, from the definition of $\overline{V}^\star_h$'s we have that

$$\overline{V}^\star_{0:h+1}(s_0) = \max_{\pi_0} \max_{\{\pi_t\}_{t=1}^{h+1}} \min_{\widehat{P}_0} \min_{\{\widehat{P}_t\}} \mathbb{E}_{\pi,\widehat{P}} \sum_{t=0}^{h+1} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t}) \right)$$
$$\le \max_{\pi_0} \min_{\widehat{P}_0} \max_{\{\pi_t\}_{t=1}^{h+1}} \min_{\{\widehat{P}_t\}} \mathbb{E}_{\pi,\widehat{P}} \sum_{t=0}^{h+1} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t}) \right) \quad (\max_x \min_y f(x,y) \le \min_y \max_x f(x,y))$$
$$= \max_{\pi_0} \min_{\widehat{P}_0} \mathbb{E}_{a_0,s_1\sim\pi_0,\widehat{P}_0} \left[r(s_0,a_0) + \gamma D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \left[\max_{\{\pi_t\}_{t=1}^{h+1}} \min_{\{\widehat{P}_t\}} \mathbb{E}_{\pi,\widehat{P}} \sum_{t=1}^{h+1} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t}) \Big) \right] \right]$$
$$= \max_{\pi_0} \min_{\widehat{P}_0} \mathbb{E}_{a_0,s_1\sim\pi_0,\widehat{P}_0} \left[r(s_0,a_0) + \gamma D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \gamma \overline{V}^\star_{0:h}(s_1) \right].$$

Further, from the statement that $\pi^\star_{t|h}, \widehat{P}^\star_{t|h;s,a}$ solves the min-max problem (19), we have

$$\overline{V}^\star_{0:h+1}(s_0) = \max_{\pi_0} \max_{\{\pi_t\}_{t=1}^{h+1}} \min_{\widehat{P}_0} \min_{\{\widehat{P}_t\}} \mathbb{E}_{\pi,\widehat{P}} \sum_{t=0}^{h+1} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t}) \right)$$
$$\ge \max_{\pi_0} \min_{\widehat{P}_0} \min_{\{\widehat{P}_t\}} \mathbb{E}_{\pi_0,\{\pi^\star_{t|h+1}\}_{t=1}^{h+1},\widehat{P}} \sum_{t=0}^{h+1} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t}) \right)$$
$$= \max_{\pi_0} \min_{\widehat{P}_0} \mathbb{E}_{a_0,s_1\sim\pi_0,\widehat{P}_0} \Bigg[r(s_0,a_0) + \gamma D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0})$$
$$+ \left[\min_{\{\widehat{P}_t\}_{t=1}^{h+1}} \mathbb{E}_{\{\pi^\star_{t|h+1}\}_{t=1}^{h+1},\widehat{P}} \sum_{t=1}^{h+1} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t}) \right) \right] \Bigg]$$
$$= \max_{\pi_0} \min_{\widehat{P}_0} \mathbb{E}_{a_0,s_1\sim\pi_0,\widehat{P}_0} \left[r(s_0,a_0) + \gamma D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \mathbb{E}_{\{\pi^\star_{t|h+1}\}_{t=1}^{h+1},\widehat{P}^\star_{t|h+1}} \sum_{t=1}^{h+1} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t}) \Big) \right]$$
$$= \max_{\pi_0} \min_{\widehat{P}_0} \mathbb{E}_{a_0,s_1\sim\pi_0,\widehat{P}_0} \left[r(s_0,a_0) + \gamma D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \gamma \overline{V}^\star_{0:h}(s_1) \right] \quad (\pi^\star_{t+1|h+1} = \pi^\star_{t|h}, \widehat{P}^\star_{t+1|h+1} = \widehat{P}^\star_{t|h}).$$

Thus, we may conclude that

$$\begin{aligned}
\overline{V}^\star_{0:h+1}(s_0) &= \max_{a_0} \min_{\widehat{P}_0} r(s_0,a_0) + \gamma \left(D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \mathbb{E}_{s_1\sim\widehat{P}_0} \overline{V}^\star_{0:h}(s_1) \right) \\
&= \max_{a_0} r(s_0,a_0) - \gamma\sigma(P_{s_0,a_0}, \overline{V}^\star_{0:h}) \qquad \text{(dual representation theorem)} \\
&= \widetilde{\mathcal{T}}^\star \overline{V}^\star_{0:h},
\end{aligned}$$

and that the max-min policies and probability transitions can be taken as:

$$\pi^\star_{0|h+1}(s) = \operatorname*{argmax}_a r(s,a) - \gamma\sigma(P_{s,a}, \overline{V}^\star_{0:h})$$
$$\pi^\star_{t|h+1}(s) = \pi^\star_{t-1|h}(s) = \operatorname*{argmax}_a r(s,a) - \gamma\sigma(P_{s,a}, \overline{V}^\star_{0:h-t}), \quad t \ge 1$$

$$\widehat{P}^{\star}_{0|h+1;s,a} = \operatorname*{argmin}_{\widehat{P}} D(\widehat{P}, P_{s,a}) + \mathbb{E}_{s'\sim\widehat{P}}\overline{V}^{\star}_{0:h}(s')$$
$$\widehat{P}^{\star}_{t|h+1;s,a} = \widehat{P}^{\star}_{t-1|h;s,a} = \operatorname*{argmin}_{\widehat{P}} D(\widehat{P}, P_{s,a}) + \mathbb{E}_{s'\sim\widehat{P}}\overline{V}^{\star}_{0:h-t}(s'), \quad t \geq 1.$$

The above arguments complete the proof. □

Lemma 5. *For any stationary and Markovian policy π, we have that*

$$\overline{V}^{\pi}_{0:h+1} = \widetilde{\mathcal{T}}^{\pi}\overline{V}^{\pi}_{0:h},$$

where $\widetilde{\mathcal{T}}^{\pi}$ is defined as in (16)*. Further, for all state s, minimal solution of* (18) *is given by the same set of probability transitions:*

$$\widehat{P}^{\pi}_{t|h;s,a} = \operatorname*{argmin}_{\widehat{P}} D(\widehat{P}, P_{s,a}) + \mathbb{E}_{s'\sim\widehat{P}}\overline{V}^{\pi}_{0:h-t-1}(s').$$

Proof. Our proof is largely similar to Lemma 4 and is again by induction. The statements are trivial for $h = 0$. Assume that the statements are true for $0 \leq t \leq h$, then for $h + 1$, from the definition of V^{π}_{h}'s we have that

$$\begin{aligned}
&\overline{V}^{\pi}_{h+1}(s_0) = \min_{\widehat{P}_0}\min_{\{\widehat{P}_t\}}\mathbb{E}_{\pi,\widehat{P}}\sum_{t=0}^{h+1}\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)\\
&= \min_{\widehat{P}_0}\min_{\{\widehat{P}_t\}}\mathbb{E}_{\pi_0,\widehat{P}_0}\left[r(s_0,a_0) + \gamma D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \mathbb{E}_{\pi_{1:h+1},\widehat{P}_{1:h+1}}\sum_{t=1}^{h+1}\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)\right]\\
&= \min_{\widehat{P}_0}\mathbb{E}_{\pi_0,\widehat{P}_0}\left[r(s_0,a_0) + \gamma D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \min_{\{\widehat{P}_t\}}\mathbb{E}_{\pi_{1:h+1},\widehat{P}_{1:h+1}}\sum_{t=1}^{h+1}\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)\right]\\
&= \min_{\widehat{P}_0}\mathbb{E}_{a_0,s_1\sim\pi_0,\widehat{P}_0}\left(r(s_0,a_0) + \gamma D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \gamma\overline{V}^{\pi}_{0:h}(s_1)\right)\\
&= \mathbb{E}_{a_0\sim\pi_0}\left(r(s_0,a_0) + \gamma\min_{\widehat{P}_0}\left(D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \mathbb{E}_{s_1\sim\widehat{P}_{0;s_0,a_0}}\overline{V}^{\pi}_{0:h}(s_1)\right)\right)\\
&= \mathbb{E}_{a_0\sim\pi_0}\left(r(s_0,a_0) - \gamma\sigma(P_{s_0,a_0}, \overline{V}^{\pi}_{0:h})\right).
\end{aligned}$$

Here the last step follows from dual representation theorem. Further, the minimal probability transitions can be taken as:

$$\widehat{P}^{\pi}_{0|h+1;s,a} = \operatorname*{argmin}_{\widehat{P}} D(\widehat{P}, P_{s,a}) + \mathbb{E}_{s'\sim\widehat{P}}\overline{V}^{\pi}_{0:h}(s')$$
$$\widehat{P}^{\pi}_{t|h+1;s,a} = \widehat{P}^{\pi}_{t-1|h;s,a} = \operatorname*{argmin}_{\widehat{P}} D(\widehat{P}, P_{s,a}) + \mathbb{E}_{s'\sim\widehat{P}}\overline{V}^{\pi}_{0:h-t}(s'), \quad t \geq 1.$$

The above arguments complete the proof. □

E.2 Proof of Theorem 2 (infinite horizon case)

Proof. We first verify that for $\overline{V}^{\star}, \overline{V}^{\pi}$ defined in (3), (4),

$$\lim_{h\to+\infty}\overline{V}^{\star}_{0:h} = \overline{V}^{\star}, \quad \lim_{h\to+\infty}\overline{V}^{\pi}_{0:h} = \overline{V}^{\pi}.$$

From the definition of $\overline{V}^{\pi}$, we have that

$$\begin{aligned}
\overline{V}^{\pi}(s) &= \inf_{\{\widehat{P}_t\}_{t\geq 0}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)\Big|s_0 = s\right]\\
&\leq \inf_{\{\widehat{P}_t\}_{t=0}^{h}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{h}\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)\Big|s_0 = s\right]
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}_{s_t,a_t\sim\pi,P}\left[\sum_{t=h}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(P_{s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\\
&=\overline{V}^{\pi}_{0:h}(s)+\mathbb{E}_{s_t,a_t\sim\pi,P}\left[\sum_{t=h}^{+\infty}\gamma^t\left(r(s_t,a_t)\right)\Big|s_0=s\right]\\
&\le\overline{V}^{\pi}_{0:h}(s)+\frac{\gamma^h}{1-\gamma},\\
\overline{V}^{\pi}(s)&=\inf_{\{\widehat{P}_t\}_{t\ge0}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\\
&\ge\inf_{\{\widehat{P}_t\}_{t\ge0}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{h}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\\
&=\inf_{\{\widehat{P}_t\}_{t=0}^{h}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{h}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\\
&=\overline{V}^{\pi}_{0:h}(s)\\
\implies&\ |\overline{V}^{\pi}(s)-\overline{V}^{\pi}_{0:h}(s)|\le\frac{\gamma^h}{1-\gamma}\implies\lim_{h\to+\infty}\overline{V}^{\pi}_{0:h}=\overline{V}^{\pi}.
\end{aligned}
$$

And similarly, for $\overline{V}^{\star}$, we have

$$
\begin{aligned}
\overline{V}^{\star}(s)&=\sup_{\pi}\inf_{\{\widehat{P}_t\}_{t\ge0}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right].\\
&\le\sup_{\pi}\Bigg[\inf_{\{\widehat{P}_t\}_{t=0}^{h}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{h}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\\
&\qquad+\mathbb{E}_{s_t,a_t\sim\pi,P}\left[\sum_{t=h}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(P_{s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\Bigg]\\
&\le\sup_{\pi}\inf_{\{\widehat{P}_t\}_{t=0}^{h}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{h}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\\
&\qquad+\sup_{\pi}\mathbb{E}_{s_t,a_t\sim\pi,P}\left[\sum_{t=h}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(P_{s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\\
&=\overline{V}^{\pi}_{0:h}(s)+\sup_{\pi}\mathbb{E}_{s_t,a_t\sim\pi,P}\left[\sum_{t=h}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(P_{s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\\
&\le\overline{V}^{\pi}_{0:h}(s)+\frac{\gamma^h}{1-\gamma},\\
\overline{V}^{\star}(s)&=\sup_{\pi}\inf_{\{\widehat{P}_t\}_{t\ge0}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right].\\
&\ge\sup_{\pi}\left[\inf_{\{\widehat{P}_t\}_{t=0}^{h}}\mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\left[\sum_{t=0}^{h}\gamma^t\left(r(s_t,a_t)+\gamma D(\widehat{P}_{t;s_t,a_t},P_{s_t,a_t})\right)\Big|s_0=s\right]\right]\\
&=V^{\star}_{0:h}(s)\\
\implies&\ |\overline{V}^{\star}(s)-\overline{V}^{\star}_{0:h}(s)|\le\frac{\gamma^h}{1-\gamma}\implies\lim_{h\to+\infty}\overline{V}^{\star}_{0:h}=\overline{V}^{\star}.
\end{aligned}
$$

Then from Lemma 4 and Lemma 5 we have that

$$
\overline{V}^{\star}_{0:h+1}=\widetilde{\mathcal{T}}^{\star}\overline{V}^{\star}_{0:h},\ \ \overline{V}^{\pi}_{0:h+1}=\widetilde{\mathcal{T}}^{\pi}\overline{V}^{\pi}_{0:h}.
$$

Since $\widetilde{\mathcal{T}}^\star, \widetilde{\mathcal{T}}^\pi$ is a continuous mapping, taking the limit on both sides of the equations we get

$$\overline{V}^\star = \widetilde{\mathcal{T}}^\star \overline{V}^\star, \overline{V}^\pi = \widetilde{\mathcal{T}}^\pi \overline{V}^\pi,$$

i.e., (7) and (4) obtains the same solution $\overline{V}^\pi$ and (8) and (3) obtains the same solution $\overline{V}^\star$.

Next, we will show that the claim that the minimal solution of (4) is given by (11). Since $\overline{V}^\pi$ satisfies $\overline{V}^\pi = \widetilde{\mathcal{T}}^\pi \overline{V}^\pi$, we have

$$\begin{aligned}
\overline{V}^\pi(s_0) &= \mathbb{E}_{a_0\sim\pi(\cdot|s_0)} \left(r(s_0,a_0) + \gamma \min_{\widehat{P}_0} D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \mathbb{E}_{s_1\sim \widehat{P}_{0;s_0,a_0}} \overline{V}^\pi(s_1) \right) \\
&= \mathbb{E}_{a_0\sim\pi(\cdot|s_0), s_1\sim \widehat{P}^\pi_{s_0,a_0}} \left(r(s_0,a_0) + \gamma D(\widehat{P}^\pi_{s_0,a_0}, P_{s_0,a_0}) + \gamma \overline{V}^\pi(s_1) \right).
\end{aligned}$$

Apply this equation iteratively, we get

$$\begin{aligned}
\overline{V}^\pi(s_0) &= \mathbb{E}_{a_t\sim\pi(\cdot|s_t), s_{t+1}\sim \widehat{P}^\pi_{s_t,a_t}} \left(r(s_0,a_0) + \gamma D(\widehat{P}^\pi_{s_0,a_0}, P_{s_0,a_0}) + \gamma r(s_1,a_1) + \gamma^2 D(\widehat{P}^\pi_{s_1,a_1}, P_{s_1,a_1}) + \gamma^2 \overline{V}^\pi(s_2) \right) \\
&= \ldots \\
&= \mathbb{E}_{a_t\sim\pi(\cdot|s_t), s_{t+1}\sim \widehat{P}^\pi_{s_t,a_t}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{s_t,a_t}, P_{s_t,a_t}) \right),
\end{aligned}$$

which concludes that the minimal solution is given by $\widehat{P}^\pi$ defined in (11).

For $\overline{V}^\star$. We aim to show that $\overline{V}^\star = \overline{V}^{\pi^\star}$. From the definition of $\pi^\star$ and the fact that $\overline{V}^\star = \widetilde{\mathcal{T}}^\star \overline{V}^\star$,

$$\begin{aligned}
\overline{V}^\star(s_0) &= \max_{a_0} r(s_0,a_0) - \gamma\sigma(P_{s_0,a_0}, \overline{V}^\star) \\
&\overset{(10)}{=} \mathbb{E}_{a_0\sim\pi^\star(\cdot|s_0)} \left(r(s_0,a_0) - \gamma\sigma(P_{s_0,a_0}, \overline{V}^\star) \right) \\
&= \mathbb{E}_{a_0\sim\pi^\star(\cdot|s_0)} \left(r(s_0,a_0) + \gamma \min_{\widehat{P}_0} \left(D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \mathbb{E}_{s_1\sim \widehat{P}_{0;s_0,a_0}} \overline{V}^\star(s_1) \right) \right) \\
&= \min_{\widehat{P}_0} \mathbb{E}_{a_0\sim\pi^\star(\cdot|s_0)} \left(r(s_0,a_0) + \gamma \left(D(\widehat{P}_{0;s_0,a_0}, P_{s_0,a_0}) + \mathbb{E}_{s_1\sim \widehat{P}_{0;s_0,a_0}} \overline{V}^\star(s_1) \right) \right).
\end{aligned}$$

Apply the above equation iteratively we get

$$\begin{aligned}
\overline{V}^\star(s_0) &= \min_{\{\widehat{P}_t\}_{t=1}^{+\infty}} \mathbb{E}_{a_t\sim\pi^\star(\cdot|s_t), s_{t+1}\sim \widehat{P}_{t;s_t,a_t}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{s_t,a_t}, P_{s_t,a_t}) \right) \\
&= \overline{V}^{\pi^\star}(s_0),
\end{aligned}$$

which implies that the optimal value function can be obtained by the stationary policy $\pi^\star$. Thus, the minimal transition probability is given by

$$\widehat{P}^\star_{s,a} = \widehat{P}^{\pi^\star}_{s,a} = \operatorname*{argmin}_{\widehat{P}} D(\widehat{P}, P_{s,a}) + \mathbb{E}_{s'\sim\widehat{P}} \overline{V}^\star(s'),$$

which completes the proof. □

F Proof of Theorem 3

Proof of Theorem 3. We first prove the differentiability of V^θ with respect to θ by the implicit function theorem [52]. We define the $|\mathcal{S}|$-dimentional multivariate function $F(\theta, V)$ as follows:

$$[F(\theta,V)](s) = V(s) - \sum_a \pi_\theta(a|s)(r(s,a) - \gamma\sigma(P_{s,a}, V)).$$

From the definition of the value function for Markov risk measures, V^θ is given by the following implicit function:

$$F(\theta, V^\theta) = 0.$$

Thus, from the implicit function theorem, to prove the differentiability of V^θ with respect to θ, it suffices to prove that the Jacobian matrix

$$J_{F,V}(\theta, V^\theta) = \left[\frac{\partial F_s}{\partial V_{s'}}\Big|_{V=V^\theta}\right]_{s,s'\in\mathcal{S}}$$

is invertible. Here F_s denotes the s-th entry of F and $V_{s'}$ the s'-th entry of V. From Lemma 6,

$$J_{F,V}(\theta, V^\theta) = I - \gamma \widehat{P}^\theta_{\mathcal{S}},$$

where $\widehat{P}^\theta_{\mathcal{S}}$ is a stochastic matrix. Thus, $\|\gamma \widehat{P}^\theta_{\mathcal{S}}\|_\infty \le \gamma < 1$, which implies that $I - \gamma \widehat{P}^\theta_{\mathcal{S}}$ is invertible, thus from the implicit function theorem V^θ is differentiable w.r.t. θ.

Given the differentiability, what is left is to calculate the gradient. We can further use implicit function theorem to compute the gradient, yet another easier way is through the following algebraic manipulation:

$$\begin{aligned}
V^\theta(s_0) &= \sum_{a_0} \pi_\theta(a_0|s_0)\left(r(s_0,a_0) - \gamma\sigma(P_{s_0,a_0}, V^\theta)\right) \\
\implies \quad \nabla_\theta V^\theta(s_0) &= \sum_{a_0} \nabla_\theta \pi_\theta(a_0|s_0)\left(r(s_0,a_0) - \gamma\sigma(P_{s_0,a_0}, V^\theta)\right) \\
&\quad - \gamma \sum_{a_0} \pi_\theta(a_0|s_0) \sum_{s_1} \frac{\partial \sigma(P_{s_0,a_0},\cdot)}{\partial V_{s_1}}\Big|_{V=V^\theta} \nabla_\theta V^\theta(s_1) \\
&= \sum_{a_0} \pi_\theta(a_0|s_0) Q^\theta(s_0,a_0) \nabla_\theta \log \pi_\theta(a_0|s_0) \\
&\quad + \gamma \sum_{a_0} \pi_\theta(a_0|s_0) \sum_{s_1} \widehat{P}^\theta_{s_0,a_0}(s_1) \nabla_\theta V^\theta(s_1) \quad \text{(Lemma 6)} \\
&= \mathbb{E}_{a_0\sim\pi_\theta(\cdot|s_0)} Q^\theta(s_0,a_0)\nabla_\theta \log\pi_\theta(a_0|s_0) + \gamma \mathbb{E}_{s_1\sim \widehat{P}^\theta_{s_0,a_0}} \nabla_\theta V^\theta(s_1)
\end{aligned}$$

Applying the above equation iteratively we get:

$$\begin{aligned}
&V^\theta(s_0) = \mathbb{E}_{a_0\sim\pi_\theta(\cdot|s_0)} Q^\theta(s_0,a_0)\nabla_\theta \log\pi_\theta(a_0|s_0) + \gamma \mathbb{E}_{s_1\sim \widehat{P}^\theta_{s_0,a_0}} \nabla_\theta V^\theta(s_1) \\
&= \mathop{\mathbb{E}}_{a_t\sim\pi_\theta(\cdot|s_t), s_{t+1}\sim\widehat{P}^\theta_{s_t,a_t}, t=0,1} Q^\theta(s_0,a_0)\nabla_\theta \log\pi_\theta(a_0|s_0) + \gamma Q^\theta(s_1,a_1)\nabla_\theta\log\pi_\theta(a_1|s_1) + \gamma^2 \nabla_\theta V^\theta(s_2) \\
&= \cdots \\
&= \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t), s_{t+1}\sim\widehat{P}^\theta_{s_t,a_t}} \sum_{t=1}^{+\infty} \gamma^t Q^\theta(s_t,a_t)\nabla_\theta\log\pi_\theta(a_t|s_t),
\end{aligned}$$

which completes the proof. $\square$

Lemma 6.

$$\frac{\partial \sigma(P_{s,a},\cdot)}{\partial V_{s'}}\Big|_{V=V^\theta} = -\widehat{P}^\theta_{s,a}(s'),$$

which implies that

$$J_{F,V}(\theta, V^\theta) = I - \gamma \widehat{P}^\theta_{\mathcal{S}},$$

where $\widehat{P}^\theta_{\mathcal{S}}$ is a stochastic matrix given by

$$[\widehat{P}^\theta_{\mathcal{S}}]_{s,s'} = \widehat{P}^\theta_{\mathcal{S}}(s'|s) = \sum_a \pi_\theta(a|s)\widehat{P}^\theta_{s,a}(s').$$

Proof. From the definition of $\widehat{P}^\theta_{s,a}(s')$ (11) and the dual representation theorem (Theorem 1) we have

$$\sigma(P_{s,a}, V^\theta) = -\left(\min_{\widehat{P}} \mathbb{E}_{s'\sim\widehat{P}} V^\theta(s') + D(\widehat{P}, P_{s,a})\right) = -\mathbb{E}_{s'\sim\widehat{P}^\theta_{s,a}} V^\theta(s') - D(\widehat{P}^\theta_{s,a}, P_{s,a})$$

$$\implies D(\widehat{P}^\theta_{s,a}, P_{s,a}) = -\sigma(P_{s,a}, V^\theta) - \mathbb{E}_{s'\sim\widehat{P}^\theta_{s,a}} V^\theta(s').$$

From the definition of D:

$$D(\widehat{P}^\theta_{s,a}, P_{s,a}) = \sup_V -\sigma(P_{s,a}, V) - \mathbb{E}_{s'\sim\widehat{P}^\theta_{s,a}} V(s')$$
$$\implies V^\theta = \operatorname*{argmax}_V -\sigma(P_{s,a}, V) - \mathbb{E}_{s'\sim\widehat{P}^\theta_{s,a}} V(s'),$$

thus

$$\nabla_V \left(\sigma(P_{s,a}, V) + \mathbb{E}_{s'\sim\widehat{P}^\theta_{s,a}} V(s')\right)\Big|_{V=V^\theta} = 0$$
$$\implies \frac{\partial\sigma(P_{s,a},\cdot)}{\partial V_{s'}}\Big|_{V=V^\theta} = -\frac{\partial\mathbb{E}_{s''\sim\widehat{P}^\theta_{s,a}} V(s'')}{\partial V_{s'}}\Big|_{V=V^\theta} = -\widehat{P}^\theta_{s,a}(s').$$

Then we have

$$\begin{aligned}\frac{\partial F_s}{\partial V_{s'}}\Big|_{V=V^\theta} &= \frac{\partial\left(V(s) - \sum_a \pi_\theta(a|s)(r(s,a) - \gamma\sigma(P_{s,a}, V))\right)}{\partial V_{s'}}\\ &= \mathbf{1}\{s'=s\} + \gamma\sum_a \pi_\theta(a|s)\frac{\partial\sigma(P_{s,a},\cdot)}{\partial V_{s'}}\Big|_{V=V^\theta}\\ &= \mathbf{1}\{s'=s\} - \gamma\sum_a \pi_\theta(a|s)\widehat{P}^\theta_{s,a}(s')\\ &= \mathbf{1}\{s'=s\} - \gamma[\widehat{P}^\theta_{\mathcal{S}}]_{s,s'},\end{aligned}$$

which completes the proof. □

G PROOF OF LEMMA 2 AND THEOREM 4

Before proving Lemma 2 and Theorem 4, we first introduce the performance difference lemma for soft RMDPs, which will play an important role in the following proofs. The lemma adopts from the performance difference lemma for risk-neutral MDPs (c.f. [43])

Lemma 7 (Performance Difference Lemma for soft RMDPs)**.** *Given stationary policies π', π, we have that*

$$\mathbb{E}_{s_t,a_t\sim\pi',\widehat{P}^\pi}\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t})\right) - \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}^\pi}\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t})\right)$$
$$= \frac{1}{1-\gamma}\sum_{s,a} d^{\pi^\star,\widehat{P}^\pi}(s)(\pi'(a|s) - \pi(a|s))Q^\pi(s,a)$$

Proof. For notational simplicity in this proof we also define the value function $V^{\pi,\widehat{P}}$ and Q-function $Q^{\pi,\widehat{P}}$ for a given policy π under a given probability transition $\widehat{P}$ as follows:

$$V^{\pi,\widehat{P}}(s) := \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\sum_{t=0}^{+\infty}\left[\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)|s_0 = s\right],$$
$$Q^{\pi,\widehat{P}}(s) := \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}}\sum_{t=0}^{+\infty}\left[\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)|s_0 = s, a_0 = a\right]$$

Also from Theorem 2 we know that $V^{\pi,\widehat{P}^\pi} = V^\pi, Q^{\pi,\widehat{P}^\pi} = Q^\pi$. Then we only need to show that

$$V^{\pi',\widehat{P}^\pi}(s) - V^{\pi,\widehat{P}^\pi}(s) = \frac{1}{1-\gamma}\sum_{s,a} d^{\pi^\star,\widehat{P}^\pi}(s)(\pi'(a|s) - \pi(a|s))Q^\pi(s,a).$$

The left hand side of the equation can be decomposed as

$$\mathbb{E}_{s_t,a_t\sim\pi',\widehat{P}^\pi}\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t})\right) - \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}^\pi}\sum_{t=0}^{+\infty}\gamma^t\left(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t})\right)$$

$$= \underbrace{\mathbb{E}_{s_t a_t \sim \pi', \widehat{P}^\pi} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t}) \Big) - \mathbb{E}_{a_0 \sim \pi', s_t,a_t \sim \pi, \widehat{P}^\pi, t \geq 1} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t}) \Big)}_{\text{Part A}}$$

$$+ \underbrace{\mathbb{E}_{a_0 \sim \pi', s_t,a_t \sim \pi, \widehat{P}^\pi, t \geq 1} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t}) \Big) - \mathbb{E}_{s_t,a_t \sim \pi, \widehat{P}^\pi} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t}) \Big)}_{\text{Part B}}.$$

Note that

$$\begin{aligned}
&\text{Part A} \\
&= \mathbb{E}_{a_0 \sim \pi'} \Bigg(\mathbb{E}_{s_t,a_t \sim \pi', \widehat{P}^\pi, t \geq 1} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t}) \Big) \\
&\qquad - \mathbb{E}_{s_t,a_t \sim \pi, \widehat{P}^\pi, t \geq 1} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\pi_{t;s_t,a_t}, P_{s_t,a_t}) \Big) \Bigg) \\
&= \gamma \mathbb{E}_{a_0 \sim \pi', s_1 \sim \widehat{P}^\pi} (V^{\pi', \widehat{P}^\pi}(s_1) - V^{\pi', \widehat{P}^\pi}(s_1))
\end{aligned}$$

and

$$\begin{aligned}
\text{Part B} &= \mathbb{E}_{a_0 \sim \pi'} Q^{\pi, \widehat{P}^\pi}(s_0, a_0) - \mathbb{E}_{a_0 \sim \pi} Q^{\pi, \widehat{P}^\pi}(s_0, a_0) \\
&= \sum_{a_0} (\pi'(a_0|s_0) - \pi(a_0|s_0)) Q^{\pi, \widehat{P}^\pi}(s_0, a_0).
\end{aligned}$$

Thus we get

$$\begin{aligned}
&V^{\pi', \widehat{P}^\pi}(s_0) - V^{\pi, \widehat{P}^\pi}(s_0) \\
&= \gamma \mathbb{E}_{a_0 \sim \pi', s_1 \sim \widehat{P}^\pi} (V^{\pi', \widehat{P}^\pi}(s_1) - V^{\pi', \widehat{P}^\pi}(s_1)) + \sum_{a_0} (\pi'(a_0|s_0) - \pi(a_0|s_0)) Q^{\pi, \widehat{P}^\pi}(s_0, a_0).
\end{aligned}$$

Applying this equality iteratively we get

$$\begin{aligned}
V^{\pi', \widehat{P}^\pi}(s_0) - V^{\pi, \widehat{P}^\pi}(s_0) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{a_\tau, s_\tau \sim \pi', \widehat{P}^\pi} \sum_{a_t} (\pi'(a_t|s_t) - \pi(a_t|s_t)) Q^{\pi, \widehat{P}^\pi}(s_t, a_t) \\
&= \frac{1}{1-\gamma} \sum_{s,a} d^{\pi^\star, \widehat{P}^\pi}(s) (\pi'(a|s) - \pi(a|s)) Q^\pi(s,a),
\end{aligned}$$

which completes the proof. □

Proof of Lemma 2. From Theorem 2, we have

$$\begin{aligned}
\mathbb{E}_{s_0 \sim \rho} V^\star(s_0) &= \min_{\{\widehat{P}_t\}_{t=0}^h} \mathbb{E}_{s_t,a_t \sim \pi^\star, \widehat{P}^\theta} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\theta_{t;s_t,a_t}, P_{s_t,a_t}) \Big) \\
&\leq \mathbb{E}_{s_t,a_t \sim \pi^\star, \widehat{P}^\theta} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\theta_{t;s_t,a_t}, P_{s_t,a_t}) \Big) \\
\mathbb{E}_{s_0 \sim \rho} V^\theta(s_0) &= \mathbb{E}_{s_t,a_t \sim \pi_\theta, \widehat{P}^\theta} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\theta_{t;s_t,a_t}, P_{s_t,a_t}) \Big)
\end{aligned}$$

.

Thus

$$\begin{aligned}
&\mathbb{E}_{s_0 \sim \rho} V^\star(s_0) - V^\theta(s_0) \\
&\leq \mathbb{E}_{s_t,a_t \sim \pi^\star, \widehat{P}^\theta} \sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t,a_t) + \gamma D(\widehat{P}^\theta_{t;s_t,a_t}, P_{s_t,a_t}) \Big)
\end{aligned}$$

$$
\begin{aligned}
&- \mathbb{E}_{s_t,a_t\sim\pi_\theta,\widehat{P}^\theta} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}^\theta_{t;s_t,a_t}, P_{s_t,a_t}) \right) \\
&= \frac{1}{1-\gamma} \sum_{s,a} d^{\pi^\star,\widehat{P}^\theta}(s)(\pi^\star(a|s) - \pi_\theta(a|s))Q^\theta(s,a) \quad \text{(by Lemma 7)} \\
&\leq \frac{1}{1-\gamma} \sum_{s,a} d^{\pi^\star,\widehat{P}^\theta}(s) \max_{\overline{\pi}}(\overline{\pi}(a|s) - \pi_\theta(a|s))Q^\theta(s,a) \qquad (20) \\
&\leq \frac{1}{1-\gamma} \left\| \frac{d^{\pi^\star,\widehat{P}^\theta}}{d^{\pi_\theta,\widehat{P}^\theta}} \right\|_\infty \sum_{s,a} d^{\pi_\theta,\widehat{P}^\theta}(s) \max_{\overline{\pi}}(\overline{\pi}(a|s) - \pi_\theta(a|s))Q^\theta(s,a) \\
&= \left\| \frac{d^{\pi^\star,\widehat{P}^\theta}}{d^{\pi_\theta,\widehat{P}^\theta}} \right\|_\infty \max_{\overline{\pi}} \left\langle \overline{\pi} - \pi_\theta, \mathbb{E}_{s_0\sim\rho} \nabla_\theta V^\theta(s_0) \right\rangle
\end{aligned}
$$

□

Proof of Theorem 4. For notational simplicity, we define use $\pi_s := \pi(\cdot|s)$ to denote the $|\mathcal{A}|$-dimentional probability distribution. We also use the abbreviation $Q^{(k)}, \widehat{P}^{(k)}$ to denote $Q^{\pi^{(k)}}, \widehat{P}^{\pi^{(k)}}$. We also define the following variable that will be useful throughout the proof.

$$
G_\eta^{(k)} := \frac{1}{\eta}\left(\pi_s^{(k+1)} - \pi_s^{(k)}\right)
$$

Similar to the proof of Lemma 2, we have

$$
\begin{aligned}
\mathbb{E}_{s_0\sim\rho} V^{(k)}(s_0) &= \min_{\{\widehat{P}_t\}_{t=0}^h} \mathbb{E}_{s_t,a_t\sim\pi^{(k)},\widehat{P}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t}) \right) \\
&\leq \mathbb{E}_{s_t,a_t\sim\pi^{(k)},\widehat{P}^{(k+1)}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}^{(k+1)}_{t;s_t,a_t}, P_{s_t,a_t}) \right) \\
\mathbb{E}_{s_0\sim\rho} V^{(k+1)}(s_0) &= \mathbb{E}_{s_t,a_t\sim\pi^{(k+1)},\widehat{P}^{(k+1)}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}^{(k+1)}_{t;s_t,a_t}, P_{s_t,a_t}) \right).
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\mathbb{E}_{s_0\sim\rho} V^{(k+1)}(s_0) - V^{(k)}(s_0) \geq \mathbb{E}_{s_t,a_t\sim\pi^{(k+1)},\widehat{P}^{(k+1)}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}^{(k+1)}_{t;s_t,a_t}, P_{s_t,a_t}) \right) \\
&\qquad - \mathbb{E}_{s_t,a_t\sim\pi^{(k)},\widehat{P}^{(k+1)}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t) + \gamma D(\widehat{P}^{(k+1)}_{t;s_t,a_t}, P_{s_t,a_t}) \right) \\
&= \frac{1}{1-\gamma} \sum_{s,a} d^{\pi^{(k)},\widehat{P}^{(k+1)}}(s)(\pi^{(k+1)}(a|s) - \pi^{(k)}(a|s))Q^{(k+1)}(s,a) \quad \text{(by Lemma 7)} \\
&= \underbrace{\frac{1}{1-\gamma} \sum_{s,a} d^{\pi^{(k)},\widehat{P}^{(k+1)}}(s)(\pi^{(k+1)}(a|s) - \pi^{(k)}(a|s))Q^{(k)}(s,a)}_{\text{Part I}} \\
&\quad \underbrace{+ \frac{1}{1-\gamma} \sum_{s,a} d^{\pi^{(k)},\widehat{P}^{(k+1)}}(s)(\pi^{(k+1)}(a|s) - \pi^{(k)}(a|s))(Q^{(k+1)}(s,a) - Q^{(k)}(s,a))}_{\text{Part II}}.
\end{aligned}
$$

From Lemma 13, we have

$$
\sum_a (\pi^{(k+1)}(a|s) - \pi^{(k)}(a|s))Q^{(k)}(s,a) = \langle \pi_s^{(k+1)} - \pi_s^{(k)}, Q^{(k)}(s,\cdot) \rangle
$$

$$
\begin{aligned}
&= \left\langle \mathrm{Proj}_{\Delta^{|\mathcal{A}|}} \left(\pi_s^{(k)} + \eta \frac{1}{1-\gamma} d^{\pi^{(k)},\widehat{P}^{(k)}}(s) Q^{(k)}(s,\cdot) \right), Q^{(k)}(s,\cdot) \right\rangle \\
&\geq \frac{1-\gamma}{\eta d^{\pi^{(k)},\widehat{P}^{(k)}}(s)} \|\pi_s^{(k+1)} - \pi_s^{(k)}\|_2^2.
\end{aligned}
$$

Thus

$$
\begin{aligned}
\text{Part I} &\geq \frac{1}{\eta} \sum_s \frac{d^{\pi^{(k)},\widehat{P}^{(k+1)}}(s)}{d^{\pi^{(k)},\widehat{P}^{(k)}}(s)} \|\pi_s^{(k+1)} - \pi_s^{(k)}\|_2^2 \\
&\geq \frac{1}{\eta M} \|\pi^{(k+1)} - \pi^{(k)}\|_2^2 \\
&= \frac{\eta}{M} \|G_\eta^{(k)}\|_2^2.
\end{aligned}
$$

Remark 7. *Note that for standard MDP, the corresponding Part I can be bounded by $\eta \|G_\eta^{(k)}\|_2^2$ instead of $\frac{\eta}{M}\|G_\eta^{(k)}\|_2^2$. This is the key reason why the dependency on M is worse for our setting.*

For Part II, we can bound it using Lemma 12

$$
\begin{aligned}
|\text{Part II}| &\leq \frac{1}{1-\gamma} \sum_{s,a} d^{\pi^{(k)},\widehat{P}^{(k+1)}}(s) \left| \pi^{(k+1)}(a|s) - \pi^{(k)}(a|s) \right| \left| Q^{(k+1)}(s,a) - Q^{(k)}(s,a) \right| \\
&\leq \frac{1}{(1-\gamma)^3} \sum_{s,a} d^{\pi^{(k)},\widehat{P}^{(k+1)}}(s) \left| \pi^{(k+1)}(a|s) - \pi^{(k)}(a|s) \right| \max_s \|\pi_s^{(k+1)} - \pi_s^{(k)}\|_1 \\
&\leq \frac{1}{(1-\gamma)^3} \left(\max_s \|\pi_s^{(k+1)} - \pi_s^{(k)}\|_1 \right)^2 \\
&\leq \frac{|\mathcal{A}|}{(1-\gamma)^3} \|\pi^{(k+1)} - \pi^{(k)}\|_2^2 \\
&= \frac{\eta^2 |\mathcal{A}|}{(1-\gamma)^3} \|G_\eta^{(k)}\|_2^2.
\end{aligned}
$$

Combining the bounds on Part I and II we get, for $\eta = \frac{(1-\gamma)^3}{2|\mathcal{A}|M}$

$$
\mathbb{E}_{s_0\sim\rho} V^{(k+1)}(s_0) - V^{(k)}(s_0) \geq \left(\frac{\eta}{M} - \frac{\eta^2|\mathcal{A}|}{(1-\gamma)^3} \right) \|G_\eta^{(k)}\|_2^2 = \frac{(1-\gamma)^3}{4|\mathcal{A}|M} \|G_\eta^{(k)}\|_2^2
$$

Using the telescoping technique we get

$$
\begin{aligned}
\sum_{k=0}^{K-1} \|G_\eta^{(k)}\|_2^2 &\leq \frac{4|\mathcal{A}|M^2}{(1-\gamma)^3} \sum_{k=0}^{K-1} \mathbb{E}_{s_0\sim\rho}(V^{(k+1)}(s_0) - V^{(k)}(s_0)) \\
&\leq \frac{4|\mathcal{A}|M^2}{(1-\gamma)^3} \left(V^{(K)}(s_0) - V^{(0)}(s_0) \right) \leq \frac{4|\mathcal{A}|M^2}{(1-\gamma)^4}
\end{aligned} \tag{21}
$$

where the last inequality uses the fact that $0 \leq V^\pi(s) \leq \mathbb{E}_{s_t,a_t\sim\pi,P}[\sum_{t=0}^{+\infty} \gamma^t r(s_t,a_t) \Big| s_0{=}s] \leq \frac{1}{1-\gamma}$.
Claim:

$$
\mathbb{E}_{s_0\sim\rho} V^\star(s_0) - V^{(k+1)}(s) \leq \left(M + \frac{\eta|\mathcal{A}|}{(1-\gamma)^3} \right) \|G_\eta^{(k)}\|_2
$$

Proof of Claim. From the definition of projection, for any $\pi_s' \in \Delta^{|\mathcal{A}|}$,

$$
\begin{aligned}
&\langle \pi_s^{(k)} + \eta \frac{1}{1-\gamma} d^{\pi^{(k)},\widehat{P}^{(k)}}(s) Q^{(k)}(s,\cdot) - \pi_s^{(k+1)}, \pi_s' - \pi_s^{(k+1)} \rangle \leq 0 \\
&\qquad \implies \langle Q^{(k)}(s,\cdot), \pi_s' - \pi_s^{(k+1)} \rangle \leq \frac{1-\gamma}{\eta d^{\pi^{(k)},\widehat{P}^{(k)}}(s)} \langle \pi_s^{(k+1)} - \pi_s^{(k)}, \pi_s' - \pi^{(k+1)} \rangle
\end{aligned}
$$

$$\leq \frac{(1-\gamma)M}{\eta}\|\pi_s^{(k+1)} - \pi_s^{(k)}\|_2\|\pi'_s - \pi^{(k+1)}\|_2$$
$$\leq (1-\gamma)M\|G_\eta^{(k)}\|_2\|\pi'_s - \pi^{(k+1)}\|_2.$$

Thus

$$\langle Q^{(k+1)}(s,\cdot), \pi'_s - \pi_s^{(k+1)}\rangle \leq \langle Q^{(k)}(s,\cdot), \pi'_s - \pi_s^{(k+1)}\rangle + \langle Q^{(k+1)}(s,\cdot) - Q^{(k)}(s,\cdot), \pi'_s - \pi_s^{(k+1)}\rangle$$
$$\leq (1-\gamma)M\|G_\eta^{(k)}\|_2\|\pi'_s - \pi^{(k+1)}\|_2 + \|Q^{(k+1)}(s,\cdot) - Q^{(k)}(s,\cdot)\|_\infty\|\pi'_s - \pi_s^{(k+1)}\|_1$$
$$\leq (1-\gamma)M\|G_\eta^{(k)}\|_2\|\pi'_s - \pi^{(k+1)}\|_2 + \sqrt{|\mathcal{A}|}\|Q^{(k+1)}(s,\cdot) - Q^{(k)}(s,\cdot)\|_\infty\|\pi'_s - \pi_s^{(k+1)}\|_2.$$

Further, from Lemma 12,

$$\|Q^{(k+1)}(s,\cdot) - Q^{(k)}(s,\cdot)\|_\infty \leq \frac{1}{(1-\gamma)^2}\max_{\overline{s}\in\mathcal{S}}\|\pi_{\overline{s}}^{(k+1)} - \pi_{\overline{s}}^{(k)}\|_1 \leq \frac{\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\|\pi^{(k+1)} - \pi^{(k)}\|_2$$
$$= \frac{\eta\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}\|G_\eta^{(k)}\|_2.$$

Thus

$$\langle Q^{(k+1)}(s,\cdot), \pi'_s - \pi_s^{(k+1)}\rangle \leq (1-\gamma)M\|G_\eta^{(k)}\|_2\|\pi'_s - \pi^{(k+1)}\|_2 + \frac{\eta|\mathcal{A}|}{(1-\gamma)^2}\|G_\eta^{(k)}\|_2\|\pi'_s - \pi_s^{(k+1)}\|_2$$
$$= \left((1-\gamma)M + \frac{\eta|\mathcal{A}|}{(1-\gamma)^2}\right)\|G_\eta^{(k)}\|_2\|\pi'_s - \pi^{(k+1)}\|_2$$

From the proof of Lemma 2 (inequality (20)), we have

$$\mathbb{E}_{s_0\sim\rho}V^\star(s_0) - V^{(k+1)}(s) \leq \frac{1}{1-\gamma}\sum_{s.a} d^{\pi^\star,\widehat{P}^{(k+1)}}(s)\max_{\pi'}(\pi'(a|s) - \pi^{k+1}(a|)s)Q^{(k+1)}(s,a)$$
$$= \frac{1}{1-\gamma}\sum_{s.a} d^{\pi^\star,\widehat{P}^{(k+1)}}(s)\max_{\pi'}\langle Q^{(k+1)}(s,\cdot), \pi'_s - \pi_s^{(k+1)}\rangle$$
$$\leq \frac{1}{1-\gamma}\max_s\max_{\pi'}\langle Q^{(k+1)}(s,\cdot), \pi'_s - \pi_s^{(k+1)}\rangle$$
$$\leq \frac{1}{1-\gamma}\left((1-\gamma)M + \frac{\eta|\mathcal{A}|}{(1-\gamma)^2}\right)\|G_\eta^{(k)}\|_2\|\pi'_s - \pi^{(k+1)}\|_2$$
$$= \left(M + \frac{\eta|A|}{(1-\gamma)^3}\right)\|G_\eta^{(k)}\|_2,$$

which completes the proof of the claim. □

Substitute the claim into (21) we get

$$\sum_{k=1}^K\left(\mathbb{E}_{s_0\sim\rho}V^\star(s_0) - V^{(k)}(s_0)\right)^2 \leq \frac{16|\mathcal{A}|M^4}{(1-\gamma)^4},$$

and thus

$$\min_{1\leq k\leq K}\mathbb{E}_{s_0\sim\rho}V^\star(s_0) - V^{(k)}(s_0) \leq \sqrt{\frac{16|\mathcal{A}|M^4}{(1-\gamma)^4K}}.$$

By setting

$$K \geq \frac{16|\mathcal{A}|M^4}{(1-\gamma)^4\epsilon^2},$$

we get

$$\min_{1\leq k\leq K}\mathbb{E}_{s_0\sim\rho}V^\star(s_0) - V^{(k)}(s_0) \leq \epsilon.$$

□

H Proof of Theorem 5

H.1 Proof Sketches

Before providing the full proof, in this section we first give a brief proof sketch of Theorem 5. The lemmas in the proof sketch is proved in the following sections. We define the following auxiliary variables:

$$Q_k(s,a) := r(s,a) - \gamma\beta^{-1}\log Z_k(s,a), \quad \pi_k(s) := \operatorname*{argmax}_a Q_k(s,a).$$

The proof of Theorem 5 can be decoupled into the following four steps:

Step 1: Decomposition of the performance difference. In this step, we first decompose the performance difference $\mathbb{E}_{s_0\sim\rho}(V^\star - V^{\pi_k})(s_0)$ in terms of $[Q^\star - Q_k]$ and $[Q_k - Q^\star]$.

Lemma 8.

$$\mathbb{E}_{s_0\sim\rho}(V^\star - V^{\pi_k})(s_0) \leq$$
$$\mathbb{E}_{s_0\sim\rho, s_{\tau+1}\sim\widehat{P}^{\pi_k}_{s_\tau,a_\tau}, a_\tau\sim\pi_k(s_\tau)}\sum_{t=0}^{+\infty}\gamma^t([Q^\star - Q_k](s_t,\pi^\star(s_t))) + ([Q_k - Q^\star](s_t,\pi_k(s_t))).$$

where $\widehat{P}^{\pi_k}$ is defined as $\widehat{P}^{\pi_k}(s'|s,a) \propto P(s'|s,a)\exp(-\gamma V^{\pi_k}(s'))$.

Step 2: Bound $[Q^\star - Q_k]$ and $[Q_k - Q^\star]$. Given Lemma 8, the next step is to further upper-bound $[Q^\star - Q_k]$ and $[Q_k - Q^\star]$. We have the following lemma:

Lemma 9. *For any distribution ν that is admissible,*

$$\mathbb{E}_{s,a\sim\nu}(Q_k - Q^\star)(s,a),\ \ \mathbb{E}_{s,a\sim\nu}(Q^\star - Q_k)(s,a) \leq \frac{\gamma^k}{1-\gamma} + C\sum_{m=1}^{k}\gamma^{k-m}\|Q_m - \widetilde{\mathcal{T}}_Q Q_{m-1}\|_{1,\mu},$$

where C is defined in Assumption 2.

The upper bound consists of two parts. The first part is caused by the contraction mapping of Bellman operator, and the second part captures the error of replacing the Bellman operation $\widetilde{\mathcal{T}}_Q Q_{m-1}$ with its approximation Q_m.

Step 3: Bound $\|Q_k - \widetilde{\mathcal{T}}_Q Q_{k-1}\|_{1,\mu}$. According to Lemma 9, to bound $[Q^\star - Q_k]$ and $[Q_k - Q^\star]$, we need to have a better understanding of $\|Q_k - \widetilde{\mathcal{T}}_Q Q_{k-1}\|_{1,\mu}$.

Lemma 10.

$$\|Q_k - \widetilde{\mathcal{T}}_Q Q_{k-1}\|_{1,\mu} \leq \gamma\beta^{-1}e^{\frac{\beta}{1-\gamma}}\left(\|(\widehat{\mathcal{T}}_{Z,\mathcal{F}} - \mathcal{T}_{Z,\mathcal{F}})Z_{k-1}\|_{1,\mu} + \|(\mathcal{T}_{Z,\mathcal{F}} - \mathcal{T}_Z)Z_{k-1}\|_{1,\mu}\right).$$

Lemma 10 suggests that the error $\|Q_k - \widetilde{\mathcal{T}}_Q Q_{k-1}\|_{1,\mu}$ can be bounded by two parts, the first part $\|(\mathcal{T}_{Z,\mathcal{F}} - \mathcal{T}_Z)Z_{k-1}\|_{1,\mu}$ is the error caused by function approximation, i.e., replacing the Bellman operator with the project Bellman operator. The second part is the error of replacing the projected Bellman operator with the empirical projected Bellman operator.

Step 4: Bound $\|(\widehat{\mathcal{T}}_{Z,\mathcal{F}} - \mathcal{T}_{Z,\mathcal{F}})Z_{k-1}\|_{1,\mu}$ and $\|(\mathcal{T}_{Z,\mathcal{F}} - \mathcal{T}_Z)Z_{k-1}\|_{1,\mu}$. The last step closes the proof by bounding $\|(\widehat{\mathcal{T}}_{Z,\mathcal{F}} - \mathcal{T}_{Z,\mathcal{F}})Z_{k-1}\|_{1,\mu}$ and $\|(\mathcal{T}_{Z,\mathcal{F}} - \mathcal{T}_Z)Z_{k-1}\|_{1,\mu}$ which shows up on the right hand side of Lemma 10.

Lemma 11.

$$\|(\mathcal{T}_{Z,\mathcal{F}} - \mathcal{T}_Z)Z_{k-1}\|_{2,\mu} \leq \epsilon_c.$$

With probability at least $1-\delta$,

$$\|(\widehat{\mathcal{T}}_{Z,\mathcal{F}} - \mathcal{T}_{Z,\mathcal{F}})Z_{k-1}\|_{2,\mu} \leq 4\sqrt{\frac{2\log(|\mathcal{F}|)}{N}} + 5\sqrt{\frac{2\log(8/\delta)}{N}}$$

Combining the four steps finishes the proof.

H.2 Proof of Lemmas in Proof Sketches

In this section we define the following operator $\mathcal{T}_{V\to Q}:\mathbb{R}^{|\mathcal{S}|}\to\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ for notational simplicity:

$$[\mathcal{T}_{V\to Q}V](s,a)=r(s,a)-\gamma\beta^{-1}\log\mathbb{E}_{s'\sim P(\cdot|s,a)}e^{-\beta V(s')} \tag{22}$$

Proof of Lemma 8. Define $V_k(s):=\max_a Q_k(s,a)$, then

$$\begin{aligned}
&(V^\star-V^{\pi_k})(s_0)=(V^\star-V_k)(s_0)+(V_k-V^{\pi_k})(s_0)\\
&=(Q^\star(s_0,\pi^\star(s_0))-Q_k(s_0,\pi_k(s_0)))+(Q_k(s_0,\pi_k(s_0))-Q^{\pi_k}(s_0,\pi_k(s_0)))\\
&\le(Q^\star(s_0,\pi^\star(s_0))-Q_k(s_0,\pi^\star(s_0)))+(Q_k(s_0,\pi_k(s_0))-Q^\star(s_0,\pi_k(s_0)))+(Q^\star(s_0,\pi_k(s_0))-Q^{\pi_k}(s_0,\pi_k(s_0)))\\
&\le([Q^\star-Q_k](s_0,\pi^\star(s_0)))+([Q_k-Q^\star](s_0,\pi_k(s_0)))+[\mathcal{T}_{V\to Q}(V^\star-V^{\pi_k})](s_0,\pi_k(s_0))\\
&\overset{\text{Lemma 14}}{\le}\gamma\mathbb{E}_{s_1\sim\widehat{P}^{\pi_k}_{s_0,a_0},a_0\sim\pi_k(s_0)}[V^\star-V^{\pi_k}](s_1)+([Q^\star-Q_k](s_0,\pi^\star(s_0)))+([Q_k-Q^\star](s_0,\pi_k(s_0)))
\end{aligned}$$

Apply the above inequality recursively, and we get:

$$\mathbb{E}_{s_0\sim\rho}(V^\star-V^{\pi_k})(s_0)\le\mathbb{E}_{s_0\sim\rho,s_{\tau+1}\sim\widehat{P}^{\pi_k}_{s_\tau,a_\tau},a_\tau\sim\pi_k(s_\tau)}\sum_{t=0}^{+\infty}\gamma^t([Q^\star-Q_k](s_t,\pi^\star(s_t)))+([Q_k-Q^\star](s_t,\pi_k(s_t))).$$

□

Proof of Lemma 9. We use proof by induction. The inequality naturally holds for $k=0$. Assume that it holds for $k-1$, then

$$\begin{aligned}
\mathbb{E}_{s,a\sim\nu}(Q_k-Q^\star)(s,a)&=\mathbb{E}_{s,a\sim\nu}[Q_k-\widetilde{\mathcal{T}}_QQ_{k-1}](s,a)+\mathbb{E}_{s,a\sim\nu}[\widetilde{\mathcal{T}}_Q(Q_{k-1}-Q^\star)](s,a)\\
&\overset{\text{Lemma 14}}{\le}\mathbb{E}_{s,a\sim\nu}[Q_k-\widetilde{\mathcal{T}}_QQ_{k-1}](s,a)+\gamma\mathbb{E}_{s,a\sim\nu}\mathbb{E}_{s'\sim\widehat{P}^\star_{s,a}}\max_{a'}(Q_{k-1}-Q^\star)(s',a')\\
&=\mathbb{E}_{s,a\sim\nu}[Q_k-\widetilde{\mathcal{T}}_QQ_{k-1}](s,a)+\gamma\mathbb{E}_{s',a'\sim\nu'}(Q_{k-1}-Q^\star)(s',a'),
\end{aligned}$$

where ν' is the marginal distribution on (s',a') given the joint distribution on (s,a,s',a') by $s,a\sim\nu,s'\sim\widehat{P}^\star_{s,a},a'=\mathrm{argmax}_{a'}(Q_{k-1}-Q^\star)(s',a')$. Since ν is admissible, ν' is also admissible. Thus from the induction assumption, we have

$$\begin{aligned}
\mathbb{E}_{s,a\sim\nu}(Q_k-Q^\star)(s,a)&\le\mathbb{E}_{s,a\sim\nu}[Q_k-\widetilde{\mathcal{T}}_QQ_{k-1}](s,a)+\gamma\left(\frac{\gamma^{k-1}}{1-\gamma}+C\sum_{m=1}^{k-1}\gamma^{k-1-m}\|Q_m-\widetilde{\mathcal{T}}_QQ_{m-1}\|_{1,\mu}\right)\\
&\le\frac{\gamma^k}{1-\gamma}+C\sum_{m=1}^{k}\gamma^{k-m}\|Q_m-\widetilde{\mathcal{T}}_QQ_{m-1}\|_{1,\mu}
\end{aligned}$$

□

Proof of Lemma 10. From the definition of Q_k, we have

$$\begin{aligned}
Q_k(s,a)&=r(s,a)-\gamma\beta^{-1}\log Z_k(s,a)\\
\widetilde{\mathcal{T}}_QQ_{k-1}(s,a)&=r(s,a)-\gamma\beta^{-1}\log\mathbb{E}_{s'\sim P(\cdot|s,a)}e^{-\beta\max_{a'}Q_{k-1}(s',a')}\\
&=r(s,a)-\gamma\beta^{-1}\log\mathbb{E}_{s'\sim P(\cdot|s,a)}e^{-\beta\max_{a'}\left(r(s',a')-\gamma\beta^{-1}\log Z_{k-1}(s',a')\right)}\\
&=r(s,a)-\gamma\beta^{-1}\log([\mathcal{T}_ZZ_{k-1}](s,a)).\\
\implies\quad\|Q_k-\widetilde{\mathcal{T}}_QQ_{k-1}\|_{1,\mu}&=\gamma\beta^{-1}\|\log Z_k-\log\mathcal{T}_ZZ_{k-1}\|_{1,\mu}\\
&\overset{\text{Lemma 15}}{\le}\frac{\gamma\beta^{-1}}{\min_{s,a}\min\{Z_k(s,a),\mathcal{T}_ZZ_{k-1}(s,a)\}}\|Z_k-\mathcal{T}_ZZ_{k-1}\|_{1,\mu}\\
&\le\gamma\beta^{-1}e^{\frac{\beta}{1-\gamma}}\|\widehat{\mathcal{T}}_{Z,\mathcal{F}}Z_{k-1}-\mathcal{T}_ZZ_{k-1}\|_{1,\mu}\\
&\le\gamma\beta^{-1}e^{\frac{\beta}{1-\gamma}}\left(\|(\widehat{\mathcal{T}}_{Z,\mathcal{F}}-\mathcal{T}_{Z,\mathcal{F}})Z_{k-1}\|_{1,\mu}+\|(\mathcal{T}_{Z,\mathcal{F}}-\mathcal{T}_Z)Z_{k-1}\|_{1,\mu}\right).
\end{aligned}$$

□

Proof of Lemma 11. The first inequality can be obtained by the approximate completeness assumption (Assumption 3).

$$\begin{aligned} & \sup_{Z\in\mathcal{F}} \inf_{Z'\in\mathcal{F}} \|Z' - \mathcal{T}_Z Z\|_{2,\mu} \le \epsilon_c \\ \implies & \inf_{Z'\in\mathcal{F}} \|Z' - \mathcal{T}_Z Z_{k-1}\|_{2,\mu} \le \epsilon_c \\ \implies & \text{LHS} = \|\widehat{\mathcal{T}}_{Z,\mathcal{F}} Z_{k-1} - \mathcal{T}_Z Z_{k-1}\|_{2,\mu} \le \epsilon_c \end{aligned}$$

The second inequality can be obtained by applying Lemma 3, eq(12) in [64]. Here we can set $l(Z,(s,a,s')) = \left(Z(s,a) - \exp\left(-\beta\gamma \max_{a'}(r(s',a') - \beta^{-1}\log Z_{k-1}(s',a'))\right)\right)^2$. Then it is not hard to verify that $l(Z,(s,a,s')) \le 1$, $l(Z,(s,a,s'))$ is 2-Lipschitz in Z and that $|Z(s,a)| \le 1$, thus we can set $c_1 = 1, c_2 = 1, c_3 = 2$ in eq(12) in [64] and obtain that with probability at least $1-\delta$,

$$\begin{aligned} & \mathbb{E}_{s,a\sim\mu}\mathbb{E}_{s'\sim P_{s,a}}\left[l(\widehat{\mathcal{T}}_{Z,\mathcal{F}} Z_{k-1},(s,a,s')) - l(\mathcal{T}_{Z,\mathcal{F}} Z_{k-1},(s,a,s'))\right] \le 4\sqrt{\frac{2\log(|\mathcal{F}|)}{N}} + 5\sqrt{\frac{2\log(8/\delta)}{N}} \\ \implies & \mathcal{L}(\widehat{\mathcal{T}}_{Z,\mathcal{F}} Z_{k-1}, Z_{k-1}) - \mathcal{L}(\mathcal{T}_{Z,\mathcal{F}} Z_{k-1}, Z_{k-1}) \\ & = \mathbb{E}_{s,a\sim\mu}\left(\widehat{\mathcal{T}}_{Z,\mathcal{F}} Z_{k-1}(s,a) - \mathbb{E}_{s'\sim P_{s,a}} e^{-\beta\gamma\max_{a'}(r(s',a') - \beta^{-1}\log Z(s',a'))}\right)^2 \\ & = \|(\widehat{\mathcal{T}}_{Z,\mathcal{F}} - \mathcal{T}_{Z,\mathcal{F}})Z_{k-1}\|_{2,\mu} \le 4\sqrt{\frac{2\log(|\mathcal{F}|)}{N}} + 5\sqrt{\frac{2\log(8/\delta)}{N}} \end{aligned}$$

□

H.3 Proof of Theorem 5

Proof of Theorem 5.

$$\begin{aligned} & \mathbb{E}_{s_0\sim\rho}(V^\star - V^{\pi_K})(s_0) \\ \overset{\text{Lemma 8}}{=} & \mathbb{E}_{s_0\sim\rho, s_{\tau+1}\sim\widehat{P}^{\pi_K}_{s_\tau,a_\tau}, a_\tau\sim\pi_K(s_\tau)} \sum_{t=0}^{+\infty} \gamma^t([Q^\star - Q_K](s_t,\pi^\star(s_t))) + ([Q_K - Q^\star](s_t,\pi_K(s_t))) \\ \overset{\text{Lemma 9}}{\le} & 2\sum_{t=0}^{+\infty}\gamma^t\left(\frac{\gamma^K}{1-\gamma} + C\sum_{k=1}^K \gamma^{K-k}\|Q_k - \widetilde{\mathcal{T}}_Q Q_{k-1}\|_{1,\mu}\right) \\ = & \frac{2\gamma^K}{(1-\gamma)^2} + \frac{2C}{1-\gamma}\sum_{k=1}^K \gamma^{K-k}\|Q_k - \widetilde{\mathcal{T}}_Q Q_{k-1}\|_{1,\mu} \\ \overset{\text{Lemma 10}}{\le} & \frac{2\gamma^K}{(1-\gamma)^2} + \frac{2C}{1-\gamma}\gamma\beta^{-1}e^{\frac{\beta}{1-\gamma}}\sum_{k=1}^K\gamma^{K-k}\left(\|(\widehat{\mathcal{T}}_{Z,\mathcal{F}} - \mathcal{T}_{Z,\mathcal{F}})Z_{k-1}\|_{1,\mu} + \|(\mathcal{T}_{Z,\mathcal{F}} - \mathcal{T}_Z)Z_{k-1}\|_{1,\mu}\right) \\ \le & \frac{2\gamma^K}{(1-\gamma)^2} + \frac{2C}{1-\gamma}\gamma\beta^{-1}e^{\frac{\beta}{1-\gamma}}\sum_{k=1}^K\gamma^{K-k}\left(\|(\widehat{\mathcal{T}}_{Z,\mathcal{F}} - \mathcal{T}_{Z,\mathcal{F}})Z_{k-1}\|_{2,\mu} + \|(\mathcal{T}_{Z,\mathcal{F}} - \mathcal{T}_Z)Z_{k-1}\|_{2,\mu}\right) \\ \overset{\text{Lemma 11}}{\le} & \frac{2\gamma^K}{(1-\gamma)^2} + \frac{2C}{1-\gamma}\gamma\beta^{-1}e^{\frac{\beta}{1-\gamma}}\sum_{k=1}^K\gamma^{K-k}\left(4\sqrt{\frac{2\log(|\mathcal{F}|)}{N}} + 5\sqrt{\frac{2\log(8/\delta)}{N}} + \epsilon_c\right) \\ \le & \frac{2\gamma^K}{(1-\gamma)^2} + \gamma\beta^{-1}e^{\frac{\beta}{1-\gamma}}\frac{2C}{(1-\gamma)^2}\left(4\sqrt{\frac{2\log(|\mathcal{F}|)}{N}} + 5\sqrt{\frac{2\log(8/\delta)}{N}} + \epsilon_c\right) \end{aligned}$$

□

I Auxiliaries

Lemma 12.

$$|Q^{\pi'}(s_0,a_0) - Q^\pi(s_0,a_0)| \le \frac{1}{(1-\gamma)^2}\max_s \|\pi'_s - \pi_s\|_1,$$

where π_s denotes $\pi(\cdot|s) \in \Delta^{|\mathcal{S}|}$ and $\|\pi'_s - \pi_s\|_1 = \sum_a |\pi'(a|s) - \pi(a|s)|$.

Proof.

$$
\begin{aligned}
Q^{\pi'}(s_0,a_0) &= \min_{\{\widehat{P}_t\}_{t\geq 1}} \mathbb{E}_{s_t,a_t\sim\pi',\widehat{P}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t)+\gamma d(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right) \\
&\leq \mathbb{E}_{s_t,a_t\sim\pi',\widehat{P}^{\pi}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t)+\gamma d(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right). \\
Q^{\pi}(s_0,a_0) &= \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}^{\pi}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t)+\gamma d(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right)
\end{aligned}
$$

Thus

$$
\begin{aligned}
Q^{\pi'}(s_0,a_0) - Q^{\pi}(s_0,a_0) &\leq \mathbb{E}_{s_t,a_t\sim\pi',\widehat{P}^{\pi}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t)+\gamma d(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right) \\
&\quad - \mathbb{E}_{s_t,a_t\sim\pi,\widehat{P}^{\pi}} \sum_{t=0}^{+\infty} \gamma^t \left(r(s_t,a_t)+\gamma d(\widehat{P}_{t;s_t,a_t}, P_{s_t,a_t})\right) \\
&= \mathbb{E}_{s_t\sim\pi',\widehat{P}} \sum_{t=0}^{+\infty} \gamma^t \sum_a (\pi'(a|s)-\pi(a|s)) Q^{\pi}(s,a) \quad \text{(by Lemma 7)} \\
&\leq \frac{1}{1-\gamma} \mathbb{E}_{s_t\sim\pi',\widehat{P}} \sum_{t=0}^{+\infty} \gamma^t \sum_a |\pi'(a|s)-\pi(a|s)| \\
&\leq \frac{1}{(1-\gamma)^2} \max_s \|\pi'_s - \pi_s\|_1
\end{aligned}
$$

□

Lemma 13. *For any convex set $\mathcal{X} \subset \mathbb{R}^n$ and $x \in \mathcal{X}, f \in \mathbb{R}^n, \eta > 0$,*

$$
\langle \mathrm{Proj}_{\mathcal{X}}(x+\eta f), f\rangle \geq \frac{1}{\eta} \|\mathrm{Proj}_{\mathcal{X}}(x+\eta f) - x\|_2^2.
$$

Proof. From the definition of projection, for any $y \in \mathcal{X}$,

$$
\langle y - \mathrm{Proj}_{\mathcal{X}}(x+\eta f), x+\eta f - \mathrm{Proj}_{\mathcal{X}}(x+\eta f)\rangle \leq 0,
$$

Set $y = x$ we get:

$$
\|\mathrm{Proj}_{\mathcal{X}}(x+\eta f) - x\|_2^2 \leq \langle \mathrm{Proj}_{\mathcal{X}}(x+\eta f), \eta f\rangle,
$$

which completes the proof. □

Lemma 14. *The operators $\mathcal{T}_Q, \mathcal{T}_{V\to Q}$ defined in* (15), (22) *satisfies*

$$
\begin{aligned}
[\mathcal{T}_{V\to Q}(\overline{V}-V)](s,a) &\leq \gamma \mathbb{E}_{s'\sim\widehat{P}_{s,a}} (\overline{V}-V)(s') \\
[\mathcal{T}_Q(\overline{Q}-Q)](s,a) &\leq \gamma \mathbb{E}_{s'\sim\widehat{P}_{s,a}} \max_{a'} (\overline{Q}-Q)(s',a'),
\end{aligned}
$$

where $\widehat{P}_{s,a}$ is defined as:

$$
\widehat{P}(s'|s,a) \propto P(s'|s,a)\exp(-\beta V(s')) \quad \text{or} \quad \widehat{P}(s'|s,a) \propto P(s'|s,a)\exp(-\beta \max_{a'} Q(s',a')).
$$

Proof. Let $\sigma(P_{s,a}, V) := \beta^{-1} \log \mathbb{E}_{s'\sim P_{s,a}} e^{-\beta V(s')}$, then from Example 1 and the dual representation theorem 1

$$
\sigma(P_{s,a}, V) := \sup_{\widetilde{P}_{s,a}} -\mathbb{E}_{s'\sim\widetilde{P}_{s,a}} V(s') - \beta^{-1} \mathrm{KL}(\widetilde{P}_{s,a} || P_{s,a}).
$$

$$
[\mathcal{T}_{V\to Q}(\overline{V}-V)] = -\gamma \left(\sigma(P_{s,a}, \overline{V}) - \sigma(P_{s,a}, V)\right) \tag{23}
$$

$$
\begin{aligned}
&= \gamma \left(\inf_{\widetilde{P}_{s,a}} \left(\mathbb{E}_{s' \sim \widetilde{P}_{s,a}} \overline{V}(s') + \beta^{-1} \mathrm{KL}(\widetilde{P}_{s,a} || P_{s,a}) \right) - \inf_{\widetilde{P}_{s,a}} \left(\mathbb{E}_{s' \sim \widetilde{P}_{s,a}} V(s') + \beta^{-1} \mathrm{KL}(\widetilde{P}_{s,a} || P_{s,a}) \right) \right) \\
&= \gamma \left(\inf_{\widetilde{P}_{s,a}} \left(\mathbb{E}_{s' \sim \widetilde{P}_{s,a}} \overline{V}(s') + \beta^{-1} \mathrm{KL}(\widetilde{P}_{s,a} || P_{s,a}) \right) - \left(\mathbb{E}_{s' \sim \widehat{P}_{s,a}} V(s') + \beta^{-1} \mathrm{KL}(\widehat{P}_{s,a} || P_{s,a}) \right) \right) \\
&\leq \gamma \left(\left(\mathbb{E}_{s' \sim \widehat{P}_{s,a}} \overline{V}(s') + \beta^{-1} \mathrm{KL}(\widehat{P}_{s,a} || P_{s,a}) \right) - \left(\mathbb{E}_{s' \sim \widehat{P}_{s,a}} V(s') + \beta^{-1} \mathrm{KL}(\widehat{P}_{s,a} || P_{s,a}) \right) \right) \\
&= \gamma \mathbb{E}_{s' \sim \widehat{P}_{s,a}} (\overline{V} - V)(s')
\end{aligned} \tag{24}
$$

Similarly,

$$
\begin{aligned}
[\mathcal{T}_Q(\overline{Q} - Q)](s,a) &= -\gamma\beta^{-1} \log \mathbb{E}_{s' \sim P(\cdot|s,a)} e^{-\beta \max_{a'} \overline{Q}(s',a')} + \gamma\beta^{-1} \log \mathbb{E}_{s' \sim P(\cdot|s,a)} e^{-\beta \max_{a'} Q(s',a')} \\
&= -\gamma \left(\sigma(P_{s,a}, \max_a Q(\cdot,a)) - \sigma(P_{s,a}, \max_a \overline{Q}(\cdot,a)) \right)
\end{aligned}
$$

Using the same inequality from (23) to (24) we get

$$
\begin{aligned}
-\gamma \left(\sigma(P_{s,a}, \max_a Q(\cdot,a)) - \sigma(P_{s,a}, \max_a \overline{Q}(\cdot,a)) \right) &\leq \mathbb{E}_{s' \sim \widehat{P}_{s,a}} \max_a \left(\overline{Q}(s',a) - \max_a \overline{Q}(s',a) \right) \\
&\leq \mathbb{E}_{s' \sim \widehat{P}_{s,a}} \max_a \left(\overline{Q}(s',a) - \overline{Q}(s',a) \right),
\end{aligned}
$$

which completes the proof. □

Lemma 15.

$$
|\log x - \log y| \leq \frac{1}{\min\{x,y\}} |x - y|
$$

Proof. Without loss of generality, we assume $x \geq y$, then

$$
\begin{aligned}
|\log x - \log y| &= \log x - \log y = \log(1 + \frac{x-y}{y}) \\
&\leq \frac{x-y}{y} = \frac{1}{y}(x-y) = \frac{1}{\min\{x,y\}} |x-y|,
\end{aligned}
$$

which completes the proof. □