Near-Optimal Sample Complexities of Divergence-Based S-Rectangular Distributionally Robust Reinforcement Learning

Zhenghao Li
HKUST

Shengbo Wang
University of Southern California

Nian Si
HKUST

Abstract

Distributionally robust reinforcement learning (DR-RL) has recently gained significant attention as a principled approach that addresses discrepancies between training and testing environments. To balance robustness, conservatism, and computational traceability, the literature has introduced DR-RL models with SA-rectangular and S-rectangular adversaries. While most existing statistical analyses focus on SA-rectangular models, owing to their algorithmic simplicity and the optimality of deterministic policies, S-rectangular models more accurately capture distributional discrepancies in many real-world applications and often yield more effective robust randomized policies. In this paper, we study the empirical value iteration algorithm for divergence-based S-rectangular DR-RL and establish near-optimal sample complexity bounds of $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-4}\varepsilon^{-2})$, where ε is the target accuracy, $|\mathcal{S}|$ and $|\mathcal{A}|$ denote the cardinalities of the state and action spaces, and γ is the discount factor. To the best of our knowledge, these are the first sample complexity results for divergence-based S-rectangular models that achieve optimal dependence on $|\mathcal{S}|$, $|\mathcal{A}|$, and ε simultaneously. We further validate this theoretical dependence through numerical experiments on a robust inventory control problem and a theoretical worst-case example, demonstrating the fast learning performance of our proposed algorithm.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 Introduction

Reinforcement learning (RL) Sutton and Barto (2018) is a powerful machine learning framework in which agents learn to make optimal sequential decisions through continuous interaction with an environment. While RL has achieved remarkable success across various domains, its practical deployment faces a significant challenge: real-world deployment conditions often differ from the training environment (e.g., simulations), resulting in fragile policies that fail to generalize. This mismatch undermines RLs applicability in real-world settings, where discrepancies between training and deployment are the norm.

The framework of distributionally robust reinforcement learning (DR-RL) was thus proposed in Zhou et al. (2021a) to address this mismatch and has since been further developed in a series of works, including Panaganti and Kalathil (2021); Yang et al. (2022); Xu et al. (2023); Blanchet et al. (2024); Liu et al. (2022); Wang et al. (2023a); Yang et al. (2023); Wang et al. (2024c); Shi and Chi (2023).

Popular models in distributionally robust reinforcement learning (DR-RL) include those based on SA-rectangular and S-rectangular uncertainty sets. The notion of *rectangularity*, originally introduced in the robust MDP literature to describe the adversarys temporal flexibility in selecting distributions (Iyengar, 2005), has since been refined. With the incorporation of various information structures and a growing focus on constraining adversarial power, rectangularity now serves to impose structural limitations on uncertainty sets, as elaborated in Le Tallec (2007) and Wiesemann et al. (2013). In particular, SA-rectangularity allows the adversary to choose separate distributions for each state-action pair, whereas S-rectangularity enforces consistency across actions within a given state, thereby offering a more confined modeling choice.

Existing statistical analyses of DR-RL predominantly focus on the SA-rectangular setting, primarily due to its computational tractability. Moreover, it has been

shown that SA-rectangular models always admit deterministic optimal policies. However, as illustrated in the example below, the S-rectangular formulation can be more appropriate and less conservative in certain applications, such as inventory management.

Example 1 (Inventory Model)**.** Consider a classical inventory control problem where the inventory evolves according to $S_{t+1} = S_t + A_t - D_t$, with $\{D_t : t \geq 0\}$ representing the stochastic demand process and A_t denoting the replenishment decision at time t. The reward function is $R(S_t, A_t, S_{t+1}) = p(S_t - S_{t+1} + A_t) + b\min(S_{t+1}, 0) - h\max(S_{t+1}, 0) - cA_t$, where p is the sales price, c is the purchase cost, h is the holding cost, and b is the penalty of backlog. To address the uncertainty in demand, distributionally robust reinforcement learning (DR-RL) provides a natural framework for enhancing robustness. In this context, it is reasonable to assume that the adversary can only modify the distribution of the demand D_t independently of the controllers action A_t, leading to an S-rectangular uncertainty set. By contrast, the SA-rectangular formulation allows the adversary to choose different distributions for D_t based on the controllers action $A_t$for example, assigning low demand when A_t is large and high demand when A_t is smallgranting the adversary excessive power and resulting in an unrealistic model.

This example highlights how S-rectangularity constrains the adversarys power by preventing it from adapting to the controllers actions, making it a more practical and less conservative modeling choice in applications such as inventory management.

While suitable for many applications, the S-rectangular formulation in DR-RL is more challenging than its SA-rectangular counterpart, both statistically and computationally, due to the possibility of randomized optimal policies. Computationally, this requires solving a full min-max problem rather than a simpler maximization. Fortunately, Ho et al. (2018, 2022) proposed an efficient method for performing Bellman updates in this setting. Statistically, the challenge arises from the fact that the space of randomized policies is exponentially larger than the space of deterministic policies typically sufficient under SA-rectangularity.

Another feature of Example 1 is that the reward depends on the current state S_t, the current action A_t, and the next state S_{t+1}. In contrast, the literature typically considers reward functions of the form $R(S_t, A_t)$, which depend only on the current state and action. The inventory management example highlights the necessity of adopting a reward function of the form $R(S_t, A_t, S_{t+1})$ to accurately capture the underlying dynamics.

In this work, we study the problem of learning the optimal value function in a divergence-based S-rectangular robust MDP, where the uncertainty set is defined as the sum of divergences across all actions. This formulation is well motivated in practice, as divergence-based uncertainty sets preserve absolute continuity and are widely adopted in the literature (Ho et al., 2022; Yang et al., 2022), where efficient algorithms for computing the robust value function have been developed.

However, a satisfactory analysis of the minimax statistical complexity for learning the value function remains missing. To the best of our knowledge, the current state-of-the-art upper bound in Yang et al. (2022) contains a sample complexity dependence on $|\mathcal{S}|$ and $|\mathcal{A}|$ in the form of $O(|\mathcal{S}|^2|\mathcal{A}|^2)$, where $|\mathcal{S}|$ and $|\mathcal{A}|$ are the cardinalities of the state and action spaces. This significantly deviates from the known lower bound of $\Omega(|\mathcal{S}||\mathcal{A}|)$. In addition, we have pointed out that in many models of practical interest (e.g., Example 1), the reward function depends naturally on the next state S_{t+1}, a structural feature that is often overlooked in the existing sample complexity literature.

We contribute to the literature by analyzing divergence-based S-rectangular robust MDPs with reward functions that depend on the current state, current action, and next state, i.e., $R(S_t, A_t, S_{t+1})$. We establish a sample complexity bound of $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-4}\varepsilon^{-2})$, where ε is the target accuracy and γ is the discount factor. This bound is optimal in its dependence on $|\mathcal{S}|$, $|\mathcal{A}|$, and ε, and it holds uniformly over the entire range of uncertainty sizes $\rho \in (0, +\infty)$ and discount factors $\gamma \in (0, 1)$. To the best of our knowledge, this is the first sample complexity upper bound for divergence-based S-rectangular models that simultaneously achieves optimal dependence on $|\mathcal{S}|$, $|\mathcal{A}|$, and ε.

To place our contribution in the context of the current literature, we summarize the currently available finite-sample upper and lower bounds for S-rectangular robust MDPs in Table 1.

Table 1 shows that, for divergence-based S-rectangular models, our upper bounds improve the dependence on the problem dimension from the previously known quadratic-type scaling to a linear $|\mathcal{S}||\mathcal{A}|$ dependence, up to logarithmic and instance-dependent factors.

To achieve the optimal $|\mathcal{S}||\mathcal{A}|$ dependence, we develop a refined sensitivity analysis that improves upon the metric entropy bounds derived from the covering numbers of the randomized policy class $\Pi = \{(\pi(\cdot|s))_{s\in\mathcal{S}} \mid \pi(\cdot|s) \in \Delta(\mathcal{A})\}$, where $\Delta(\mathcal{A})$ denotes the probability simplex over $\mathcal{A}$, as used in Yang et al. (2022). Moreover, our analyses advance the techniques of Wang et al. (2024c) by relaxing the mutual absolute continuity requirement, thereby extending the allow-

Type	Ref.	Set	Sample complexity for S-rec
Upper Bound	Yang et al. (2022)	KL	$\widetilde{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{\varepsilon^2\rho^2 p^2(1-\gamma)^4}\right)$
Upper Bound	Yang et al. (2022)	χ^2	$\widetilde{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^3(1+\rho)^2}{\varepsilon^2(\sqrt{1+\rho}-1)^2(1-\gamma)^4}\right)$
Upper Bound	This paper	KL, f_k	$\widetilde{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\mathfrak{p}_\wedge\varepsilon^2}\right)$
Lower Bound	Yang et al. (2022)	χ^2	$\widetilde{\Omega}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\varepsilon^2(1-\gamma)^2}\min\left\{\frac{1}{1-\gamma},\frac{1}{\rho}\right\}\right)$

Table 1: Comparison of currently available finite-sample upper and lower bounds for S-rectangular robust MDPs. The currently available lower bound is for the χ^2 ambiguity set from Yang et al. (2022). Our upper bounds improve the dependence on the state-action dimension for divergence-based ambiguity sets, up to logarithmic and instance-dependent factors.

able range of the uncertainty radius to $\mathbb{R}+$, beyond the previously restrictive regime of $\rho = O(\mathfrak{p}_\wedge)$, while retaining an $O(1)$ dependence on ρ as $\rho \downarrow 0$.

The remainder of this paper is organized as follows: Section 2 briefly reviews related work on SA-rectangular and S-rectangular distributionally robust reinforcement learning. Section 3 introduces the framework for learning S-rectangular distributionally robust Markov Decision Processes. Section 4 establishes sample complexity upper bounds for value function estimation. Section 5 presents numerical experiments to support our theoretical results.

2 Literature Review

In this section, we briefly survey SA-rectangular and S-rectangular distributionally robust reinforcement learning.

SA-rectangular DR-RL: The dynamic programming principles for SA-rectangular distributionally robust Markov decision processes (DR-MDPs) have been gradually established through a series of works under different information structures (González-Trejo et al., 2002; Iyengar, 2005; Nilim and El Ghaoui, 2005; Shapiro, 2022; Wang et al., 2024b). Recent advances in SA-rectangular distributionally robust reinforcement learning (DR-RL) have explored sample complexity in various settings. Broadly speaking, model-based approaches have been studied in Zhou et al. (2021b); Panaganti and Kalathil (2021); Yang et al. (2022); Shi and Chi (2022); Xu et al. (2023); Shi et al. (2023); Blanchet et al. (2024), while the statistical properties of model-free algorithms are presented in Liu et al. (2022); Wang et al. (2023a,b); Yang et al. (2023).

S-rectangular DR-RL: To extend the flexibility of robust MDP models, S-rectangularity was introduced in Xu and Mannor (2010); Wiesemann et al. (2013) as an overarching theoretical framework to constrain the adversary while retaining a dynamic programming equation. Ho et al. (2018, 2022) developed an efficient optimization algorithm to solve the Bellman update within this framework. Subsequently, Kumar et al. (2024) improved upon their work by proposing a faster algorithm for the L_p uncertainty sets. On the statistical side, Yang et al. (2022) provided the first sample complexity result for S-rectangular DR-RL, achieving a rate of $\widetilde{O}(|\mathcal{S}|^2|\mathcal{A}|^2(1-\gamma)^{-4}\varepsilon^{-2})$, which is suboptimal in its dependence on the number of states and actions. More recently, Clavier et al. (2024) established near-optimal rates for the S-rectangular setting under general L_p norm uncertainty sets. However, their analysis does not directly extend to divergence-based uncertainty sets.

3 Learning S-rectangular Robust Markov Decision Processes

3.1 Classical Markov Decision Processes

We briefly review and establish notation for classical tabular MDP models. Let $\Delta(\mathcal{S}), \Delta(\mathcal{A})$ denote the probability simplex over the finite state space $\mathcal{S}$ and action space $\mathcal{A}$ respectively. An infinite horizon MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the finite state and action spaces, respectively; $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the reward function; $P = \{P_{s,a}(\cdot) \in \Delta(\mathcal{S}) : (s, a) \in \mathcal{S} \times \mathcal{A}\}$ is the controlled transition kernel; and $\gamma \in (0, 1)$ is the discount factor. Throughout the paper, given a controlled transition kernel P, we denote $P_s := (P_{s,a})_{a\in\mathcal{A}}$ which is seen as a function $P_s : A \to \Delta(\mathcal{S})$.

We define the measurable space $(\Omega, \mathcal{F})$ to be the canonical space $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$ equipped with the σ-field generated by cylinder sets. Define state-action process $(S_t, A_t)_{t\geq 0}$ by the point evaluation $S_t(\omega) = s_t, A_t(\omega) = a_t$ for all $t \geq 0$ and $\omega = (s_0, a_0, s_1, a_1, \dots) \in \Omega$.

An agent may optimize over the class of history-dependent policies, denoted by Π_{HD}, where each policy $\pi = (\pi_t)_{t\geq 0} \in \Pi_{\text{HD}}$ is a sequence of decision rules. Each decision rule π_t at time t specifies the conditional

distribution of the action A_t given the full history, that is, a mapping $\pi_t : (\mathcal{S}\times\mathcal{A})^t \times \mathcal{S} \to \Delta(\mathcal{A})$. In the setting of classical infinite-horizon discounted MDPs, it is well known that optimal decision-making can be achieved using stationary, Markov, deterministic policies, denoted Π_{D}, where each policy is a mapping $\pi : \mathcal{S} \to \mathcal{A}$ (Puterman, 2009).

However, in the context of S-rectangular DRMDPs, policies in Π_{D} may fail to attain the optimal performance achievable within the broader class Π_{HD} (Wiesemann et al., 2013). In this setting, it suffices to consider stationary, Markov, randomized policies, which we denote by Π throughout the paper. Each $\pi \in \Pi$ is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, specifying a conditional distribution over actions given the current state S_t, uniformly for all $t \geq 0$. Given this sufficiency, we restrict our attention to policies in the class Π for the remainder of the paper.

Given a controlled transition kernel P of a classical MDP, a policy $\pi \in \Pi$ and an initial distribution $\mu \in \Delta(\mathcal{S})$ uniquely defines a probability measure on $(\Omega, \mathcal{F})$. We will always assume that μ is the uniform distribution over $\mathcal{S}$. The expectation under this measure is denoted by E_P^π. The infinite horizon discounted value V_P^π is defined as:

$$V_P^\pi(s) := E_P^\pi\left[\sum_{t=0}^{\infty}\gamma^t R(S_t, A_t, S_{t+1})\middle| S_0 = s\right].$$

An optimal policy $\pi^* \in \Pi$ achieves the optimal value $V_P^*(s) := \max_{\pi\in\Pi} V_P^\pi(s)$.

It is well known that the optimal value function is the unique solution of the following *Bellman equation*:

$$v(s) = \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} P_{s,a}(s')(R(s,a,s') + \gamma v(s')). \quad (3.1)$$

Let v^* be the unique solution, then any deterministic policy $\pi^* : \mathcal{S} \to \mathcal{A}$ with $\pi^*(s) \in \arg\max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} P_{s,a}(s')(R(s,a,s') + \gamma v^*(s'))$ will achieve the optimal value $V_P^*(s)$.

3.2 Robust MDPs and S-Rectangularity

Robust MDPs extend standard MDP models by introducing an adversary that perturbs the transition dynamics within a prescribed uncertainty set $\mathcal{P}$, aiming to minimize the control value achieved by the decision maker. This formulation gives rise to a dynamic zero-sum game between the controller and the adversary. Consequently, the controller must account for potential model misspecifications represented by the adversary perturbation, leading to the design of more robust policies.

The statistical complexity of policy learning in robust MDPs has been primarily studied under SA- and S-rectangular uncertainty sets. As discussed in the previous section, S-rectangularity generalizes SA-rectangular models and provides a more expressive framework for modeling adversarial perturbations, constraining the adversary in a structured way while preserving the dynamic programming principle. From this point forward, we will be focusing on S-rectangular robust MDPs.

Definition 1 (Wiesemann et al. (2013), S-rectangularity)**.** The uncertainty set $\mathcal{P}$ is S-rectangular if $\mathcal{P} = \bigtimes_{s\in\mathcal{S}} \mathcal{P}_s$ for some $\mathcal{P}_s \subseteq \{(\psi_a)_{a\in\mathcal{A}} | \psi_a \in \Delta(\mathcal{S}), \forall a \in \mathcal{A}\}$ for all $s \in \mathcal{S}$.

We focus on a special class of S-rectangular adversarial uncertainty sets, where the controlled transition kernels are perturbations of a nominal kernel $\overline{P}$. These sets are defined via a divergence function f and a radius parameter ρ. The computational methods and statistical complexity associated with this type of uncertainty structure have been extensively studied in the literature (Yang et al., 2022; Ho et al., 2018).

Specifically, given a divergence function f, i.e. $f : \mathbb{R}_+ \to \mathbb{R}$ is convex with $f(1) = 0$ and $f(0) = \lim_{t\downarrow 0} f(t)$, we consider the S-rectangular uncertainty set $\mathcal{P}(f,\rho) = \bigtimes_{s\in\mathcal{S}} \mathcal{P}_s(f,\rho)$ under f-divergence and radius ρ where

$$\mathcal{P}_s(f,\rho) = \left\{ P_s \in \Delta(\mathcal{S})^{|\mathcal{A}|} \middle| P_{s,a} \ll \overline{P}_{s,a}, \sum_{s'\in\mathcal{S}, a\in\mathcal{A}} \overline{P}_{s,a}(s') f\left(\frac{P_{s,a}(s')}{\overline{P}_{s,a}(s')}\right) \leq |\mathcal{A}|\rho \right\}. \quad (3.2)$$

Here, $\ll$ denotes absolute continuity; i.e. a probability measure $p \in \Delta(\mathcal{S})$ is absolutely continuous with respect to $q \in \Delta(\mathcal{S})$, denoted by $p \ll q$, if $q(s) = 0$ implies $p(s) = 0$ for any $s \in \mathcal{S}$. The dependence of the uncertainty set on (f,ρ) is suppressed when there is no ambiguity.

Given a policy $\pi \in \Pi_{\mathrm{HD}}$ and uncertainty set $\mathcal{P} = \mathcal{P}(f,\rho)$, the robust value function of π is

$$V_{\mathcal{P}}^\pi(s) = \inf_{P\in\mathcal{P}} E_P^\pi\left[\sum_{t=0}^{\infty}\gamma^t R(S_t, A_t, S_{t+1})\middle| S_0 = s\right] \quad (3.3)$$

for all $s \in \mathcal{S}$. The optimal value, defined as $V_{\mathcal{P}}^*(s) := \sup_{\pi\in\Pi_{\mathrm{HD}}} V_{\mathcal{P}}^\pi(s)$, is achieved by $\pi^* \in \Pi$.

Definition 2 (DR Bellman Equation)**.** Given S-rectangular $\mathcal{P} = \bigtimes_{s\in\mathcal{S}} \mathcal{P}_s$, the DR Bellman equation

is the following fixed-point equation for $v : \mathcal{S} \to \mathbb{R}$

$$v(s) = \sup_{\phi \in \Delta(\mathcal{A})} \inf_{P_s \in \mathcal{P}_s} \sum_{a \in \mathcal{A}} \phi(a) \Bigg[\sum_{s' \in \mathcal{S}} P_{s,a}(s') \left(R(s,a,s') + \gamma v(s') \right) \Bigg]. \tag{3.4}$$

It is well known (Wiesemann et al., 2013) that for $\mathcal{P} = \mathcal{P}(f, \rho)$ the optimal value $V^*_{\mathcal{P}}$ is the unique solution v^* to (3.4).

We note that the value function in (3.3) assumes an adversary that fixes a controlled transition kernel over the entire control horizon, a setting commonly referred to as a static or time-homogeneous adversarial model (Iyengar, 2005; Wiesemann et al., 2013; Wang et al., 2024b). This framework can be extended to more general Markovian or history-dependent adversarial models, while still preserving Markov optimality (Wang et al., 2024b).

To facilitate our analysis, we define the DR Bellman operators as follows.

Definition 3 (DR Bellman Operators)**.** Given uncertainty set $\mathcal{P} = \mathcal{P}(f, \rho)$ and $\pi \in \Pi$ the (population) DR Bellman operator is defined as

$$\mathcal{T}^\pi(v)(s) := \inf_{P \in \mathcal{P}} \left(\sum_{a \in \mathcal{A}} \pi(a|s) \Bigg[\sum_{s' \in \mathcal{S}} P_{s,a}(s') \left(R(s,a,s') + \gamma v(s') \right) \Bigg] \right) \tag{3.5}$$

for all $s \in \mathcal{S}$. The optimal DR Bellman operator is $\mathcal{T}^*(v)(s) := \max_{\pi \in \Pi} \mathcal{T}^\pi(v)(s)$ where, for each fixed $s \in \mathcal{S}$, the quantity $\mathcal{T}^*(v)(s)$ depends only on the state-wise action distribution $\pi(\cdot \mid s)$, and the maximum is attained by some $\pi^*_v(\cdot \mid s) \in \Delta(\mathcal{A})$.

3.3 Generative Model and the Empirical Bellman Estimator

The sample complexity analysis in this paper assumes the availability of a *generative model*, a.k.a. a simulator, which allows us to sample independently from the nominal controlled transition kernel $\overline{P}_{s,a}$, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. In particular, given sample size n, we sample i.i.d. $\{S^{(1)}_{s,a}, \cdots, S^{(n)}_{s,a}\}$ from $\overline{P}_{s,a}$ and construct the empirical transition probability

$$\overline{P}_{s,a,n}(s') := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left\{ S^{(i)}_{s,a} = s' \right\}. \tag{3.6}$$

Then, we define $\overline{P}_n := \{\overline{P}_{s,a,n} | (s,a) \in \mathcal{S} \times \mathcal{A}\}$ as the empirical nominal controlled transition kernel based on n samples. We define the empirical uncertainty set $\mathcal{P}_n(f, \rho) := \bigtimes_{s \in \mathcal{S}} \mathcal{P}_{s,n}(f, \rho)$ where $\mathcal{P}_{s,n}(f, \rho)$ is from (3.2) by replacing $\overline{P}_{s,a}$ with $\overline{P}_{s,a,n}$. Again, the dependence on (f, ρ) will be suppressed for simplicity.

The empirical value function $V^\pi_{\hat{\mathcal{P}}}$ is defined in (3.3) with $\mathcal{P}$ replaced by $\mathcal{P}_n$. The empirical DR Bellman operator $\hat{\mathbf{T}}^\pi$ is defined as in (3.5) with $\mathcal{P}$ replaced by $\mathcal{P}_n$. The corresponding optimal empirical DR Bellman operator is $\hat{\mathbf{T}}^*(v)(s) := \max_{\pi \in \Pi} \hat{\mathbf{T}}^\pi(v)(s)$, defined for each $v : S \to \mathbb{R}$.

Equipped with these definitions, we present our strategy to estimate the optimal value and policy of the S-rectangular robust MDP via the empirical value function. This is motivated by the fact that $V^*_{\mathcal{P}} = v^*$ where v^* solves (3.4).

Definition 4 (Empirical Bellman Estimators)**.** We define the empirical Bellman estimator $\hat{v}$ to $V^*_{\mathcal{P}}$ as the unique fixed point of $\hat{\mathbf{T}}^*$; i.e. $\hat{\mathbf{T}}^*(\hat{v}) = \hat{v}$. Moreover, we define an empirical optimal policy to be any $\hat{\pi}^* \in \Pi$ that attains the maximum in $\hat{\mathbf{T}}^*(\hat{v})$, i.e., $\hat{\mathbf{T}}^{\hat{\pi}^*}(\hat{v})(s) = \hat{\mathbf{T}}^*(\hat{v})(s)$ for all $s \in \mathcal{S}$.

The remainder of the paper is devoted to theoretical analysis and numerical validation of the statistical efficiency of estimating the optimal value $V^*_{\mathcal{P}} = v^*$ using $\hat{v}$, and of estimating the optimal policy using $\hat{\pi}^*$. We conclude this section by introducing the following important proposition that provides an upper bound on the l_∞ estimation error.

Proposition 1. *Let T, $\hat{T}$ be any γ-contraction operators, and u^*, $\hat{u}$ be the solution of $T(u) = u$ and $\hat{T}(u) = u$ respectively. Then, the estimation error is upper bounded by*

$$\|\hat{u} - u^*\|_\infty \leq \frac{1}{1-\gamma} \left\| \hat{T}(u^*) - T(u^*) \right\|_\infty$$

Corollary 1. *Let v^*,$\hat{v}$ be the solution of $\mathcal{T}^*(v) = v$ and $\hat{\mathbf{T}}^*(v) = v$, respectively. Then, the estimation error is upper bounded by*

$$\|\hat{v} - v^*\|_\infty \leq \frac{1}{1-\gamma} \left\| \hat{\mathbf{T}}^*(v^*) - \mathcal{T}^*(v^*) \right\|_\infty$$

Corollary 2. *Let $V^\pi_{\mathcal{P}}$, $V^\pi_{\hat{\mathcal{P}}}$ be the solution of $v = \mathcal{T}^\pi(v)$ and $v = \hat{\mathbf{T}}^\pi(v)$ respectively. Then, the estimation error is upper bounded by*

$$\sup_{\pi \in \Pi} \left\| V^\pi_{\hat{\mathcal{P}}} - V^\pi_{\mathcal{P}} \right\|_\infty \leq \frac{1}{1-\gamma} \sup_{\pi \in \Pi} \left\| \hat{\mathbf{T}}^\pi(V^\pi_{\mathcal{P}}) - \mathcal{T}^\pi(V^\pi_{\mathcal{P}}) \right\|_\infty$$

The proof of Proposition 1 is deferred to Appendix A.

4 Sample Complexity Bounds for the Empirical Bellman Estimator

In this section, we establish sample complexity upper bounds to achieve an absolute ε error in l_∞ distance when estimating $V^*_{\mathcal{P}}$ using $\hat{v}$. We focus on two specific f-divergence uncertainty models. When $f_{\mathrm{KL}}(t) = f(t) = t\log t$, the corresponding uncertainty set $\mathcal{P}_s(f_{\mathrm{KL}}, \rho)$ is based on the KullbackLeibler (KL) divergence, which is widely used in the machine learning literature. Alternatively, when $f = f_k$ as defined in Definition 6, the resulting f_k-divergence model captures another well-studied class of uncertainty sets (Duchi and Namkoong, 2021).

We note that our analysis techniques are applicable to a broader class of smooth divergence functions f. However, we focus on these two representative cases for demonstration purposes. This reflects that achieving near-tight sample complexity bounds often requires leveraging specific structural properties of the divergence. In particular, we highlight the desirable feature that, in the regime where the radius $\rho \downarrow 0$, our bounds remain $O(1)$ in ρ, avoiding the diverging sample complexity upper bounds established in earlier results, as discussed in (Wang et al., 2024c).

To facilitate our analysis and establish sample complexity results, we define the minimum support probability as a complexity metric parameter as follows.

Definition 5. Define the minimum support probability as

$$\mathfrak{p}_\wedge := \min_{s,a\in\mathcal{S}\times\mathcal{A}} \min_{s'\in\mathcal{S}:\overline{P}_{s,a}(s')>0} \overline{P}_{s,a}(s')$$

As noted in the literature, the use of $\mathfrak{p}_\wedge$ as a complexity metric is well justified. In the KL case, the convergence rate of the estimation error can degrade arbitrarily, depending on the specific MDP instance, if there is no lower bound on the minimum support probability. In particular, the rate can be as slow as $\Omega(n^{-1/\beta})$ for any $\beta \geq 2$ as the sample size n tends to infinity (Si et al., 2020). Similar negative results hold in the f_k-divergence setting when the parameter k approaches 1 (Duchi and Namkoong, 2020), highlighting the necessity of such a complexity measure.

4.1 The Kullback-Leibler Divergence Uncertainty Set

In this section, we present sample complexity results under the KL-divergence uncertainty set. Our analysis relies on the following dual representation of the DR Bellman operator and its empirical version. The strong duality underlying this representation follows from the ϕ-divergence duality result of Shapiro (2017, Sec. 3.2).

Lemma 1. *With $\mathcal{P} = \mathcal{P}(f_{\mathrm{KL}}, \rho)$ where $f_{\mathrm{KL}}(t) = t\log t$ and $\rho \in (0,\infty)$, for any $\pi \in \Pi$ and $s \in \mathcal{S}$, the dual form of the DR Bellman operator with KL uncertainty set $\mathcal{P}$ is*

$$\mathcal{T}^\pi(v)(s) = \sup_{\lambda\geq 0}\Bigg(-\lambda|\mathcal{A}|\rho - \sum_{a\in\mathcal{A}} \lambda \log \mathbb{E}_{\overline{P}_{s,a}}\left[\exp\left(-\frac{\pi(a|s)w(s,a,S)}{\lambda}\right)\right]\Bigg). \tag{4.1}$$

where $w(s,a,S) = R(s,a,S) + \gamma v(S)$. The KL empirical DR Bellman operator $\hat{\mathbf{T}}^\pi$ satisfies (4.1) with $\overline{P}_{s,a}$ replaced by $\overline{P}_{s,a,n}$.

The proof of Lemma 1 is provided in Appendix B.1. Building on this dual formulation, we next analyze the statistical error between the empirical and population DR Bellman operators.

Proposition 2. *Under the KL-divergence uncertainty set with any $\rho \in (0,\infty)$, for any $v : \mathcal{S} \to \mathbb{R}$ and $n \geq 12\mathfrak{p}_\wedge^{-1}\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)$, with probability at least $1-\eta$,*

$$\begin{aligned}\|\hat{\mathbf{T}}^*(v) - \mathcal{T}^*(v)\|_\infty &\leq \sup_{\pi\in\Pi}\|\hat{\mathbf{T}}^\pi(v) - \mathcal{T}^\pi(v)\|_\infty \\ &\leq \frac{9\|R+\gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\wedge}}\sqrt{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}.\end{aligned}$$

The proof of Proposition 2 is provided in Appendix C.1. Then, combining Proposition 2 with Proposition 1, and the fact that $\|R+\gamma v^*\|_\infty \leq 1/(1-\gamma)$ under our assumption that $R \in [0,1]$, we arrive at the following theorem. The proof is presented in Appendix C.4.

Theorem 1. *Assume the adversary chooses from a KL-divergence uncertainty set with $\rho \in (0,\infty)$. If the sample size $n \geq 12\mathfrak{p}_\wedge^{-1}\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)$, then, with probability at least $1-\eta$,*

$$\|\hat{v} - v^*\|_\infty \leq \frac{9}{(1-\gamma)^2\sqrt{n\mathfrak{p}_\wedge}}\sqrt{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}.$$

Moreover, with probability at least $1-\eta$,

$$\sup_{\pi\in\Pi} V^\pi_{\mathcal{P}}(s) - V^{\hat{\pi}^*}_{\mathcal{P}}(s) \leq \frac{18}{(1-\gamma)^2\sqrt{n\mathfrak{p}_\wedge}}\sqrt{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}.$$

Remark 1. Therefore, under the KL-divergence, a total of $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-4}\mathfrak{p}_\wedge^{-1}\varepsilon^{-2})$ samples from the simulator suffices to obtain an ε-accurate estimate of v^* with $\hat{v}$, or an ε-optimal policy with high probability.

4.2 f_k-Divergence Uncertainty Set

Next, we consider a subclass of the Cressie-Read family of f_k-divergence with $k \in (1, \infty)$, as studied in Duchi and Namkoong (2021).

Definition 6. For $k \in (1, \infty)$, the f_k-divergence is defined by the divergence functions $f_k(t) := (t^k - kt + k - 1)/(k(k-1))$. We also define $k^* = k/(k-1)$.

Notably, when $k = 2$, the f_2-divergence is the χ^2-divergence, which sees extensive application in the statistical testing literature. Moreover, when $k \downarrow 1$, the f_k induced divergence converges to KL.

The analysis for f_k-divergence uncertainty sets follows the same strategy to KL-divergence in the previous subsection. Below we summarise the main results.

Lemma 2. *With $\mathcal{P} = \mathcal{P}(f_k, \rho)$ and $\rho \in (0, \infty)$, for any $\pi \in \Pi$ and $s \in \mathcal{S}$, the dual form of the DR Bellman operator with f_k uncertainty set $\mathcal{P}$ is*

$$\mathcal{T}^\pi(v)(s) = \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} \left[\sum_{a \in \mathcal{A}} \eta_a - c \left(\sum_{a \in \mathcal{A}} \mathbb{E}_{\overline{P}_{s,a}} \left[(\eta_a - \pi(a|s) w(s,a,S))_+^{k^*} \right] \right)^{1/k^*} \right]$$

where $c = c(k, \rho, |\mathcal{A}|) = |\mathcal{A}|^{1/k} \left(k(k-1)\rho + 1\right)^{1/k}$, $(\cdot)_+ = \max(\cdot, 0)$ and $w(s,a,S) = R(s,a,S) + \gamma v(S)$. The f_k empirical DR Bellman operator $\hat{\mathbf{T}}^\pi$ satisfies a similar equality with $\overline{P}_{s,a}$ replaced by $\overline{P}_{s,a,n}$.

The proof of Lemma 2 is provided in Appendix B.2. Again, with this dual representation of the DR Bellman operators and refined estimation error analysis, we arrive at the following result.

Proposition 3. *Under the f_k-divergence uncertainty set with any $\rho \in (0, \infty)$, for any $v : \mathcal{S} \to \mathbb{R}_+$ and $n \geq 12\mathfrak{p}_\wedge^{-1} \log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)$, w.p. at least $1 - \eta$,*

$$\begin{aligned} \|\hat{\mathbf{T}}^*(v) - \mathcal{T}^*(v)\|_\infty &\leq \sup_{\pi \in \Pi} \|\hat{\mathbf{T}}^\pi(v) - \mathcal{T}^\pi(v)\|_\infty \\ &\leq \frac{9\|R + \gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\wedge}} \sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}. \end{aligned}$$

The proof of Proposition 3 is provided in Appendix D.1. This, combined with Proposition 1, implies the following error bound, whose proof is deferred to Appendix D.5.

Theorem 2. *Assume the adversary chooses from a f_k-divergence uncertainty set with $\rho \in (0, \infty)$. If $n \geq 12\mathfrak{p}_\wedge^{-1} \log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)$, then, w.p. at least $1 - \eta$,*

$$\|\hat{v} - v^*\|_\infty \leq \frac{9}{(1-\gamma)^2 \sqrt{n\mathfrak{p}_\wedge}} \sqrt{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}.$$

Moreover, w.p. at least $1 - \eta$,

$$\sup_{\pi \in \Pi} V_{\mathcal{P}}^\pi(s) - V_{\mathcal{P}}^{\hat{\pi}^*}(s) \leq \frac{18}{(1-\gamma)^2 \sqrt{n\mathfrak{p}_\wedge}} \sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}.$$

Remark 2. Therefore, under the f_k-divergence, an ε-accurate estimate of v^* with $\hat{v}$, or an ε-optimal policy with high probability, can be obtained using a total of $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-4}\mathfrak{p}_\wedge^{-1}\varepsilon^{-2})$ samples.

5 Numerical Experiments

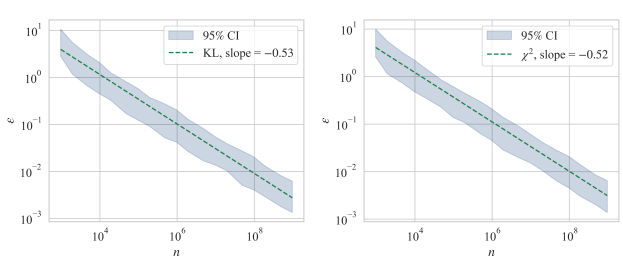
In this section, we present two sets of numerical examples. In Section 5.1, we revisit the robust inventory problem from Ho et al. (2018), which features uncertain demand, to demonstrate the $n^{-1/2}$ error decay rate. In Section 5.2, we consider an example from Yang et al. (2022) to illustrate the linear dependence on $|\mathcal{S}||\mathcal{A}|$, which matches the lower bound established in Yang et al. (2022).

5.1 Robust Inventory Control Problems

We investigate the dependency of the estimation error ε on the sample size n and evaluate our approach on a classical discrete-time inventory management problem with stochastic demand and backlog (Ho et al., 2018). In each period t, an agent decides the order quantity to maximize cumulative discounted rewards, accounting for holding costs, backlog penalties, and profits.

Let I denote the maximum inventory level, B the maximum backlog, and O the maximum order quantity per period. The state space is defined as $\mathcal{S} = \{-B, \cdots, 0, \cdots, I\}$, the action space is $\mathcal{A} = \{0, \cdots, O\}$, where $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ denote the inventory and order item at the beginning of period t. Demand $D_t \in \{0, \cdots, D_{\max}\}$ is an i.i.d. sequence, with distribution $P_D \in \Delta(D_{\max} + 1)$.

The MDP dynamics proceed as follows. Due to storage constraints, the effective order size is $\tilde{A}_t = \min(A_t, I -$



(a) Uncertainty sets based on KL-divergence (b) Uncertainty sets based on χ^2-divergence

Figure 1: Estimation error versus sample size n in the robust inventory control problem.

S_t). Then, the next state evolves as $S_{t+1} = \max(S_t + \tilde{A}_t - D_t, -B)$, ensuring that the backlog does not exceed B. The actual sales in period t are given by $X_t = S_t - S_{t+1} + \tilde{A}_t$. A one-step reward $R(S_t, A_t, S_{t+1}) = pX_t + b\min(S_{t+1}, 0) - h\max(S_{t+1}, 0) - c\tilde{A}_t$ is collected, where p is the sales price, c is the purchase cost, h is the holding cost, and b is the penalty of backlog.

For our experiments, we set the parameters as follows: $I = 10$, $B = 5$, $O = 5$, $p = 3$, $c = 2$, $h = 0.2$, $b = 3$, $\gamma = 0.9$, and use the nominal demand distribution $P_D = [0.1, 0.2, 0.3, 0.3, 0.1]$, supported on $\{0, 1, 2, 3, 4\}$. For each (s, a), we sample n_0 samples from the nominal transition kernels to generate the estimated transition probability P_n, and solve the DR-MDP problem with uncertainty size $\rho = 1/6$ using the algorithm presented in Ho et al. (2022).

Figure 1 illustrates the relationship between the sample size n and the error ε between the empirical and population value functions. As shown in the log-log plot, the slope is approximately -0.5 for both the KL and χ^2 cases, indicating that the error decreases at a rate proportional to $1/\sqrt{n}$.

5.2 MDP Instances from the Lower Bound Construction in Yang et al. (2022)

In this section, we investigate the relationship between the estimation error and the sizes of the state space $|\mathcal{S}|$ and action space $|\mathcal{A}|$. We adopt the classic MDP structure introduced in Gheshlaghi Azar et al. (2013) and Yang et al. (2022), which comprises three subsets: $\mathcal{S}$, $\mathcal{Y}_1$, and $\mathcal{Y}_2$, as illustrated in Figure 2.

Specifically, $\mathcal{S}$ denotes the set of all initial states, each associated with an action set $\mathcal{A}$. When an action $a_i \in \mathcal{A}$ is taken in state $s \in \mathcal{S}$, the system deterministically transitions (with probability 1) to the corresponding state $y_{1,s,a} \in \mathcal{Y}_1$. From each $y_{1,s,a}$, the system either remains in the same state with nominal probability p, or transitions to the corresponding absorbing state $y_{2,s,a} \in \mathcal{Y}_2$ with nominal probability $1-p$. All states in $\mathcal{Y}_2$ are absorbing, meaning that once the system enters one of these states, it remains there indefinitely via a self-loop with probability 1. The reward function is defined such that a reward of 1 is obtained only when the system is in any state within $\mathcal{Y}_1$; all other states yield a reward of 0. We solve the DR-MDP problem with $\gamma = 0.9$ and uncertainty size $\rho = 0.1$.

In our experiments, we first fix $|\mathcal{A}| = 65$ and vary the number of states from 10 to 1000, with the results shown in Figure 3. We then fix $|\mathcal{S}| = 65$ and vary $|\mathcal{A}|$ over the same range, with the corresponding results presented in Figure 4.

To align with our theoretical results, we normalize the estimation error by dividing it by $\log(|\mathcal{S}||\mathcal{A}|)$. Figures 3 and 4 display the behavior of this normalized error as $|\mathcal{S}|$ and $|\mathcal{A}|$ vary, respectively. Specifically, for each (s, a) pair, we use n_0 samples, resulting in a total of $n_0|\mathcal{S}||\mathcal{A}|$ samples. The left subfigures correspond to the KL-divergence case, while the right subfigures correspond to the χ^2-divergence case. We observe that as either $|\mathcal{S}|$ or $|\mathcal{A}|$ increases, the normalized error is non-increasing. This is consistent with our theoretical analysis, which predicts that the sample complexity scales linearly (up to logarithmic factors) with the product $|\mathcal{S}||\mathcal{A}|$.

6 Conclusion and Future Work

In this paper, we establish near-optimal sample complexity results for divergence-based S-rectangular robust MDPs under the discounted reward criterion. To the best of our knowledge, our results are the first to achieve optimal dependence on $|\mathcal{S}|$, $|\mathcal{A}|$, and ε simultaneously. Looking ahead, several promising research directions remain open. A natural extension is to move beyond the generative model assumption and develop provable guarantees in more practical settings, such as model-free reinforcement learning and offline reinforcement learning. Another important avenue is to explore richer formulations of uncertainty, for instance, by modeling uncertainty around the underlying noise process in distributionally robust stochastic control. Finally, it is also interesting to investigate S-rectangular robust MDPs under the average-reward criterion.

Acknowledgement

N. Sis research is supported in part by the Early Career Scheme [Grant 26210125] from the Hong Kong Research Grants Council and HKUST Li & Fung Supply Chain Institute Research Grant 2025.

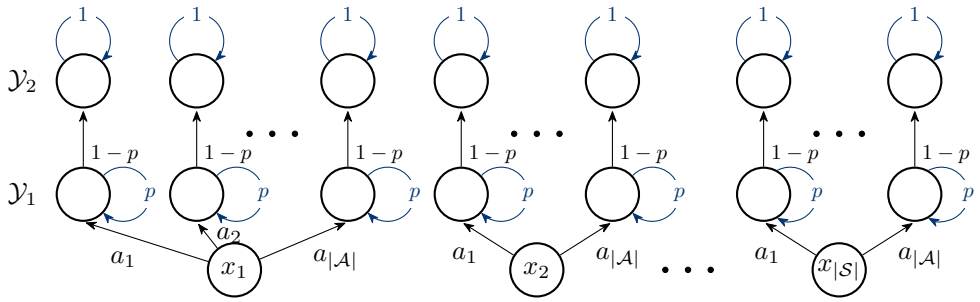
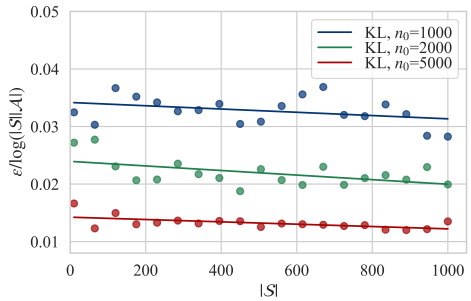
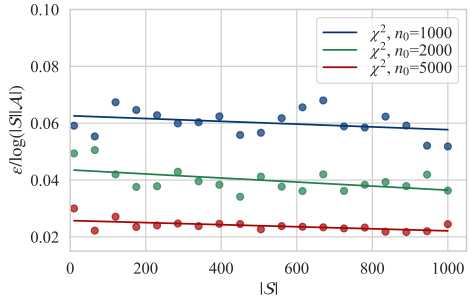


Figure 2: MDP instances from the lower bound construction in Yang et al. (2022).

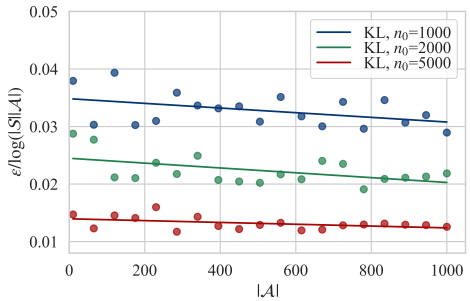


(a) Uncertainty sets based on KL-divergence

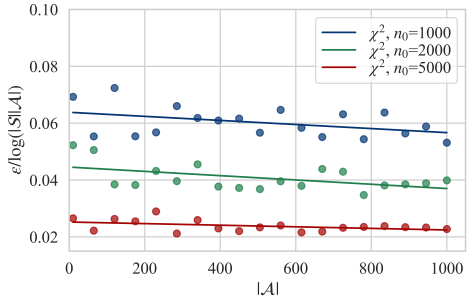


(b) Uncertainty sets based on χ^2-divergence

Figure 3: Estimation error versus the number of states $|\mathcal{S}|$ for the MDP instances based on the lower bound construction in Yang et al. (2022).



(a) Uncertainty sets based on KL-divergence



(b) Uncertainty sets based on χ^2-divergence

Figure 4: Estimation error versus the number of actions $|\mathcal{A}|$ for the MDP instances based on the lower bound construction in Yang et al. (2022).

References

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36, 2024.

Pierre Clavier, Laixi Shi, Erwan Le Pennec, Eric Mazumdar, Adam Wierman, and Matthieu Geist. Near-optimal distributionally robust reinforcement learning with general l_p norms. *Advances in Neural Information Processing Systems*, 37:1750–1810, 2024.

John Duchi and Hongseok Namkoong. Learning Models with Uniform Performance via Distributionally Robust Optimization, July 2020. URL `http://arxiv.org/abs/1810.08750`. arXiv:1810.08750 [stat].

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349, 2013.

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.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Fast bellman updates for robust mdps. In *International Conference on Machine Learning*, pages 1979–1988. PMLR, 2018.

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

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

Navdeep Kumar, Kaixin Wang, Kfir Yehuda Levy, and Shie Mannor. Efficient value iteration for s-rectangular robust markov decision processes. In *Forty-first International Conference on Machine Learning*, 2024.

Yann Le Tallec. *Robust, risk-sensitive, and data-driven control of Markov decision processes*. PhD thesis, Massachusetts Institute of Technology, 2007.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust Q-learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13623–13643. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/liu22a.html`.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

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

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model, 2021. URL `https://arxiv.org/abs/2112.01506`.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Number v.414 in Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, Hoboken, 2009. ISBN 978-0-471-72782-8.

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

Alexander Shapiro. Distributionally robust modeling of optimal control. *Operations Research Letters*, 50 (5):561–567, 2022.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity, 2022. URL `https://arxiv.org/abs/2208.05767`.

Laixi Shi and Yuejie Chi. Distributionally Robust Model-Based Offline Reinforcement Learning with Near-Optimal Sample Complexity, December 2023. URL `http://arxiv.org/abs/2208.05767`. arXiv:2208.05767 [cs].

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distri-

butional robustness in reinforcement learning with a generative model, 2023.

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally Robust Policy Evaluation and Learning in Offline Contextual Bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8884–8894. PMLR, November 2020. URL https://proceedings.mlr.press/v119/si20a.html. ISSN: 2640-3498.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust Q-learning, 2023a. URL https://arxiv.org/abs/2302.13203.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust Q-learning, 2023b.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Statistical learning of distributionally robust stochastic control in continuous state spaces. *arXiv preprint arXiv:2406.11281*, 2024a.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. On the Foundation of Distributionally Robust Reinforcement Learning, January 2024b. URL http://arxiv.org/abs/2311.09018. arXiv:2311.09018 [cs].

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample Complexity of Variance-reduced Distributionally Robust Q-learning, September 2024c. URL http://arxiv.org/abs/2305.18420. arXiv:2305.18420 [cs].

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

Huan Xu and Shie Mannor. Distributionally robust markov decision processes. In *NIPS*, pages 2505–2513, 2010.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning, 2023.

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.

Wenhao Yang, Han Wang, Tadashi Kozuno, Scott M. Jordan, and Zhihua Zhang. Avoiding model estimation in robust markov decision processes with a generative model, 2023.

Zhengqing Zhou, Qinxun Bai, Zhengyuan Zhou, 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, 2021a.

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 Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3331–3339. PMLR, 13–15 Apr 2021b. URL https://proceedings.mlr.press/v130/zhou21d.html.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

   We provide clear and complete definitions in Section 3 and Section 4.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

   All assumptions are stated explicitly in the main text, and detailed proofs of every theorem are provided in the Appendix for completeness.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

We provide the full source code and reproduction instructions in the supplemental material, and report instance parameters in Section 5. Experiments were run on a SLURM-managed cluster for efficiency, but since training can also runs on standard CPUs, we did not emphasize the hardware setup.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

   We re-implemented existing algorithms from the literature and cited their original sources in the main paper and supplemental material. Our implementation code is released in the supplemental material.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

   Our work does not involve human subjects or crowdsourced data.

Supplementary Materials

A Proofs of Value Function Error Bounds

In this section, we prove Proposition 1, Corollary 1 and Corollary 2. We first show that both the population and the empirical S-rectangular Bellman operators $\mathcal{T}^*$ and $\hat{\mathbf{T}}^*$ are γ-contractions. Furthermore, for given π, the Bellman operators $\mathcal{T}^\pi$ and $\hat{\mathbf{T}}^\pi$ are also γ-contractions. This is a well-known fact, see for example (Wang et al., 2024b). We include a proof to make the paper self-contained.

Lemma 3. *Let*

$$f_{\pi,P}(v)(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \left[\sum_{s'\in\mathcal{S}} P_{s,a}(s') \left(R(s,a,s') + \gamma v(s') \right) \right]$$

Then,

$$\sup_{\pi\in\Pi} \sup_{P\in\mathcal{P}} |f_{\pi,P}(v_1)(s) - f_{\pi,P}(v_2)(s)| \le \gamma \|v_1 - v_2\|_\infty$$

Proof. Fix s. The reward terms cancel in the difference, hence for any (π, P),

$$\begin{aligned} |f_{\pi,P}(v_1)(s) - f_{\pi,P}(v_2)(s)| &= \gamma \Big| \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} P_{s,a}(s') \big(v_1(s') - v_2(s') \big) \Big| \\ &\le \gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} P_{s,a}(s') \, \|v_1 - v_2\|_\infty \\ &= \gamma \, \|v_1 - v_2\|_\infty, \end{aligned}$$

where we used nonnegativity and normalization of $\pi(\cdot|s)$ and $P_{s,a}(\cdot)$ so the weights sum to 1. Taking $\sup_{\pi\in\Pi} \sup_{P\in\mathcal{P}}$ does not increase the right-hand side, yielding the claim. □

Lemma 4. *$\mathcal{T}^*$ and $\hat{\mathbf{T}}^*$ are γ-contraction operators on $(\mathcal{S} \to \mathbb{R}, \|\cdot\|_\infty)$; i.e. for all $v_1, v_2 : \mathcal{S} \to \mathbb{R}$,*

$$\begin{aligned} \|\mathcal{T}^*(v_1) - \mathcal{T}^*(v_2)\|_\infty &\le \gamma \|v_1 - v_2\|_\infty, \\ \|\hat{\mathbf{T}}^*(v_1) - \hat{\mathbf{T}}^*(v_2)\|_\infty &\le \gamma \|v_1 - v_2\|_\infty. \end{aligned}$$

Proof. By definition, we have

$$\begin{aligned} |\mathcal{T}^*(v_1)(s) - \mathcal{T}^*(v_2)(s)| &= \left| \sup_{\pi\in\Pi} \mathcal{T}^\pi(v_1)(s) - \sup_{\pi\in\Pi} \mathcal{T}^\pi(v_2)(s) \right| \\ &= \left| \sup_{\pi\in\Pi} \inf_{P\in\mathcal{P}} f_{\pi,P}(v_1)(s) - \sup_{\pi\in\Pi} \inf_{P\in\mathcal{P}} f_{\pi,P}(v_2)(s) \right|. \end{aligned}$$

Since $|\sup_X f - \sup_X g| \le \sup_X |f - g|$ and $|\inf_X f - \inf_X g| \le \sup_X |f - g|$, we have

$$|\mathcal{T}^*(v_1)(s) - \mathcal{T}^*(v_2)(s)| \le \sup_{\pi\in\Pi} \sup_{P\in\mathcal{P}} |f_{\pi,P}(v_1)(s) - f_{\pi,P}(v_2)(s)| \le \gamma \|v_1 - v_2\|_\infty,$$

where the last inequality follows from Lemma 3. The above inequality holds for any $s \in \mathcal{S}$, which lead to

$$\|\mathcal{T}^*(v_1) - \mathcal{T}^*(v_2)\|_\infty \le \gamma \|v_1 - v_2\|_\infty.$$

We replace P_s with $P_{s,n}$, therefore

$$|\hat{\mathbf{T}}^*(v_1)(s) - \hat{\mathbf{T}}^*(v_2)(s)| \le \sup_{\pi\in\Pi} \sup_{P\in\mathcal{P}_n} |f_{\pi,P}(v_1)(s) - f_{\pi,P}(v_2)(s)| \le \gamma \|v_1 - v_2\|_\infty.$$

which lead to

$$\|\hat{\mathbf{T}}^*(v_1) - \hat{\mathbf{T}}^*(v_2)\|_\infty \le \gamma \|v_1 - v_2\|_\infty.$$

□

Lemma 5. *$\mathcal{T}^\pi$ and $\hat{\mathbf{T}}^\pi$ are γ-contraction operators on $(\mathcal{S} \to \mathbb{R}, \|\cdot\|_\infty)$; i.e. for all $v_1, v_2 : \mathcal{S} \to \mathbb{R}$,*

$$\|\mathcal{T}^\pi(v_1) - \mathcal{T}^\pi(v_2)\|_\infty \le \gamma \|v_1 - v_2\|_\infty,$$
$$\|\hat{\mathbf{T}}^\pi(v_1) - \hat{\mathbf{T}}^\pi(v_2)\|_\infty \le \gamma \|v_1 - v_2\|_\infty.$$

Proof. By definition, we have

$$|\mathcal{T}^\pi(v_1)(s) - \mathcal{T}^\pi(v_2)(s)| = \left| \inf_{P \in \mathcal{P}} f_{\pi,P}(v_1)(s) - \inf_{P \in \mathcal{P}} f_{\pi,P}(v_2)(s) \right|$$

Since $|\inf_X f - \inf_X g| \le \sup_X |f - g|$, we have

$$|\mathcal{T}^\pi(v_1)(s) - \mathcal{T}^\pi(v_2)(s)| \le \sup_{P \in \mathcal{P}} |f_{\pi,P}(v_1)(s) - f_{\pi,P}(v_2)(s)| \le \sup_{\tilde{\pi} \in \Pi} \sup_{P \in \mathcal{P}} |f_{\tilde{\pi},P}(v_1)(s) - f_{\tilde{\pi},P}(v_2)(s)| \le \gamma \|v_1 - v_2\|_\infty$$

where the last inequality follows from Lemma 3. The above inequality holds for any $s \in \mathcal{S}$, which lead to

$$\|\mathcal{T}^\pi(v_1) - \mathcal{T}^\pi(v_2)\|_\infty \le \gamma \|v_1 - v_2\|_\infty.$$

We replace P_s with $P_{s,n}$, therefore,

$$|\hat{\mathbf{T}}^\pi(v_1)(s) - \hat{\mathbf{T}}^\pi(v_2)(s)| \le \sup_{\pi \in \Pi} \sup_{P \in \mathcal{P}_n} |f_{\pi,P}(v_1)(s) - f_{\pi,P}(v_2)(s)| \le \gamma \|v_1 - v_2\|_\infty.$$

which lead to

$$\|\hat{\mathbf{T}}^\pi(v_1) - \hat{\mathbf{T}}^\pi(v_2)\|_\infty \le \gamma \|v_1 - v_2\|_\infty.$$

□

A.1 Proof of Proposition 1

Proof. The proof of Proposition 1 follows a similar argument to that used for the continuous-case operator in Wang et al. (2024a).

Let $u_0 \equiv 0$ and $u_{k+1} = \hat{T}(u_k)$. $\hat{u}$ and u^* is defined as the fixed point of $\hat{T}$ and T respectively.

$$\begin{aligned}
\Delta_{k+1} &= u_{k+1} - u^* \\
&= \hat{T}(u_k) - \hat{T}(u^*) + \hat{T}(u^*) - T(u^*) \\
&= \left[\hat{T}(u^* + \Delta_k) - \hat{T}(u^*)\right] + \left[\hat{T}(u^*) - T(u^*)\right] \\
&:= \mathbf{H}(\Delta_k) + V
\end{aligned}$$

Since $\hat{T}$ is a γ-contraction operator, we have

$$\|\mathbf{H}(\Delta_1) - \mathbf{H}(\Delta_2)\|_\infty = \left\|\hat{T}(u^* + \Delta_1) - \hat{T}(u^* + \Delta_2)\right\|_\infty \le \gamma \|\Delta_1 - \Delta_2\|_\infty,$$

therefore, $\mathbf{H}$ is also a γ-contraction operator. Then we show

$$\|\Delta_k\|_\infty \le \gamma^k \|u^*\|_\infty + \sum_{j=0}^{k-1} \gamma^j \|V\|_\infty$$

by induction: for $k = 1$,

$$\begin{aligned}
\|\Delta_1\|_\infty &\le \|\mathbf{H}(\Delta_0)\|_\infty + \|V\|_\infty \\
&= \|\mathbf{H}(\Delta_0) - \mathbf{H}(0)\|_\infty + \|V\|_\infty \\
&\le \gamma \|u^*\|_\infty + \|V\|_\infty .
\end{aligned}$$

For any k, we have

$$\begin{aligned}\|\Delta_{k+1}\|_\infty &\le \|\mathbf{H}(\Delta_k)\|_\infty + \|V\|_\infty \\ &= \|\mathbf{H}(\Delta_k) - \mathbf{H}(0)\|_\infty + \|V\|_\infty \\ &\le \gamma \|\Delta_k\|_\infty + \|V\|_\infty \\ &\le \gamma \left(\gamma^k \|u^*\|_\infty + \sum_{j=0}^{k-1} \gamma^j \|V\|_\infty\right) + \|V\|_\infty \\ &= \gamma^{k+1}\|u^*\|_\infty + \sum_{j=0}^{k} \gamma^j \|V\|_\infty.\end{aligned}$$

Therefore,

$$\|\hat{u} - u^*\|_\infty = \lim_{k\to\infty} \|\Delta_k\|_\infty \le \sum_{j=0}^{\infty} \gamma^j \|V\|_\infty = \frac{1}{1-\gamma} \left\|\hat{T}(u^*) - T(u^*)\right\|_\infty.$$

□

A.2 Proof of Corollary 1

Proof. By Lemma 4, both $\mathcal{T}^*$ and $\hat{\mathbf{T}}^*$ are γ-contractions. Applying Proposition 1 with $T = \mathcal{T}^*$, $\hat{T} = \hat{\mathbf{T}}^*$ and $u^* = v^*$ yields

$$\|\hat{v} - v^*\|_\infty \;\le\; \frac{1}{1-\gamma} \|\hat{\mathbf{T}}^*(v^*) - \mathcal{T}^*(v^*)\|_\infty,$$

which proves the statement. □

A.3 Proof of Corollary 2

Proof. Fix any $\pi \in \Pi$. By Lemma 5, $\mathcal{T}^\pi$ and $\hat{\mathbf{T}}^\pi$ are γ-contractions. Applying Proposition 1 with $T = \mathcal{T}^\pi$, $\hat{T} = \hat{\mathbf{T}}^\pi$ and $u^* = V^\pi_{\mathcal{P}}$ gives

$$\|V^\pi_{\hat{\mathcal{P}}} - V^\pi_{\mathcal{P}}\|_\infty \;\le\; \frac{1}{1-\gamma} \left\|\hat{\mathbf{T}}^\pi(V^\pi_{\mathcal{P}}) - \mathcal{T}^\pi(V^\pi_{\mathcal{P}})\right\|_\infty.$$

Since the above inequality holds for any $\pi \in \Pi$, taking the supremum over $\pi \in \Pi$ on both sides yields

$$\sup_{\pi\in\Pi} \|V^\pi_{\hat{\mathcal{P}}} - V^\pi_{\mathcal{P}}\|_\infty \;\le\; \frac{1}{1-\gamma} \sup_{\pi\in\Pi} \left\|\hat{\mathbf{T}}^\pi(V^\pi_{\mathcal{P}}) - \mathcal{T}^\pi(V^\pi_{\mathcal{P}})\right\|_\infty,$$

which proves the claim. □

B Strong Duality for Divergence-Based S-Rectangular Bellman Operators

The proofs for all f-divergence-based uncertainty sets follow a unified framework. We first present Lemma 6, which gives a general dual formulation for any convex f-divergence. For the KL-divergence and the f_k-divergence, we specialise this result by substituting the corresponding conjugate functions f^*. The detailed derivations for the KL-divergence and the f_k-divergence are provided in Appendix B.1 and B.2, respectively.

Lemma 6. *For any f-divergence uncertainty set, where $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex function and $f(1) = 0$ and satisfies $f(0) = \lim_{t\downarrow 0} f(t)$, the convex optimization problem*

$$\inf_{P\in\mathcal{P}} \sum_{a\in\mathcal{A}} \pi(a|s)\mathbb{E}_{P_{s,a}}\left[R(s,a,S) + \gamma v(S)\right]$$

can be reformulated as:

$$\sup_{\lambda\ge 0, \boldsymbol{\eta}\in\mathbb{R}^{|\mathcal{A}|}} -\lambda \sum_{a\in\mathcal{A}} \mathbb{E}_{\overline{P}_{s,a}}\left[f^*\left(\frac{\eta_a - \pi(a|s)\left(R(s,a,S) + \gamma v(S)\right)}{\lambda}\right)\right] - \lambda|\mathcal{A}|\rho + \sum_{a\in\mathcal{A}} \eta_a$$

where $f^*(t) = -\inf_{s\ge 0}\left(f(s) - st\right)$.

Proof. We follow the proof of Lemma 8.5 in Yang et al. (2022), however, in our case, R is determined by the next state. We do a change of variables, let $L_{s,a}(s') = \frac{P_{s,a}(s')}{\overline{P}_{s,a}(s')}$. The original optimization problem can be reformulated as:

$$\begin{aligned}
\inf_{L_s \geq 0} \quad & \sum_{a\in\mathcal{A}} \pi(a|s)\mathbb{E}_{\overline{P}_{s,a}}\left[L_{s,a}\left(R(s,a,S)+\gamma v(S)\right)\right] \\
\text{s.t.} \quad & \sum_{a\in\mathcal{A}} \mathbb{E}_{\overline{P}_{s,a}}[f(L_{s,a})] \leq |\mathcal{A}|\rho \\
& \mathbb{E}_{\overline{P}_{s,a}}[L_{s,a}] = 1 \quad \text{for all } a \in \mathcal{A}
\end{aligned}$$

The Lagrange function of primal problem is

$$\begin{aligned}
\mathcal{L}(L,\lambda,\boldsymbol{\eta}) = & \sum_{a\in\mathcal{A}} \pi(a|s)\mathbb{E}_{\overline{P}_{s,a}}\left[L_{s,a}\left(R(s,a,S)+\gamma v(S)\right)\right] \\
& + \lambda\left(\sum_{a\in\mathcal{A}} \mathbb{E}_{\overline{P}_{s,a}}[f(L_{s,a})] - |\mathcal{A}|\rho\right) - \sum_{a\in\mathcal{A}} \eta_a\left(\mathbb{E}_{\overline{P}_{s,a}}[L_{s,a}] - 1\right)
\end{aligned}$$

Denoting $f^*(t) = -\inf_{s\geq 0}\left(f(s) - st\right)$,

$$\begin{aligned}
& \inf_{L_s\geq 0} \ \mathcal{L}(L,\lambda,\boldsymbol{\eta}) \\
& = \inf_{L_s\geq 0}\left(\sum_{a\in\mathcal{A}} \mathbb{E}_{\overline{P}_{s,a}}\left[\pi(a|s)L_{s,a}\left(R(s,a,S)+\gamma v(S)\right) + \lambda f(L_{s,a}) - \eta_a L_{s,a}\right]\right) - \lambda|\mathcal{A}|\rho + \sum_{a\in\mathcal{A}}\eta_a \\
& = \lambda\sum_{a\in\mathcal{A}} \inf_{L_{s,a}\geq 0} \mathbb{E}_{\overline{P}_{s,a}}\left[\frac{\pi(a|s)\left(R(s,a,S)+\gamma v(S)\right) - \eta_a}{\lambda} L_{s,a} + f(L_{s,a})\right] - \lambda|\mathcal{A}|\rho + \sum_{a\in\mathcal{A}}\eta_a \\
& = -\lambda\sum_{a\in\mathcal{A}} \mathbb{E}_{\overline{P}_{s,a}}\left[f^*\left(\frac{\eta_a - \pi(a|s)\left(R(s,a,S)+\gamma v(S)\right)}{\lambda}\right)\right] - \lambda|\mathcal{A}|\rho + \sum_{a\in\mathcal{A}}\eta_a
\end{aligned}$$

□

B.1 Proof of Lemma 1

Proof. Recall that for the KL-divergence, $f(t) = t\log t$, whose conjugate function $f^*(s) = e^{s-1}$. Substituting f^* into Lemma 6, we obtain the following dual form:

$$\begin{aligned}
& \sup_{\lambda\geq 0, \boldsymbol{\eta}\in\mathbb{R}^{|\mathcal{A}|}} -\lambda\sum_{a\in\mathcal{A}} \mathbb{E}_{\overline{P}_{s,a}}\left[f^*\left(\frac{\eta_a - \pi(a|s)\left(R(s,a,S)+\gamma v(S)\right)}{\lambda}\right)\right] - \lambda|\mathcal{A}|\rho + \sum_{a\in\mathcal{A}}\eta_a \\
& = \sup_{\lambda\geq 0, \boldsymbol{\eta}\in\mathbb{R}^{|\mathcal{A}|}} -\lambda\sum_{a\in\mathcal{A}} \exp\left(\frac{\eta_a - \lambda}{\lambda}\right)\mathbb{E}_{\overline{P}_{s,a}}\left[\exp\left(\frac{-\pi(a|s)\left(R(s,a,S)+\gamma v(S)\right)}{\lambda}\right)\right] \\
& \quad - \lambda|\mathcal{A}|\rho + \sum_{a\in\mathcal{A}}\eta_a.
\end{aligned} \tag{B.1}$$

We first note that for each action a, the term $\lambda\mathbb{E}_{\overline{P}_{s,a}}[\cdot]$ is a positive constant with respect to η_a, while the term $-\lambda\mathbb{E}_{\overline{P}_{s,a}}[\cdot]\exp((\eta_a - \lambda)/\lambda)$ is concave in η_a, since for any $c > 0$, the function $-c\exp(x)$ is concave. Moreover, the term $\sum_a \eta_a$ is affine, and hence concave. As the sum of concave functions is concave, we conclude that (B.1) is concave in $\boldsymbol{\eta}$. Next, we optimize with respect to $\boldsymbol{\eta}$ by setting the gradient with respect to each η_a to zero:

$$-\exp\left(\frac{\eta_a - \lambda}{\lambda}\right)\mathbb{E}_{\overline{P}_{s,a}}\left[\exp\left(\frac{-\pi(a|s)\left(R(s,a,S)+\gamma v(S)\right)}{\lambda}\right)\right] + 1 = 0$$

Solving for η_a, we obtain

$$\eta_a = \lambda - \lambda\log\mathbb{E}_{\overline{P}_{s,a}}\left[\exp\left(\frac{-\pi(a|s)\left(R(s,a,S)+\gamma v(S)\right)}{\lambda}\right)\right]. \tag{B.2}$$

Substituting (B.2) into (B.1), we obtain

$$\sup_{\lambda\geq 0,\boldsymbol{\eta}\in\mathbb{R}^{|\mathcal{A}|}} -\lambda\sum_{a\in\mathcal{A}}\log\mathbb{E}_{\overline{P}_{s,a}}\left[\exp\left(\frac{-\pi(a|s)\left(R(s,a,S)+\gamma v(S)\right)}{\lambda}\right)\right]-\lambda|\mathcal{A}|\rho.$$

□

B.2 Proof of Lemma 2

Proof. We first introduce the conjugate function of f_k, which will be instrumental for deriving the dual representation of DR Bellman operator.

Lemma 7 (Duchi and Namkoong (2021), Section 2)**.** *Recall that in f_k-divergence,*

$$f_k(t) := \frac{t^k - kt + k - 1}{k(k-1)}$$

The conjugate function $f_k^*(s) = \sup_{t\geq 0}\left(st - f_k(t)\right)$ is given by

$$f_k^*(s) := \frac{1}{k}\left[((k-1)s+1)_+^{k^*} - 1\right]$$

where $(x)_+ = \max(x,0)$.

Substituting f_k^* into Lemma 6, and let $w_{s,a}(S) := \pi(a|s)\left(R(s,a,S)+\gamma v(S)\right)$, we obtain

$$\begin{aligned}
&\sup_{\lambda\geq 0,\eta\in\mathbb{R}^{|\mathcal{A}|}} -\sum_{a\in\mathcal{A}}\lambda\mathbb{E}_{\overline{P}_{s,a}}\left[f^*\left(\frac{\eta_a - w_{s,a}(S)}{\lambda}\right)\right]-\lambda|\mathcal{A}|\rho+\sum_{a\in\mathcal{A}}\eta_a\\
=&\sup_{\lambda\geq 0,\eta\in\mathbb{R}^{|\mathcal{A}|}} -\sum_{a\in\mathcal{A}}\lambda\mathbb{E}_{\overline{P}_{s,a}}\left[\frac{1}{k}\left[\left((k-1)\frac{\eta_a - w_{s,a}(S)}{\lambda}+1\right)_+^{k^*}-1\right]\right]
\end{aligned}$$

Since $k-1>0$ and $\lambda>0$ are constants with respect to the random variable S, we can factor them out of the expectation and the positive-part operator $(\cdot)_+$.

$$=\sup_{\lambda\geq 0,\eta\in\mathbb{R}^{|\mathcal{A}|}} -\frac{(k-1)^{k^*}}{k\lambda^{k^*-1}}\sum_{a\in\mathcal{A}}\mathbb{E}_{\overline{P}_{s,a}}\left[\left(\eta_a - w_{s,a}(S)+\frac{\lambda}{k-1}\right)_+^{k^*}\right]-\lambda|\mathcal{A}|\left(\rho-\frac{1}{k}\right)+\sum_{a\in\mathcal{A}}\eta_a$$

Finally, we perform the change of variables, let $\tilde{\eta}_a = \eta_a + \frac{\lambda}{k-1}$, we obtain

$$=\sup_{\lambda\geq 0,\tilde{\eta}\in\mathbb{R}^{|\mathcal{A}|}} -\frac{(k-1)^{k^*}}{k\lambda^{k^*-1}}\sum_{a\in\mathcal{A}}\mathbb{E}_{\overline{P}_{s,a}}\left[(\tilde{\eta}_a - w_{s,a}(S))_+^{k^*}\right]-\lambda|\mathcal{A}|\left(\rho+\frac{1}{k(k-1)}\right)+\sum_{a\in\mathcal{A}}\tilde{\eta}_a \tag{B.3}$$

Since $-\lambda^{-\alpha}$ is concave in λ for any $\alpha>0$, and $\lambda|\mathcal{A}|\left(\rho+\frac{1}{k(k-1)}\right)$ is an affine function of λ, it follows that (B.3) is concave with respect to λ. To optimize over λ, we take the derivative with respect to λ and set it to zero, which yields:

$$\frac{(k-1)^{k^*}}{k(k-1)\lambda^{k^*}}\sum_{a\in\mathcal{A}}\mathbb{E}_{\overline{P}_{s,a}}\left[(\tilde{\eta}_a - w_{s,a}(S))_+^{k^*}\right]-|\mathcal{A}|\left(\rho+\frac{1}{k(k-1)}\right)=0$$

Multiply $k(k-1)$ on both side of the equation, we have

$$\frac{(k-1)^{k^*}}{\lambda^{k^*}}\sum_{a\in\mathcal{A}}\mathbb{E}_{\overline{P}_{s,a}}\left[(\tilde{\eta}_a - w_{s,a}(S))_+^{k^*}\right]-|\mathcal{A}|\left(k(k-1)\rho+1\right)=0$$

Therefore, we obtain

$$\lambda^* = (k-1)|\mathcal{A}|^{-1/k^*}\left(k(k-1)\rho+1\right)^{-1/k^*}\left(\sum_{a\in\mathcal{A}}\mathbb{E}_{\overline{P}_{s,a}}\left[(\tilde{\eta}_a - w_{s,a}(S))_+^{k^*}\right]\right)^{1/k^*}$$

By substituting λ^* into the equation (B.3) , we have

$$\begin{aligned}
&\sup_{\lambda\geq 0,\eta\in\mathbb{R}^{|\mathcal{A}|}} -\sum_{a\in\mathcal{A}}\lambda\mathbb{E}_{\overline{P}_{s,a}}\left[f^*\left(\frac{\eta_a - w_{s,a}(S)}{\lambda}\right)\right] - \lambda|\mathcal{A}|\rho + \sum_{a\in\mathcal{A}}\tilde{\eta}_a \\
&= \sup_{\tilde{\eta}\in\mathbb{R}^{|\mathcal{A}|}} -\frac{k-1}{k}|\mathcal{A}|^{1/k}\left(k(k-1)\rho+1\right)^{1/k}\left(\sum_{a\in\mathcal{A}}\mathbb{E}_{\overline{P}_{s,a}}\left[(\tilde{\eta}_a - w_{s,a}(S))_+^{k^*}\right]\right)^{1/k^*} \\
&- \frac{1}{k}|\mathcal{A}|^{1/k}\left(k(k-1)\rho+1\right)^{1/k}\left(\sum_{a\in\mathcal{A}}\mathbb{E}_{\overline{P}_{s,a}}\left[(\tilde{\eta}_a - w_{s,a}(S))_+^{k^*}\right]\right)^{1/k^*} + \sum_{a\in\mathcal{A}}\tilde{\eta}_a \\
&= \sup_{\tilde{\eta}\in\mathbb{R}^{|\mathcal{A}|}} -|\mathcal{A}|^{1/k}\left(k(k-1)\rho+1\right)^{1/k}\left(\sum_{a\in\mathcal{A}}\mathbb{E}_{\overline{P}_{s,a}}\left[(\tilde{\eta}_a - w_{s,a}(S))_+^{k^*}\right]\right)^{1/k^*} + \sum_{a\in\mathcal{A}}\tilde{\eta}_a
\end{aligned}$$

□

C Proofs of Properties of the Empirical Bellman Operator: KL Case

Our techniques in this section refine that in Wang et al. (2024c). To follow the constructions in Wang et al. (2024c), we introduce some notations. Consider $\mu_{s,a} \in \Delta(\mathcal{S})$ and its empirical version $\mu_{s,a,n}$ constructed from n i.i.d samples from $\mu_{s,a}$. Define the collection of these measures under state s as $\boldsymbol{\mu}_s := \{\mu_{s,a} : a \in \mathcal{A}\}$. For a function $u : \mathcal{S} \to \mathbb{R}$ and for each $s \in \mathcal{S}$, we define:

$$\begin{aligned}
\|u\|_{\infty,\boldsymbol{\mu}_s} &= \max_{a\in\mathcal{A}}\|u\|_{L^\infty(\mu_{s,a})}, \\
\left\|\frac{dm_n}{d\mu_n(t)}\right\|_{\infty,\boldsymbol{\mu}_s} &= \max_{a\in\mathcal{A}}\left\|\frac{dm_{a,n}}{d\mu_{a,n}(t)}\right\|_{L^\infty(\mu_{s,a})}.
\end{aligned}$$

For the supremum over all states, we define

$$\|u\|_\infty = \sup_{s\in\mathcal{S}}\|u\|_{\infty,\boldsymbol{\mu}_s}.$$

We define a "good event" under which the empirical measure $\mu_{s,a,n}$ uniformly approximates the population measure $\mu_{s,a}$ with relative error bounded by δ_0 across all actions $a \in \mathcal{A}$. Formally, this event is given by

$$\Omega_{n,\delta_0}(\boldsymbol{\mu}_s) = \left\{\omega : \sup_{a\in\mathcal{A}}\sup_{s':\mu_{s,a}(s')>0}\left|\frac{\mu_{s,a,n}(\omega)(s') - \mu_{s,a}(s')}{\mu_{s,a}(s')}\right| \leq \delta_0\right\}.$$

Further, the good event over all states is defined as

$$\Omega_{n,\delta_0} = \bigcap_{s\in\mathcal{S}}\Omega_{n,\delta_0}(\boldsymbol{\mu}_s) = \left\{\omega : \sup_{s\in\mathcal{S}}\sup_{a\in\mathcal{A},s':\mu_{s,a}(s')>0}\left|\frac{\mu_{s,a,n}(\omega)(s') - \mu_{s,a}(s')}{\mu_{s,a}(s')}\right| \leq \delta_0\right\}.$$

For notation simplicity, we suppress the dependence on the state variable s. Consider a function $u : \mathcal{S} \to \mathbb{R}$. The dual function under KL-divergence is given by:

$$f(\boldsymbol{\mu}, u, \lambda) := -\lambda|\mathcal{A}|\rho - \sum_{a\in\mathcal{A}}\lambda\log\mu_a\left[e^{-d_a u/\lambda}\right], \tag{C.1}$$

where $\lambda > 0$ is the dual regularization parameter, and we denote $d_a := \pi(a|s)$ for simplicity.

We define the deviation between empirical and true measures as

$$m_{a,n} = \mu_{a,n} - \mu_a,$$

and their convex interpolation by

$$\mu_{a,n}(t) = t\mu_a + (1-t)\mu_{a,n}.$$

C.1 Proof of Proposition 2

Proof. Using (C.1) to express Bellman operator, we obtain

$$\begin{aligned}\sup_{\pi\in\Pi}\left|\hat{\mathbf{T}}^{\pi}(v)(s) - \mathcal{T}^{\pi}(v)(s)\right| &= \sup_{d\in\Delta(\mathcal{A})}\left|\sup_{\lambda>0} f(P_{s,n}, R(s,\cdot,\cdot)+\gamma v, \lambda) - \sup_{\lambda>0} f(P_s, R(s,\cdot,\cdot)+\gamma v, \lambda)\right| \\ &\le \sup_{\lambda>0, d\in\Delta(\mathcal{A})} \left|f(P_{s,n}, R(s,\cdot,\cdot)+\gamma v, \lambda) - f(P_s, R(s,\cdot,\cdot)+\gamma v, \lambda)\right|.\end{aligned} \tag{C.2}$$

We analyze the sensitivity of the mapping $\boldsymbol{\mu} \to f(\boldsymbol{\mu}, u, \lambda)$. For any fixed $u, \boldsymbol{\mu}$ and $\boldsymbol{\mu}_n$, define

$$g_n(t,\lambda) = f\left(\boldsymbol{\mu}_n(t), u, \lambda\right).$$

According to mean value theorem, there exists $\tau \in (0,1)$ satisfies:

$$\begin{aligned}|f(\boldsymbol{\mu}_n, u, \lambda) - f(\boldsymbol{\mu}, u, \lambda)| &= |g_n(0,\lambda) - g_n(1,\lambda)| \\ &= \left|\partial_t g_n(t,\lambda)\right|_{t=\tau}\Big| \\ &= \left|\sum_{a\in\mathcal{A}} \lambda \frac{m_{a,n}[e^{-d_a u/\lambda}]}{\mu_{a,n}(\tau)[e^{-d_a u/\lambda}]}\right|\end{aligned}$$

To bound the difference above, we invoke the following lemma.

Lemma 8. *For any fixed u and π, $\mu_n \ll \mu$, we have that*

$$\sup_{\lambda\ge 0}\left|\sum_{a\in\mathcal{A}} \lambda \frac{m_{a,n}[e^{-d_a u/\lambda}]}{\mu_{a,n}(t)[e^{-d_a u/\lambda}]}\right| \le 2\|u\|_\infty \left\|\frac{dm_n}{d\mu_n(t)}\right\|_{\infty,\boldsymbol{\mu}}.$$

The proof is deferred to Appendix C.2. According to lemma 8, we have

$$\sup_{\lambda\ge 0}|f(\boldsymbol{\mu}_n, u, \lambda) - f(\boldsymbol{\mu}, u, \lambda)| \le 2\|u\|_\infty \left\|\frac{dm_n}{d\mu_n(t)}\right\|_{\infty,\boldsymbol{\mu}}.$$

We decomposed the probability using the event $\Omega_{n,\delta_0}(\boldsymbol{\mu})$ where the empirical estimates are close to the population measures:

$$\begin{aligned}&P\left(\sup_{\lambda\ge 0, d\in\Delta(\mathcal{A})} |f(\mu_n, u, \lambda) - f(\mu, u, \lambda)| > t\right) \\ &\le P(\Omega_{n,\delta_0}(\boldsymbol{\mu})^c) + P\left(2\,\|u\|_\infty \left\|\frac{dm_n}{d\mu_n(\tau)}\right\|_{\infty,\boldsymbol{\mu}} > t, \Omega_{n,\delta_0}(\boldsymbol{\mu})\right)\end{aligned}$$

To control the denominator $\mu_{a,n}(\tau)(s')$ appearing in the bound, we use the following lemma, which asserts that under the good event, the empirical and population measures remain close for all $t \in [0,1]$:

Lemma 9. *For any s' with $\mu(s') > 0$, the measure $\mu_n(t)(s')$ satisfies*

$$(1-\delta_0)\mu(s') \le \mu_n(t)(s') \le (1+\delta_0)\mu(s'), \quad \forall t \in [0,1].$$

The proof is deferred to Appendix C.3. By using lemma 9, we have $\mu_{a,n}(\tau)(s') \geq (1-\delta_0)\mu_a(s')$, therefore,

$$\leq P\left(\sup_{a,s'}\left|\frac{\mu_{a,n}(s')-\mu_a(s')}{\mu_a(s')}\right| > \delta_0\right) + P\left(2\|u\|_\infty \sup_{a,s'}\left|\frac{\mu_{a,n}(s')-\mu_a(s')}{(1-\delta_0)\mu_a(s')}\right| > t\right).$$

By using the multiplicative Chernoff bound and Bernstein inequality, we have

$$\begin{aligned}
&\leq P\left(\sup_{a,s'}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}(S_i = s') - \mu_a(s')\right| > \delta_0\mu_a(s')\right) \\
&+ P\left(\frac{2}{1-\delta_0}\|u\|_\infty \sup_{a,s'}\frac{1}{\mu_a(s')}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}(S_i = s') - \mu_a(s')\right| > t\right) \\
&\leq 2\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\left(\exp\left(-\frac{\delta_0^2 n\mu_a(s')}{3}\right) + \exp\left(-\frac{t^2}{2}\left(\frac{4\|u\|_\infty^2}{(1-\delta_0)^2 n\mu_a(s')} + \frac{2\|u\|_\infty t}{3(1-\delta_0)n\mu_a(s')}\right)^{-1}\right)\right).
\end{aligned}$$

Since $\mu_a(s') \geq \mathfrak{p}_\wedge$, and both exponential term above is monotonically decreasing over $\mu_a(s')$, we have

$$\leq 2|\mathcal{A}||\mathcal{S}|\left(\exp\left(-\frac{\delta_0^2 n\mathfrak{p}_\wedge}{3}\right) + \exp\left(-\frac{t^2}{2}\left(\frac{4\|u\|_\infty^2}{(1-\delta_0)^2 n\mathfrak{p}_\wedge} + \frac{2\|u\|_\infty t}{3(1-\delta_0)n\mathfrak{p}_\wedge}\right)^{-1}\right)\right).$$

Recall from (C.2) that

$$P\left(\sup_{\pi\in\Pi}\left|\hat{\mathbf{T}}^\pi(v)(s) - \mathcal{T}^\pi(v)(s)\right| > t\right) \leq P\left(\sup_{\lambda>0, d_a\in\Delta(\mathcal{A})}|f(P_s, R(s,\cdot,\cdot)+\gamma v, \lambda) - f(P_s, R(s,\cdot,\cdot)+\gamma v, \lambda)|\right)$$

Replacing $\boldsymbol{\mu}$ with P_s and $\boldsymbol{\mu}_n$ with $P_{s,n}$ and choose $\delta_0 = \frac{1}{2}$, by union bound, we have

$$\begin{aligned}
&P\left(\sup_{\pi\in\Pi}\left\|\hat{\mathbf{T}}^\pi(v) - \mathcal{T}^\pi(v)\right\|_\infty > t\right) \\
&\leq P\left(\sup_{s}\sup_{\lambda\geq 0, d\in\Delta(\mathcal{A})}|f(P_{s,n}, R(s,\cdot,\cdot)+\gamma v, \lambda) - f(P_s, R(s,\cdot,\cdot)+\gamma v, \lambda)| > t\right) \\
&\leq 2|\mathcal{S}|^2|\mathcal{A}|\exp\left(-\frac{n\mathfrak{p}_\wedge}{12}\right) + 2|\mathcal{S}|^2|\mathcal{A}|\exp\left(-\frac{t^2}{2}\left(\frac{16\|R(s,\cdot,\cdot)+\gamma v\|_\infty^2}{n\mathfrak{p}_\wedge} + \frac{4\|R(s,\cdot,\cdot)+\gamma v\|_\infty t}{3\gamma n\mathfrak{p}_\wedge}\right)^{-1}\right).
\end{aligned}$$

Set each term to be less than $\eta/2$, we need

$$n \geq \frac{12}{\mathfrak{p}_\wedge}\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right) \tag{C.3}$$

$$t \geq \frac{8\|R+\gamma v\|_\infty}{3n\mathfrak{p}_\wedge}\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right) + \frac{4\|R+\gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\wedge}}\sqrt{2\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}. \tag{C.4}$$

Under (C.3), we have

$$\frac{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}{n\mathfrak{p}_\wedge} \leq \sqrt{\frac{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}{n\mathfrak{p}_\wedge}}.$$

By substituting this bound into (C.4), we have

$$\begin{aligned}
&\frac{8\|R+\gamma v\|_\infty}{3n\mathfrak{p}_\wedge}\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right) + \frac{4\|R+\gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\wedge}}\sqrt{2\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)} \\
&\leq \left(\frac{8}{3} + 4\sqrt{2}\right)\frac{\|R+\gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\wedge}}\sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)} \\
&\leq \frac{9\|R+\gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\wedge}}\sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}
\end{aligned}$$

Therefore, for when n specifies (C.3) and t satisfies

$$t \geq \frac{9\|R+\gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\wedge}}\sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)},$$

we have

$$P\left(\sup_{\pi\in\Pi}\left\|\hat{\mathbf{T}}^\pi(v)-\mathcal{T}^\pi(v)\right\|_\infty > t\right) \leq \eta.$$

Since $|\sup_X f - \sup_X g| \leq \sup_X |f-g|$,

$$\begin{aligned}
P\left(\sup_{\pi\in\Pi}\left\|\hat{\mathbf{T}}^*(v)-\mathcal{T}^*(v)\right\|_\infty > t\right) &= P\left(\left\|\sup_{\pi\in\Pi}\hat{\mathbf{T}}^\pi(v)-\sup_{\pi\in\Pi}\mathcal{T}^\pi(v)\right\|_\infty > t\right)\\
&\leq P\left(\sup_{\pi\in\Pi}\left\|\hat{\mathbf{T}}^\pi(v)-\mathcal{T}^\pi(v)\right\|_\infty > t\right)\\
&\leq \eta.
\end{aligned}$$

This implies Proposition 2. □

C.2 Proof of Lemma 8

Proof. Observe that multiplying the numerator and denominator by $e^{d_a\|u\|_{L^\infty(\mu_a)}/\lambda}$ preserves the value of the fraction. This is equivalent to rewriting the exponential terms as:

$$\left|\sum_{a\in\mathcal{A}}\lambda\frac{m_{a,n}[e^{-d_a u/\lambda}]}{\mu_{a,n}(t)[e^{-d_a u/\lambda}]}\right| = \left|\sum_{a\in\mathcal{A}}\lambda\frac{m_{a,n}[e^{d_a(\|u\|_{L^\infty(\mu_a)}-u)/\lambda}]}{\mu_{a,n}(t)[e^{d_a(\|u\|_{L^\infty(\mu_a)}-u)/\lambda}]}\right|.$$

Since $m_{a,n} = \mu_{a,n} - \mu_a$, for any constant c, we have $m_{a,n}[c] = 0$, which lead to

$$= \left|\sum_{a\in\mathcal{A}}\lambda\frac{m_{a,n}[e^{d_a(\|u\|_{L^\infty(\mu_a)}-u)/\lambda}-1]}{\mu_{a,n}(t)[e^{d_a(\|u\|_{L^\infty(\mu_a)}-u)/\lambda}]}\right|.$$

For any measure m, μ and random variable w_1, w_2, the following equation holds:

$$\begin{aligned}
\left|\frac{m[w_1]}{\mu[w_2]}\right| &= \left|\frac{\sum_s m(s)w_1(s)}{\sum_s \mu(s)w_2(s)}\right|\\
&= \left|\left(\sum_s \mu(s)\frac{m(s)}{\mu(s)}w_2(s)\frac{w_1(s)}{w_2(s)}\right)\left(\sum_s \mu(s)w_2(s)\right)^{-1}\right|\\
&\leq \left|\frac{\sum_s \mu(s)w_2(s)}{\sum_s \mu(s)w_2(s)}\right|\cdot\max_s\left|\frac{m(s)}{\mu(s)}\right|\cdot\max_s\left|\frac{w_1(s)}{w_2(s)}\right|\\
&= \left\|\frac{dm}{d\mu}\right\|_{L^\infty(\mu)}\left\|\frac{w_1}{w_2}\right\|_{L^\infty(\mu)}.
\end{aligned}$$

Applying this result and $|\sum\cdot| \leq \sum|\cdot|$, we obtain

$$\left|\sum_{a\in\mathcal{A}}\lambda\frac{m_{a,n}[e^{-d_a u/\lambda}]}{\mu_{a,n}(t)[e^{-d_a u/\lambda}]}\right| \leq \sum_{a\in\mathcal{A}}\left\|\lambda\frac{e^{d_a(\|u\|_{L^\infty(\mu_a)}-u)/\lambda}-1}{e^{d_a(\|u\|_{L^\infty(\mu_a)}-u)/\lambda}}\right\|_{L^\infty(\mu_a)}\left\|\frac{dm_{a,n}}{d\mu_{a,n}(t)}\right\|_{L^\infty(\mu_a)}.$$

Notice that when $x > 0$, we have $e^x - 1 \le xe^x$, then we obtain

$$\begin{aligned}
&\le \sum_{a\in\mathcal{A}} \left\| \lambda \frac{\frac{d_a(\|u\|_{L^\infty(\mu_a)}-u)}{\lambda} e^{d_a(\|u\|_{L^\infty(\mu_a)}-u)/\lambda}}{e^{d_a(\|u\|_{L^\infty(\mu_a)}-u)/\lambda}} \right\|_{L^\infty(\mu_a)} \left\| \frac{dm_{a,n}}{d\mu_{a,n}(t)} \right\|_{L^\infty(\mu_a)} \\
&\le \sum_{a\in\mathcal{A}} \left\| d_a(\|u\|_{L^\infty(\mu_a)} - u) \right\|_{L^\infty(\mu_a)} \left\| \frac{dm_{a,n}}{d\mu_{a,n}(t)} \right\|_{L^\infty(\mu_a)} \\
&\le \sum_{a\in\mathcal{A}} 2d_a\|u\|_\infty \left\| \frac{dm_{a,n}}{d\mu_{a,n}(t)} \right\|_{L^\infty(\mu_a)} \\
&\le 2\|u\|_\infty \left\| \frac{dm_n}{d\mu_n(t)} \right\|_{\infty,\boldsymbol{\mu}}
\end{aligned}$$

as claimed. □

C.3 Proof of Lemma 9

Proof. On the event Ω_{n,δ_0}, the empirical measure satisfies $\sup_{s'\in\mathcal{S}} \left| \frac{\mu_n(s')-\mu(s')}{\mu(s')} \right| \le \delta_0$. Hence, for any s' with $\mu(s') > 0$, we have:

$$(1-\delta_0)\mu(s') \le \mu_n(s') \le (1+\delta_0)\mu(s').$$

Substituting in the above bound on $\mu_n(s')$ into the definition of $\mu_n(t)(s')$ gives

$$(1-(1-t)\delta_0)\mu(s') \le t\mu(s') + (1-t)\mu_n(s') \le (1+(1-t)\delta_0)\mu(s').$$

For all $t \in [0,1]$, $(1-t) \le 1$, therefore, we have

$$(1-\delta_0)\mu(s') \le \mu_n(t)(s') \le (1+\delta_0)\mu(s').$$

□

C.4 Proof of Theorem 1

Proof. Substituting $\|R+\gamma v\|_\infty \le 1/(1-\gamma)$ into the bound from Proposition 2 and applying Proposition 1, we obtain the stated result.

$$\begin{aligned}
\|\hat{v} - v^*\|_\infty &\le \frac{1}{1-\gamma} \left\| \hat{\mathbf{T}}^*(v^*) - \mathcal{T}^*(v^*) \right\|_\infty \\
&\le \frac{9\|R+\gamma v\|_\infty}{(1-\gamma)\sqrt{n\mathfrak{p}_\wedge}} \sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)} \\
&\le \frac{9}{(1-\gamma)^2\sqrt{n\mathfrak{p}_\wedge}} \sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}
\end{aligned}$$

with probability $1-\eta$.

The preceding step provides a uniform bound on the value-function estimation error. We now turn to the policy suboptimality gap. We first introduce following lemma.

Lemma 10. *Let $\hat{\pi}^* \in \Pi$ satisfy*

$$\hat{\mathbf{T}}^{\hat{\pi}^*}(\hat{v})(s) = \hat{\mathbf{T}}^*(\hat{v})(s), \qquad \forall s \in \mathcal{S}.$$

Then for every $s \in \mathcal{S}$, we have

$$0 \le \sup_{\pi\in\Pi} V_{\mathcal{P}}^{\pi}(s) - V_{\mathcal{P}}^{\hat{\pi}^*}(s) \le 2\sup_{\pi\in\Pi} \left\| V_{\hat{\mathcal{P}}}^{\pi} - V_{\mathcal{P}}^{\pi} \right\|_\infty$$

According to Corollary 2, Lemma 10 and Proposition 2, we have

$$\begin{aligned}\sup_{\pi\in\Pi} V_{\mathcal{P}}^{\pi}(s) - V_{\mathcal{P}}^{\hat{\pi}^*}(s) &\le 2\sup_{\pi\in\Pi}\left\|V_{\hat{\mathcal{P}}}^{\pi} - V_{\mathcal{P}}^{\pi}\right\|_\infty \\ &\le \frac{2}{1-\gamma}\sup_{\pi\in\Pi}\left\|\hat{\mathbf{T}}^{\pi}(V_{\mathcal{P}}^{\pi}) - \mathcal{T}^{\pi}(V_{\mathcal{P}}^{\pi})\right\|_\infty \\ &\le \frac{18\|R+\gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\wedge}(1-\gamma)}\sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)} \\ &\le \frac{18}{\sqrt{n\mathfrak{p}_\wedge}(1-\gamma)^2}\sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}\end{aligned}$$

□

C.5 Proof of Lemma 10

Proof. We first show that $\hat{\pi}^*$ is optimal for the empirical DR-MDP, namely

$$V_{\hat{\mathcal{P}}}^{\hat{\pi}^*}(s) = \sup_{\pi\in\Pi} V_{\hat{\mathcal{P}}}^{\pi}(s), \qquad \forall s\in\mathcal{S}.$$

Since

$$\hat{\mathbf{T}}^{\hat{\pi}^*}(\hat{v})(s) = \hat{\mathbf{T}}^*(\hat{v})(s) = \hat{v}(s), \qquad \forall s\in\mathcal{S},$$

the vector $\hat{v}$ is a fixed point of $\hat{\mathbf{T}}^{\hat{\pi}^*}$. Because $\hat{\mathbf{T}}^{\hat{\pi}^*}$ is a γ-contraction, this fixed point is unique. By definition, that unique fixed point is $V_{\hat{\mathcal{P}}}^{\hat{\pi}^*}$. Hence

$$V_{\hat{\mathcal{P}}}^{\hat{\pi}^*} = \hat{v}.$$

Now fix any $\pi\in\Pi$. By definition of $\hat{\mathbf{T}}^*$, for every $u: S\to\mathbb{R}$ and every $s\in\mathcal{S}$,

$$\hat{\mathbf{T}}^{\pi}(u)(s) \le \hat{\mathbf{T}}^*(u)(s).$$

In particular,

$$\hat{\mathbf{T}}^{\pi}(\hat{v})(s) \le \hat{\mathbf{T}}^*(\hat{v})(s) = \hat{v}(s), \qquad \forall s\in\mathcal{S}.$$

Define the sequence $(u_k)_{k\ge 0}$ by

$$u_0 := \hat{v}, \qquad u_{k+1} := \hat{\mathbf{T}}^{\pi}(u_k), \quad k\ge 0.$$

Then $u_1(s)\le u_0(s)$ for all s. Since $\hat{\mathbf{T}}^{\pi}$ is monotone, whenever $u_k(s)\le u_{k-1}(s)$, we have

$$u_{k+1} = \hat{\mathbf{T}}^{\pi}(u_k)(s) \le \hat{\mathbf{T}}^{\pi}(u_{k-1})(s) = u_k(s), \qquad \forall s\in\mathcal{S}.$$

Thus, by induction,

$$u_k(s)\le\hat{v}(s), \qquad \forall s\in\mathcal{S}, \forall k\ge 1.$$

Since $\hat{\mathbf{T}}^{\pi}$ is a γ-contraction, (u_k) converges to the unique fixed point of $\hat{\mathbf{T}}^{\pi}$, namely $V_{\hat{\mathcal{P}}}^{\pi}$. Letting $k\to\infty$ yields

$$V_{\hat{\mathcal{P}}}^{\pi}(s) \le \hat{v}(s) = V_{\hat{\mathcal{P}}}^{\hat{\pi}^*}(s), \qquad \forall s\in\mathcal{S}.$$

Because $\pi\in\Pi$ was arbitrary,

$$V_{\hat{\mathcal{P}}}^{\hat{\pi}^*}(s) = \sup_{\pi\in\Pi} V_{\hat{\mathcal{P}}}^{\pi}(s), \qquad \forall s\in\mathcal{S}.$$

Finally, for any $s\in\mathcal{S}$,

$$\begin{aligned}0 &\le \sup_{\pi\in\Pi} V_{\mathcal{P}}^{\pi}(s) - V_{\mathcal{P}}^{\hat{\pi}^*}(s) \\ &= \sup_{\pi\in\Pi} V_{\mathcal{P}}^{\pi}(s) - \sup_{\pi\in\Pi} V_{\hat{\mathcal{P}}}^{\pi}(s) + V_{\hat{\mathcal{P}}}^{\hat{\pi}^*}(s) - V_{\mathcal{P}}^{\hat{\pi}^*}(s) \\ &\le \left\|\sup_{\pi\in\Pi} V_{\mathcal{P}}^{\pi} - \sup_{\pi\in\Pi} V_{\hat{\mathcal{P}}}^{\pi}\right\|_\infty + \left\|V_{\hat{\mathcal{P}}}^{\hat{\pi}^*} - V_{\mathcal{P}}^{\hat{\pi}^*}\right\|_\infty \\ &\le 2\sup_{\pi\in\Pi}\left\|V_{\mathcal{P}}^{\pi} - V_{\hat{\mathcal{P}}}^{\pi}\right\|_\infty\end{aligned}$$

This completes the proof. □

D Proofs of Properties of the Empirical Bellman Operator: f-Divergence Case

D.1 Proof of Proposition 3

Proof. Let

$$f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) = -c(k, \rho, |\mathcal{A}|) \left(\sum_{a \in \mathcal{A}} \mu_a \left[w_a^{k^*} \right] \right)^{1/k^*} + \sum_{a \in \mathcal{A}} \eta_a,$$

where $w_a = (\eta_a - d_a u)_+$. By definition, we have

$$\begin{aligned} & P\left(\sup_{\pi} \left| \hat{\mathbf{T}}^{\pi}(v)(s) - \mathcal{T}^{\pi}(v)(s) \right| > t \right) \\ & \leq P\left(\sup_{d \in \Delta(|\mathcal{A}|)} \left| \sup_{\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|}} f(\boldsymbol{\mu}_n, R(s,\cdot,\cdot) + \gamma v, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|}} f(\boldsymbol{\mu}, R(s,\cdot,\cdot) + \gamma v, \boldsymbol{\eta}) \right| > t \right). \end{aligned}$$

We analyze the sensitivity of the mapping $\boldsymbol{\mu} \to f(\boldsymbol{\mu}, u, \lambda)$. To control the difference between the empirical and the population objective, we establish the following lemma. The proof is deferred to Appendix D.2.

Lemma 11. *For any fixed u and π, there exists $t \in [0,1]$ such that*

$$\left| \sup_{\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|}} f(\boldsymbol{\mu}_n, u, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|}} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) \right| \leq 2\|u\|_{\infty,\boldsymbol{\mu}} \left\| \frac{dm_n}{d\mu_n(t)} \right\|_{\infty,\boldsymbol{\mu}}.$$

We decomposed the probability using the event $\Omega_{n,\delta_0}(\boldsymbol{\mu})$ where the empirical estimates are close to the population measures. By using lemma 11, we obtain

$$\begin{aligned} & P\left(\sup_{d \in \Delta(|\mathcal{A}|)} \left| \sup_{\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|}} f(\boldsymbol{\mu}_n, u, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|}} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) \right| > t \right) \\ & \leq P(\Omega_{n,\delta_0}(\boldsymbol{\mu})^c) + P\left(2\|u\|_{\infty} \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{\infty,\boldsymbol{\mu}} > t, \Omega_{n,\delta_0}(\boldsymbol{\mu}) \right) \end{aligned}$$

Again using Lemma 9, we have

$$\leq P\left(\sup_{a \in \mathcal{A}, s' \in \mathcal{S}} \left| \frac{\mu_{a,n}(s') - \mu_a(s')}{\mu_a(s')} \right| > \delta_0 \right) + P\left(2\|u\|_{\infty,\boldsymbol{\mu}} \sup_{a \in \mathcal{A}, s' \in \mathcal{S}} \left| \frac{\mu_{a,n}(s') - \mu_a(s')}{(1-\delta_0)\mu_a(s')} \right| > t \right).$$

By Chernoff Bound and Bernstein Inequality, we obtain

$$\begin{aligned} & \leq P\left(\sup_{a \in \mathcal{A}, s' \in \mathcal{S}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(S_i = s') - \mu_a(s') \right| > \delta_0 \mu_a(s') \right) \\ & + P\left(\frac{2}{1-\delta_0} \|u\|_{\infty} \sup_{a \in \mathcal{A}, s' \in \mathcal{S}} \frac{1}{\mu_a(s')} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(S_i = s') - \mu_a(s') \right| > t \right) \\ & \leq 2 \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \left(\exp\left(-\frac{\delta_0^2 n \mu_a(s')}{3} \right) + \exp\left(-\frac{t^2}{2} \left(\frac{4\|u\|_{\infty}^2}{(1-\delta_0)^2 n \mu_a(s')} + \frac{2\|u\|_{\infty} t}{3(1-\delta_0) n \mu_a(s')} \right)^{-1} \right) \right), \end{aligned}$$

Since $\mu_a(s') \geq \mathfrak{p}_{\wedge}$, and both exponential term above is monotonically decreasing over $\mu_a(s')$, we have

$$\leq 2|\mathcal{A}||\mathcal{S}| \left(\exp\left(-\frac{\delta_0^2 n \mathfrak{p}_{\wedge}}{3} \right) + \exp\left(-\frac{t^2}{2} \left(\frac{4\|u\|_{\infty}^2}{(1-\delta_0)^2 n \mathfrak{p}_{\wedge}} + \frac{2\|u\|_{\infty} t}{3(1-\delta_0) n \mathfrak{p}_{\wedge}} \right)^{-1} \right) \right)$$

Choose $\delta_0 = \frac{1}{2}$, by union bound, we obtain

$$\begin{aligned}
&P\left(\sup_{\pi\in\Pi}\left\|\hat{\mathbf{T}}^{\pi}(v)-\mathcal{T}^{\pi}(v)\right\|_{\infty}>t\right)\\
&\le P\left(\sup_{s\in\mathcal{S}}\sup_{d\in\Delta(\mathcal{A})}\left|\sup_{\boldsymbol{\eta}\in\mathbb{R}^{|\mathcal{A}|}} f(P_{s,n},R(s,\cdot,\cdot)+\gamma v,\boldsymbol{\eta})-\sup_{\boldsymbol{\eta}\in\mathbb{R}^{|\mathcal{A}|}} f(P_{s},R(s,\cdot,\cdot)+\gamma v,\boldsymbol{\eta})\right|>t\right)\\
&\le 2|\mathcal{S}|^2|\mathcal{A}|\exp\left(-\frac{n\mathfrak{p}_{\wedge}}{12}\right)+2|\mathcal{S}|^2|\mathcal{A}|\exp\left(-\frac{t^2}{2}\left(\frac{16\|R(s,\cdot,\cdot)+\gamma v\|_{\infty}^2}{n\mathfrak{p}_{\wedge}}+\frac{4\|R(s,\cdot,\cdot)+\gamma v\|_{\infty}t}{3\gamma n\mathfrak{p}_{\wedge}}\right)^{-1}\right).
\end{aligned}$$

Set each term to be less than $\eta/2$, by union bound, we need

$$n \ge \frac{12}{\mathfrak{p}_{\wedge}}\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right) \tag{D.1}$$

$$t \ge \frac{8\|R+\gamma v\|_{\infty}}{3n\mathfrak{p}_{\wedge}}\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)+\frac{4\|R+\gamma v\|_{\infty}}{\sqrt{n\mathfrak{p}_{\wedge}}}\sqrt{2\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}. \tag{D.2}$$

Under (D.1), we have

$$\frac{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}{n\mathfrak{p}_{\wedge}} \le \sqrt{\frac{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}{n\mathfrak{p}_{\wedge}}}.$$

By substituting this bound into (D.2), we obtain

$$\begin{aligned}
&\frac{8\|R+\gamma v\|_{\infty}}{3n\mathfrak{p}_{\wedge}}\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)+\frac{4\|R+\gamma v\|_{\infty}}{\sqrt{n\mathfrak{p}_{\wedge}}}\sqrt{2\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}\\
&\le\left(\frac{8}{3}+4\sqrt{2}\right)\frac{\|R+\gamma v\|_{\infty}}{\sqrt{n\mathfrak{p}_{\wedge}}}\sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}\\
&\le\frac{9\|R+\gamma v\|_{\infty}}{\sqrt{n\mathfrak{p}_{\wedge}}}\sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)}.
\end{aligned}$$

Therefore, when n satisfies (D.1) and t satisfies

$$t \ge \frac{9\|R+\gamma v\|_{\infty}}{\sqrt{n\mathfrak{p}_{\wedge}}}\sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)},$$

we have

$$P\left(\sup_{\pi\in\Pi}\left\|\hat{\mathbf{T}}^{\pi}(v)-\mathcal{T}^{\pi}(v)\right\|_{\infty}>t\right)\le\eta,$$

Since $|\sup_X f-\sup_X g|\le\sup_X|f-g|$,

$$\begin{aligned}
P\left(\sup_{\pi\in\Pi}\left\|\hat{\mathbf{T}}^{*}(v)-\mathcal{T}^{*}(v)\right\|_{\infty}>t\right)&=P\left(\left\|\sup_{\pi\in\Pi}\hat{\mathbf{T}}^{\pi}(v)-\sup_{\pi\in\Pi}\mathcal{T}^{\pi}(v)\right\|_{\infty}>t\right)\\
&\le P\left(\sup_{\pi\in\Pi}\left\|\hat{\mathbf{T}}^{\pi}(v)-\mathcal{T}^{\pi}(v)\right\|_{\infty}>t\right)\\
&\le\eta.
\end{aligned}$$

which implies the statement of the proposition. □

D.2 Proof of Lemma 11

Proof. We partition $\mathbb{R}^{|\mathcal{A}|}$ into three subsets, denoted by

$$\begin{aligned}
X_1&=\left\{\boldsymbol{\eta}\middle|\eta_a\le d_a\operatorname*{essinf}_{\mu_a}u\text{ for all }a\in\mathcal{A}\right\},\\
X_2&=\left\{\boldsymbol{\eta}\middle|d_a\operatorname*{essinf}_{\mu_a}u<\eta_a<M\text{ for all }a\in\mathcal{A}\right\},\\
X_3&=\mathbb{R}^{|\mathcal{A}|}\setminus(X_1\cup X_2).
\end{aligned}$$

where

$$M = \max_{a,s'} d_a u(s') + (c(k,\rho,|\mathcal{A}|) - |\mathcal{A}|^{1/k})^{-1} \sum_{a\in\mathcal{A}} d_a \max_{s'} u(s')$$

For any fixed $\boldsymbol{\mu}, \boldsymbol{\mu}_n$ and u, let

$$\begin{aligned} g(\boldsymbol{\eta}, t) &= f(\boldsymbol{\mu}_n(t), u, \boldsymbol{\eta}), \\ V(t) &= \sup_{\boldsymbol{\eta}\in X_2} g(\boldsymbol{\eta}, t). \end{aligned}$$

First we record the regularity properties of g that will be used to verify the assumptions of the envelope theorem.

Lemma 12. *The function $g(\boldsymbol{\eta}, t)$ has the following properties:*

- *For every fixed $t \in [0,1]$, the map $\boldsymbol{\eta} \mapsto g(\boldsymbol{\eta}, t)$ is continuous on X_2.*
- *For every fixed $\boldsymbol{\eta} \in X_2$, the map $t \mapsto g(\boldsymbol{\eta}, t)$ is absolutely continuous and differentiable on $[0,1]$.*
- *Under Ω_{n,δ_0}, there exists a finite constant b_0 such that*

$$|\partial_t g(\boldsymbol{\eta}, t)| \le b_0, \qquad \forall \boldsymbol{\eta} \in X_2,\ \forall t \in [0,1].$$

Lemma 13. *For any fixed measure vector $\boldsymbol{\mu} = (\mu_a)_{a\in\mathcal{A}}$, consider the optimization problem*

$$\sup_{\boldsymbol{\eta}\in\mathbb{R}^{|\mathcal{A}|}} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) = \sup_{\boldsymbol{\eta}\in\mathbb{R}^{|\mathcal{A}|}} -c(k,\rho,|\mathcal{A}|) \left(\sum_{a\in\mathcal{A}} \mu_a [w_a^{k^*}] \right)^{1/k^*} + \sum_{a\in\mathcal{A}} \eta_a,$$

where $w_a = (\eta_a - d_a u)_+$. Then:

- *this optimization problem admits an optimal solution $\boldsymbol{\eta}^*(\boldsymbol{\mu})$;*
- *any optimal solution belongs to $X_1 \cup X_2$;*
- *the optimal solution satisfies*

$$\left(\sum_{a\in\mathcal{A}} \mu_a [w_a^{k^*}] \right)^{1/k} = c(k,\rho,|\mathcal{A}|) \mu_i [w_i^{1/(k-1)}], \qquad \forall i \in \mathcal{A}; \tag{D.3}$$

- *if $\boldsymbol{\eta}^*(\boldsymbol{\mu}) \in X_2$, then*

$$f(\boldsymbol{\mu}, u, \boldsymbol{\eta}^*(\boldsymbol{\mu})) = -\frac{\sum_{a\in\mathcal{A}} \mu_a [w_a^{k^*}]}{\mu_i [w_i^{1/(k-1)}]} + \sum_{a\in\mathcal{A}} \eta_a^*(\boldsymbol{\mu}), \qquad \forall i \in \mathcal{A}.$$

The proof of Lemma 12 and Lemma 13 is deferred to Appendix D.3 and Appendix D.4, respectively.

When $\boldsymbol{\eta} \in X_1$, we have

$$\begin{aligned} \left| \sup_{\boldsymbol{\eta}\in X_1} f(\boldsymbol{\mu}_n, u, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta}\in X_1} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) \right| &\le \sup_{\boldsymbol{\eta}\in X_1} |f(\boldsymbol{\mu}_n, u, \boldsymbol{\eta}) - f(\boldsymbol{\mu}, u, \boldsymbol{\eta})| \\ &= \left| \left(-0 + \sum_{a\in\mathcal{A}} \eta_a \right) - \left(-0 + \sum_{a\in\mathcal{A}} \eta_a \right) \right| = 0 \end{aligned}$$

Otherwise, $\boldsymbol{\eta} \in X_2$. Before proceeding, we introduce the following version of the envelope theorem, which ensures the differentiability of $V(t)$ and provides an explicit formula for its derivative. This result allows us to apply the mean value theorem in the subsequent analysis.

Lemma 14 (Envelope theorem, Milgrom and Segal (2002), Theorem 2)**.** *Denote V as*

$$V(t) = \sup_{\mathbf{x}\in X} f(\mathbf{x}, t).$$

Suppose that $f(\mathbf{x}, \cdot)$ is absolutely continuous for all $\mathbf{x} \in X$. Suppose also that there exists an integrable function $b : [0, 1] \to \mathbb{R}_+$ such that $|\partial_t f(\mathbf{x}, t)| \le b(t)$ for all $\mathbf{x} \in X$ and almost all $t \in [0, 1]$. Then V is absolutely continuous. Suppose, in addition, that $f(\mathbf{x}, \cdot)$ is differentiable for all $\mathbf{x} \in X$, and that $X^*(t) \neq \varnothing$ almost everywhere on $[0, 1]$. Then for any selection $\mathbf{x}^*(t) \in X^*(t)$,

$$V(t) = V(0) + \int_0^t \partial_t f(\mathbf{x}^*(s), s)ds.$$

On the event Ω_{n,δ_0}, by Lemma 12, we have that $g(\boldsymbol{\eta}, \cdot)$ is absolutely continuous for all $\boldsymbol{\eta} \in X_2$, and there exists an integrable function b_0 such that $|\partial_t g(\boldsymbol{\eta}, t)| \le b_0$ for all $\boldsymbol{\eta} \in X_2$ and all $t \in [0, 1]$. Then by using Lemma 14, we have,

$$\begin{aligned}
\left|\sup_{\boldsymbol{\eta}\in X_2} f(\boldsymbol{\mu}_n, u, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta}\in X_2} f(\boldsymbol{\mu}, u, \boldsymbol{\eta})\right| &= |V(0) - V(1)| \\
&= \left|\int_0^1 \frac{d}{dt} g(\boldsymbol{\eta}^*(t), t)\, dt\right| \\
&\le \sup_{t\in[0,1]} \left|\frac{\partial}{\partial t} g(\boldsymbol{\eta}^*(t), t)\right|.
\end{aligned}$$

Next, we bound for any $\tau \in [0, 1]$

$$\begin{aligned}
\left|\frac{\partial}{\partial t} g(\boldsymbol{\eta}^*(t), t)\Big|_{t=\tau}\right| &= \left|\frac{c(k, \rho, |\mathcal{A}|)}{k^* \left(\sum_{a\in\mathcal{A}} \mu_{a,n}(\tau) \left[w_a^{k^*}\right]\right)^{1/k}} \sum_{a\in\mathcal{A}} m_{a,n} \left[w_a^{k^*}\right]\right| \\
&= \left|\sum_{a\in\mathcal{A}} \frac{c(k, \rho, |\mathcal{A}|)}{k^* \left(\sum_{\ell\in\mathcal{A}} \mu_{\ell,n}(\tau) \left[w_\ell^{k^*}\right]\right)^{1/k}} m_{a,n} \left[w_a^{k^*}\right]\right|.
\end{aligned}$$

By using (D.3), we have

$$\frac{c(k, \rho, |\mathcal{A}|)}{k^* \left(\sum_{a\in\mathcal{A}} \mu_{a,n}(t) \left[w_a^{k^*}\right]\right)^{1/k}} = \frac{1}{k^* \cdot \mu_{a,n}(t) \left[w_a^{1/(k-1)}\right]}, \qquad \forall a \in \mathcal{A}.$$

Therefore, combining the previous displays,

$$\begin{aligned}
\left|\frac{\partial}{\partial t} g(\boldsymbol{\eta}^*(t), t)\Big|_{t=\tau}\right| &= \left|\sum_{a\in\mathcal{A}} \frac{m_{a,n} \left[w_a^{k^*}\right]}{k^* \cdot \mu_{a,n}(\tau) \left[w_a^{1/(k-1)}\right]}\right| \\
&\le \sum_{a\in\mathcal{A}} \left|\frac{m_{a,n} \left[w_a^{k^*}\right]}{k^* \cdot \mu_{a,n}(\tau) \left[w_a^{1/(k-1)}\right]}\right|.
\end{aligned}$$

For $\eta_a \ge d_a \operatorname{essinf}_{\mu_a} u$, by mean value theorem, there exists $\xi \in (\eta_a - d_a\|u\|_{L^\infty(\mu_a)}, \eta_a - d_a u)$ satisfies

$$\begin{aligned}
\left|\frac{m_{a,n} \left[w_a^{k^*}\right]}{k^* \cdot \mu_{a,n}(\tau) \left[w_a^{1/(k-1)}\right]}\right| &\overset{(i)}{=} \left|\frac{m_{a,n} \left[(\eta_a - d_a u)_+^{k^*} - (\eta_a - d_a\|u\|_{L^\infty(\mu_a)})_+^{k^*}\right]}{\mu_{a,n}(\tau) \left[k^* (\eta_a - d_a u)_+^{1/(k-1)}\right]}\right| \\
&= \left|\frac{m_{a,n} \left[d_a(u - \|u\|_{L^\infty(\mu_a)}) k^* (\xi)_+^{1/(k-1)}\right]}{\mu_{a,n}(\tau) \left[k^* (\eta_a - d_a u)_+^{1/(k-1)}\right]}\right|,
\end{aligned}$$

where (i) follows from the fact that $m_{a,n}[c] = 0$ for any constant function c. Since $\xi < \eta_a - d_a u$, we have

$$\begin{aligned}
&\left| \frac{m_{a,n}\left[d_a(u - \|u\|_{L^\infty(\mu_a)})k^*(\xi)_+^{1/(k-1)}\right]}{\mu_{a,n}(\tau)\left[k^*\left(\eta_a - d_a u\right)_+^{1/(k-1)}\right]} \right| \\
&\leq \left\| \frac{d_a(u - \|u\|_{L^\infty(\mu_a)})k^*(\xi)_+^{1/(k-1)}}{k^*\left(\eta_a - d_a u\right)_+^{1/(k-1)}} \right\|_{L^\infty(\mu_a)} \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{L^\infty(\mu_a)} \\
&\leq \left\| \frac{d_a(u - \|u\|_{L^\infty(\mu_a)})(\eta_a - d_a u)_+^{1/(k-1)}}{(\eta_a - d_a u)_+^{1/(k-1)}} \right\|_{L^\infty(\mu_a)} \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{L^\infty(\mu_a)} \\
&= 2d_a\|u\|_{L^\infty(\mu_a)} \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{L^\infty(\mu_a)}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\left| \frac{\partial}{\partial t} g(\boldsymbol{\eta}^*(t), t)\Big|_{t=\tau} \right| &= \sum_{a\in\mathcal{A}} \left| \frac{m_{a,n}\left[w_a^{k^*}\right]}{k^* \cdot \mu_{a,n}(\tau)\left[w_a^{1/(k-1)}\right]} \right| \\
&\leq \sum_{a\in\mathcal{A}} 2d_a\|u\|_{L^\infty(\mu_a)} \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{L^\infty(\mu_a)} \\
&= 2\|u\|_\infty \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{\infty,\boldsymbol{\mu}}.
\end{aligned}$$

To conclude the lemma, we we observe that $\boldsymbol{\eta} \in X_1$,

$$\left| \sup_{\boldsymbol{\eta}\in X_1} f(\boldsymbol{\mu}_n, u, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta}\in X_1} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) \right| = 0.$$

When $\boldsymbol{\eta} \in X_2$,

$$\begin{aligned}
\left| \sup_{\boldsymbol{\eta}\in X_2} f(\boldsymbol{\mu}_n, u, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta}\in X_2} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) \right| &\leq \sup_{t\in[0,1]} \left| \frac{\partial}{\partial t} g(\boldsymbol{\eta}^*(t), t) \right| \\
&\leq \sup_{t\in[0,1]} 2\|u\|_\infty \left\| \frac{dm_n}{d\mu_n(t)} \right\|_{\infty,\boldsymbol{\mu}} \\
&= 2\|u\|_\infty \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{\infty,\boldsymbol{\mu}}
\end{aligned}$$

for some $\tau \in [0,1]$, where the existence of τ follows from the compactness of $[0,1]$ and the continuity of $t \to \mu_n(t)$ and norms.

Finally, using $|\max\{C, x\} - \max\{C, y\}| \leq |x - y|$, we obtain

$$\left| \sup_{\boldsymbol{\eta}\in X_1\cup X_2} f(\boldsymbol{\mu}_n, u, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta}\in X_1\cup X_2} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) \right| \leq \left| \sup_{\boldsymbol{\eta}\in X_2} f(\boldsymbol{\mu}_n, u, \boldsymbol{\eta}) - \sup_{\boldsymbol{\eta}\in X_2} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) \right| \leq 2\|u\|_\infty \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{\infty,\boldsymbol{\mu}}$$

□

D.3 Proof of Lemma 12

Let

$$\Phi(\boldsymbol{\eta}, t) := \sum_{a\in\mathcal{A}} \mu_{a,n}(t)\left[(\eta_a - d_a u)_+^{k^*}\right].$$

For every fixed $t \in [0,1]$, since $u \in L^\infty$ and $\eta_a < M$ on X_2, the map

$$\eta_a \mapsto (\eta_a - d_a u)_+$$

is uniformly Lipschitz on $(-\infty, M]$. Hence $\Phi(\cdot, t)$ is continuous on X_2, and therefore $g(\cdot, t)$ is continuous on X_2. Now fix $\boldsymbol{\eta} \in X_2$. Then

$$\Phi(\boldsymbol{\eta}, t) = \sum_{a \in \mathcal{A}} \left(t\mu_a \left[w_a^{k^*} \right] + (1-t)\mu_{a,n} \left[w_a^{k^*} \right] \right),$$

so $\Phi(\boldsymbol{\eta}, \cdot)$ is affine in t. Moreover, under Ω_{n,δ_0}, since $\eta_a > d_a \operatorname{essinf}_{\mu_a} u$, we have $\mu_a \left[w_a^{k^*} \right] > 0$, and Lemma 9 yields

$$\mu_{a,n} \left[w_a^{k^*} \right] \geq (1-\delta_0)\mu_a \left[w_a^{k^*} \right] > 0, \qquad \forall a \in \mathcal{A}, \forall t \in [0,1].$$

Therefore $\Phi(\boldsymbol{\eta}, t) > 0$ for all $t \in [0,1]$. Since $x \mapsto x^{1/k^*}$ is C^1 on $(0, \infty)$, it follows that $t \mapsto g(\boldsymbol{\eta}, t)$ is C^1 on $[0,1]$, hence absolutely continuous and differentiable on $[0,1]$.

Finally we check under Ω_{n,δ_0}, when $\boldsymbol{\eta} \in X_2$, $\partial_t g(\boldsymbol{\eta}, t)$ has an upper bound. Recall that

$$\partial_t g(\boldsymbol{\eta}, t) = -c(k, \rho, |\mathcal{A}|) \frac{\sum_{a \in \mathcal{A}} m_{a,n} \left[w_a^{k^*} \right]}{k^* \left(\sum_{a \in \mathcal{A}} \mu_{a,n}(t) \left[w_a^{k^*} \right] \right)^{1/k}}.$$

Let

$$W := \max_{a \in \mathcal{A}} \left(M - d_a \operatorname*{essinf}_{\mu_a} u \right) \leq \infty,$$

by the definition of w_a, for every $a \in \mathcal{A}$ and s,

$$w_a(s) = (\eta_a - d_a u(s))_+ \leq M - d_a \operatorname*{essinf}_{\mu_a} u \leq W.$$

Applying Lemma 9, we have

$$\left| m_{a,n} \left[w_a^{k^*} \right] \right| = \left| \mu_{a,n} \left[w_a^{k^*} \right] - \mu_a \left[w_a^{k^*} \right] \right| \leq \delta_0 \mu_{a,n} \left[w_a^{k^*} \right] \leq \frac{\delta_0}{1-\delta_0} \mu_{a,n}(t) \left[w_a^{k^*} \right].$$

Therefore,

$$\begin{aligned} |\partial_t g(\boldsymbol{\mu}, t)| &\leq \frac{c(k, \rho, |\mathcal{A}|)}{k^*} \frac{\delta_0}{1-\delta_0} \frac{\sum_{a \in \mathcal{A}} \mu_{a,n}(t) \left[w_a^{k^*} \right]}{\left(\sum_{a \in \mathcal{A}} \mu_{a,n}(t) \left[w_a^{k^*} \right] \right)^{1/k}} \\ &= \frac{c(k, \rho, |\mathcal{A}|)}{k^*} \frac{\delta_0}{1-\delta_0} \left(\sum_{a \in \mathcal{A}} \mu_{a,n}(t) \left[w_a^{k^*} \right] \right)^{1/k^*} \\ &\leq \frac{c(k, \rho, |\mathcal{A}|)}{k^*} \frac{\delta_0}{1-\delta_0} |\mathcal{A}|^{1/k^*} W < \infty. \end{aligned}$$

$|\partial_t g(\boldsymbol{\mu}, t)|$ is bounded.

D.4 Proof of Lemma 13

We prove the lemma in several steps.

Step 1: Existence of an optimizer. We choose a closed set

$$D := \left\{ \boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|} : \eta_a \geq d_a \operatorname*{essinf}_{\mu_a} u \right\}.$$

For any $\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|}$, let $\eta_a = \max\{\eta_a, d_a \operatorname{essinf}_{\mu_a} u\}$. Since $(\eta_a - d_a u)_+ = 0$ for $\eta_a \leq \operatorname{essinf}_{\mu_a} u\}$,

$$f(\boldsymbol{\mu}, u, \tilde{\boldsymbol{\eta}}) \geq f(\boldsymbol{\mu}, u, \boldsymbol{\eta}).$$

Therefore

$$\sup_{\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|}} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) = \sup_{\boldsymbol{\eta} \in D} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}).$$

Next we prove that f is coercive on D. Let

$$U_a := d_a \operatorname*{esssup}_{\mu_a} u,$$

then we have

$$(\mu_a - d_a u(s'))_+ \geq (\mu_a - U_a)_+ \qquad \forall s' \in \mathcal{S},$$

therefore

$$\mu_a \left[(\eta_a - d_a u)_+^{k^*} \right] \geq (\eta_a - U_a)_+^{k^*}.$$

Then we have

$$\begin{aligned} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) &= -c(k, \rho, |\mathcal{A}|) \left(\sum_{a \in \mathcal{A}} \mu_a \left[(\eta_a - d_a u)_+^{k^*} \right] \right)^{1/k^*} + \sum_{a \in \mathcal{A}} \eta_a \\ &\leq -c(k, \rho, |\mathcal{A}|) \left(\sum_{a \in \mathcal{A}} (\eta_a - U_a)_+^{k^*} \right)^{1/k^*} + \sum_{a \in \mathcal{A}} \eta_a. \end{aligned}$$

Since $\|x\|_k^* \geq |\mathcal{A}|^{-1/k} \|x\|_1$,

$$\begin{aligned} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) &\leq -c(k, \rho, |\mathcal{A}|) |\mathcal{A}|^{-1/k} \sum_{a \in \mathcal{A}} (\eta_a - U_a)_+ + \sum_{a \in \mathcal{A}} \eta_a \\ &\leq -c(k, \rho, |\mathcal{A}|) |\mathcal{A}|^{-1/k} \sum_{a \in \mathcal{A}} (\eta_a - U_a) + \sum_{a \in \mathcal{A}} \eta_a \\ &= \left(1 - c(k, \rho, |\mathcal{A}|) |\mathcal{A}|^{-1/k} \right) \sum_{a \in \mathcal{A}} \eta_a + c(k, \rho, |\mathcal{A}|) |\mathcal{A}|^{-1/k} \sum_{a \in \mathcal{A}} U_a. \end{aligned}$$

Since $\rho > 0$, we have $c(k, \rho, |\mathcal{A}|) > |\mathcal{A}|^{1/k}$, therefore $1 - c(k, \rho, |\mathcal{A}|)|\mathcal{A}|^{-1/k} < 0$. Therefore, when $\|\boldsymbol{\eta}\| \to \infty$, $f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) \to -\infty$. Therefore, f is coercive on D. Since f is continuous on D, there exists $\boldsymbol{\eta}^* \in D$ such that

$$\sup_{\boldsymbol{\eta} \in \mathbb{R}^{|\mathcal{A}|}} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) = \sup_{\boldsymbol{\eta} \in D} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) = f(\boldsymbol{\mu}, u, \boldsymbol{\eta}^*).$$

Step 2: First-order characterization of an optimal solution. Define

$$\Phi(\boldsymbol{\eta}) := \sum_{a \in \mathcal{A}} \mu_a \left[(\eta_a - d_a u)_+^{k^*} \right].$$

We distinguish two cases.

Case 1: $\Phi(\boldsymbol{\eta}^*) = 0$. Since each term in the sum is nonnegative, we have

$$\mu_a \left[(\eta_a^* - d_a u)_+^{k^*} \right] = 0, \qquad \forall a \in \mathcal{A}.$$

Hence

$$(\eta_a^* - d_a u)_+ = 0 \qquad \mu_a\text{-a.s.},$$

which implies

$$\eta_a^* \leq d_a \operatorname*{essinf}_{\mu_a} u, \qquad \forall a \in \mathcal{A}.$$

On the other hand, by Step 1 we already know that $\boldsymbol{\eta}^* \in D$, so

$$\eta_a^* \geq d_a \operatorname*{essinf}_{\mu_a} u, \qquad \forall a \in \mathcal{A}.$$

Therefore,

$$\eta_a^* = d_a \operatorname*{essinf}_{\mu_a} u, \qquad \forall a \in \mathcal{A},$$

and hence $\boldsymbol{\eta}^* \in X_1$. In this case,

$$\mu_i[w_i^{1/(k-1)}] = 0, \qquad \forall i \in \mathcal{A},$$

so (D.3) holds trivially.

Case 2: $\Phi(\boldsymbol{\eta}^*) > 0$. We first show that

$$\eta_i^* > d_i \operatorname*{essinf}_{\mu_i} u, \qquad \forall i \in \mathcal{A}.$$

Suppose, to the contrary, that for some $i \in \mathcal{A}$,

$$\eta_i^* = d_i \operatorname*{essinf}_{\mu_i} u.$$

Then

$$(\eta_i^* - d_i u)_+ = 0 \qquad \mu_i\text{-a.s.},$$

and therefore

$$\mu_i[w_i^{1/(k-1)}] = 0.$$

Since $\Phi(\boldsymbol{\eta}^*) > 0$, the function $f(\boldsymbol{\mu}, u, \cdot)$ is differentiable at $\boldsymbol{\eta}^*$, and

$$\frac{\partial}{\partial \eta_i} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}^*) = 1 - c(k, \rho, |\mathcal{A}|) \left(\sum_{a \in \mathcal{A}} \mu_a[w_a^{k^*}] \right)^{-1/k} \mu_i[w_i^{1/(k-1)}] = 1.$$

Moreover, since $\boldsymbol{\eta}^* \in D$, the direction $+e_i$ is feasible. Hence, for all sufficiently small $t > 0$,

$$f(\boldsymbol{\mu}, u, \boldsymbol{\eta}^* + te_i) > f(\boldsymbol{\mu}, u, \boldsymbol{\eta}^*),$$

which contradicts the optimality of $\boldsymbol{\eta}^*$. Therefore,

$$\eta_i^* > d_i \operatorname*{essinf}_{\mu_i} u, \qquad \forall i \in \mathcal{A}.$$

In particular, $\boldsymbol{\eta}^* \in X_2$, so $\boldsymbol{\eta}^*$ is an interior point of D.

Since $f(\boldsymbol{\mu}, u, \cdot)$ is concave in $\boldsymbol{\eta}$, we may now apply the first-order optimality condition at the interior maximizer $\boldsymbol{\eta}^*$, which gives

$$\frac{\partial}{\partial \eta_i} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}^*) = 1 - c(k, \rho, |\mathcal{A}|) \left(\sum_{a \in \mathcal{A}} \mu_a[w_a^{k^*}] \right)^{-1/k} \mu_i[w_i^{1/(k-1)}] = 0, \qquad \forall i \in \mathcal{A}.$$

Equivalently,

$$\left(\sum_{a \in \mathcal{A}} \mu_a[w_a^{k^*}] \right)^{1/k} = c(k, \rho, |\mathcal{A}|) \mu_i[w_i^{1/(k-1)}], \qquad \forall i \in \mathcal{A},$$

which is exactly (D.3).

Step 3: Value of the objective when $\boldsymbol{\eta}^*(\boldsymbol{\mu}) \in X_2$. When $\boldsymbol{\eta}^*(\boldsymbol{\mu}) \in X_2$, we have

$$\mu_i[w_i^{1/(k-1)}] > 0, \qquad \forall i \in \mathcal{A}.$$

Substituting (D.3) into the definition of f, we obtain

$$\begin{aligned} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}^*) &= -c(k, \rho, |\mathcal{A}|) \left(\sum_{a \in \mathcal{A}} \mu_a[w_a^{k^*}] \right)^{1/k^*} + \sum_{a \in \mathcal{A}} \eta_a^* \\ &= -\frac{\sum_{a \in \mathcal{A}} \mu_a[w_a^{k^*}]}{\mu_i[w_i^{1/(k-1)}]} + \sum_{a \in \mathcal{A}} \eta_a^*, \qquad \forall i \in \mathcal{A}. \end{aligned}$$

Step 4: Any optimizer belongs to $X_1 \cup X_2$.

We first prove that

$$\boldsymbol{\eta}^* \notin X_{3,1} := \left\{ \boldsymbol{\eta} \middle| \exists a_1, a_2 \in \mathcal{A}, \eta_{a_1} > d_{a_1} \operatorname*{essinf}_{\mu_{a_1}} u, \eta_{a_2} \leq d_{a_2} \operatorname*{essinf}_{\mu_{a_2}} u \right\}$$

by contradictory.

Suppose $\boldsymbol{\eta}^* \in X_{3,1}$. Then, according to the definition of $X_{3,1}$, $\mu_{a_2}[w_{a_2}^{1/(k-1)}] = 0$. According to (D.3),

$$\left(\sum_{a \in \mathcal{A}} \mu_a \left[w_a^{k^*} \right] \right)^{1/k} = 0,$$

which implies

$$\mu_a \left[w_a^{k^*} \right] = 0 \quad \text{for all } a \in \mathcal{A}.$$

Therefore, $\boldsymbol{\eta}^* \in X_1$, which contradicts the assumption that $\boldsymbol{\eta}^* \in X_{3,1}$. Next, we prove that

$$\boldsymbol{\eta}^* \notin X_{3,2} := \left\{ \boldsymbol{\eta} \middle| \exists a \in \mathcal{A}, \eta_a \geq M \right\}$$

where

$$M = \max_{a,s'} d_a u(s') + \frac{\sum_{a \in \mathcal{A}} d_a \max_{s'} u(s')}{c(k, \rho, |\mathcal{A}|) - |\mathcal{A}|^{1/k}}.$$

We first bound the first term in f

$$(\eta_a - d_a u(s'))_+ \geq \left(\eta_a - \max_{s'} d_a u(s') \right)_+$$

taking expectation and sum over a on both side of the inequality

$$\sum_{a \in \mathcal{A}} \mu_a \left[(\eta_a - d_a u)_+^{k^*} \right] \geq \sum_{a \in \mathcal{A}} \left(\eta_a - \max_{s'} d_a u(s') \right)_+^{k^*}.$$

Therefore,

$$\left(\sum_{a \in \mathcal{A}} \mu_a \left[w_a^{k^*} \right] \right)^{1/k^*} \geq \left(\sum_{a \in \mathcal{A}} \left(\eta_a - \max_{s'} d_a u(s') \right)_+^{k^*} \right)^{1/k^*}.$$

Then we bound the second term in f

$$\sum_{a \in \mathcal{A}} \eta_a \leq \sum_{a \in \mathcal{A}} (\eta_a - \max_{s'} d_a u(s'))_+ + \sum_{a \in \mathcal{A}} \max_{s'} d_a u(s').$$

According to Hölder's inequality, we have

$$\sum_{a \in \mathcal{A}} (\eta_a - \max_{s'} d_a u(s'))_+ \leq |\mathcal{A}|^{1/k} \left(\sum_{a \in \mathcal{A}} (\eta_a - \max_{s'} d_a u(s'))_+^{k^*} \right)^{1/k^*}.$$

Therefore,

$$\sum_{a \in \mathcal{A}} \eta_a \leq |\mathcal{A}|^{1/k} \left(\sum_{a \in \mathcal{A}} (\eta_a - \max_{s'} d_a u(s'))_+^{k^*} \right)^{1/k^*} + \sum_{a \in \mathcal{A}} \max_{s'} d_a u(s').$$

To sum up, we have

$$\begin{aligned} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) &= -c(k, \rho, |\mathcal{A}|) \left(\sum_{a \in \mathcal{A}} \mu_a \left[w_a^{k^*} \right] \right)^{1/k^*} + \sum_{a \in \mathcal{A}} \eta_a \\ &\leq -\left(c(k, \rho, |\mathcal{A}|) - |\mathcal{A}|^{1/k} \right) \left(\sum_{a \in \mathcal{A}} \left(\eta_a - \max_{s'} d_a u(s') \right)_+^{k^*} \right)^{1/k^*} + \sum_{a \in \mathcal{A}} \max_{s'} d_a u(s'). \end{aligned}$$

Note that when $\rho > 0$,

$$c(k, \rho, |\mathcal{A}|) - |\mathcal{A}|^{1/k} = |\mathcal{A}|^{1/k} \left((k(k-1)\rho + 1)^{1/k} - 1 \right) > 0$$

Suppose there exists a_1 such that $\eta_{a_1} > M$, then

$$\begin{aligned} f(\boldsymbol{\mu}, u, \boldsymbol{\eta}) &\leq -\left(c(k, \rho, |\mathcal{A}|) - |\mathcal{A}|^{1/k} \right) \left(\eta_{a_1} - \max_{s'} d_{a_1} u(s') \right)_+ + \sum_{a \in \mathcal{A}} \max_{s'} d_a u(s') \\ &< -\left(c(k, \rho, |\mathcal{A}|) - |\mathcal{A}|^{1/k} \right) \left(\frac{\sum_{a \in \mathcal{A}} d_a \max_{s'} u(s')}{c(k, \rho, |\mathcal{A}|) - |\mathcal{A}|^{1/k}} \right) + \sum_{a \in \mathcal{A}} \max_{s'} d_a u(s') \\ &< 0 \end{aligned}$$

Therefore, $\boldsymbol{\eta}$ can not be the optimal solution when $\boldsymbol{\eta} \in X_{3,2}$. It can be easily check that $X_3 = X_{3,1} \cup X_{3,2}$, and overall, we have $\boldsymbol{\eta}^* \notin X_3$.

D.5 Proof of Theorem 2

Proof. Substituting $\|R + \gamma v\|_\infty \leq 1/(1-\gamma)$ into the bound from Proposition 3 and applying Proposition 1, we obtain the stated result.

$$\begin{aligned} \|\hat{v} - v^*\|_\infty &\leq \frac{1}{1-\gamma} \left\| \hat{\mathbf{T}}^*(v^*) - \mathcal{T}^*(v^*) \right\|_\infty \\ &\leq \frac{9\|R + \gamma v\|_\infty}{(1-\gamma)\sqrt{n\mathfrak{p}_\wedge}} \sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)} \\ &\leq \frac{9}{(1-\gamma)^2\sqrt{n\mathfrak{p}_\wedge}} \sqrt{\log\left(4|\mathcal{S}|^2|\mathcal{A}|/\eta\right)} \end{aligned}$$

with probability $1 - \eta$. The preceding step provides a uniform bound on the value-function estimation error. We now turn to the policy suboptimality gap. According to Corollary 2, Lemma 10 and Proposition 3, we have

$$\begin{aligned} \sup_{\pi \in \Pi} V_{\mathcal{P}}^{\pi}(s) - V_{\mathcal{P}}^{\hat{\pi}^*}(s) &\leq 2 \sup_{\pi \in \Pi} \left\| V_{\hat{\mathcal{P}}}^{\pi} - V_{\mathcal{P}}^{\pi} \right\|_\infty \\ &\leq \frac{2}{1-\gamma} \sup_{\pi \in \Pi} \left\| \hat{\mathbf{T}}^{\pi}(V_{\mathcal{P}}^{\pi}) - \mathcal{T}^{\pi}(V_{\mathcal{P}}^{\pi}) \right\|_\infty \\ &\leq \frac{18\|R + \gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\wedge}(1-\gamma)} \sqrt{\log\left(4|S|^2|\mathcal{A}|/\eta\right)} \\ &\leq \frac{18}{\sqrt{n\mathfrak{p}_\wedge}(1-\gamma)^2} \sqrt{\log\left(4|S|^2|\mathcal{A}|/\eta\right)} \end{aligned}$$

□