
Near-Optimal Sample Complexities of Divergence-based S-rectangular Distributionally Robust Reinforcement Learning

Anonymous Author(s)

Affiliation

Address

email

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 $\tilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-4}\varepsilon^{-2})$, 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.

1 Introduction

Reinforcement learning (RL) Sutton and Barto [20] 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 RL’s 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. [32] to address this mismatch and has since been further developed in a series of works, including Panaganti and Kalathil [13], Yang et al. [30], Xu et al. [28], Blanchet et al. [1], Liu et al. [10], Wang et al. [21], Yang et al. [31], Wang et al. [25], Shi and Chi [17].

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 adversary’s temporal flexibility in selecting distributions [8], 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 [9] and Wiesemann et al. [26]. 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 controller’s 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 controller’s action A_t —for example, assigning low demand when A_t is large and high demand when A_t is small—granting the adversary excessive power and resulting in an unrealistic model.

This example highlights how S-rectangularity constrains the adversary’s power by preventing it from adapting to the controller’s 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. [7] 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 [7, 30], 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. [30] contains a sample complexity dependence on $|S|$ and $|A|$ in the form of $O(|S|^2|A|^2)$, where $|S|$ and $|A|$ are the cardinalities of the state and action spaces. This significantly deviates from the known lower bound of $\Omega(|S||A|)$. 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 $\tilde{O}(|S||A|(1 - \gamma)^{-4}\varepsilon^{-2})$, where ε is the target accuracy and γ is the discount factor. This bound is optimal in its dependence on $|S|$, $|A|$, and ε , and it holds uniformly over the entire range of uncertainty sizes $\rho \in (0, +\infty)$ and discount factors $\gamma \in (0, 1)$.

89 To the best of our knowledge, this is the first sample complexity upper bound for divergence-based
90 S-rectangular models that simultaneously achieves optimal dependence on $|\mathcal{S}|$, $|\mathcal{A}|$, and ε .

91 To achieve the optimal $|\mathcal{S}||\mathcal{A}|$ dependence, we develop a refined sensitivity analysis that improves
92 upon the metric entropy bounds derived from the covering numbers of the randomized policy
93 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 $\mathcal{A}$,
94 as used in Yang et al. [30]. Moreover, our analyses advance the techniques of Wang et al. [25] by
95 relaxing the mutual absolute continuity requirement, thereby extending the allowable range of the
96 uncertainty radius to $\mathbb{R}_+$, beyond the previously restrictive regime of $\rho = O(p_\wedge)$, while retain an
97 $O(1)$ dependence on ρ as $\rho \downarrow 0$.

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

103 2 Literature Review

104 In this section, we briefly survey SA-rectangular and S-rectangular distributionally robust reinforce-
105 ment learning.

106 **SA-rectangular DR-RL:** The dynamic programming principles for SA-rectangular distributionally
107 robust Markov decision processes (DR-MDPs) have been gradually established through a series of
108 works under different information structures [6, 8, 12, 15, 24]. Recent advances in SA-rectangular
109 distributionally robust reinforcement learning (DR-RL) have explored sample complexity in various
110 settings. Broadly speaking, model-based approaches have been studied in Zhou et al. [33], Panaganti
111 and Kalathil [13], Yang et al. [30], Shi and Chi [16], Xu et al. [28], Shi et al. [18], Blanchet et al. [1],
112 while the statistical properties of model-free algorithms are presented in Liu et al. [10], Wang et al.
113 [21, 22], Yang et al. [31].

114 **S-rectangular DR-RL:** To extend the flexibility of robust MDP models, S-rectangularity was
115 introduced in Xu and Mannor [27], Wiesemann et al. [26] as an overarching theoretical framework to
116 constrain the adversary while retaining a dynamic programming equation. Ho et al. [7] developed an
117 efficient optimization algorithm to solve the Bellman update under this structure. On the statistical
118 side, Yang et al. [30] provided the first sample complexity result for S-rectangular DR-RL, achieving
119 a rate of $\tilde{O}(|\mathcal{S}|^2|\mathcal{A}|^2(1-\gamma)^{-4}\varepsilon^{-2})$, which is suboptimal in its dependence on the number of states
120 and actions. More recently, Clavier et al. [2] established near-optimal rates for the S-rectangular
121 setting under general L_p norm uncertainty sets. However, their analysis does not directly extend to
122 divergence-based uncertainty sets.

123 3 Learning S-rectangular Robust Markov Decision Processes

124 3.1 Classical Markov Decision Processes

125 We briefly review and establish notation for classical tabular MDP models. Let $\Delta(\mathcal{S})$, $\Delta(\mathcal{A})$ denote
126 the probability simplex over the finite state space $\mathcal{S}$ and action space $\mathcal{A}$ respectively. An infinite hori-
127 zon 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,
128 respectively; $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the reward function; $P = \{P_{s,a}(\cdot) \in \Delta(\mathcal{S}) : (s, a) \in \mathcal{S} \times \mathcal{A}\}$
129 is the controlled transition kernel; and $\gamma \in (0, 1)$ is the discount factor. Through out the paper,
130 given a controlled transition kernel P , we denote $P_s := (P_{s,a})_{a \in \mathcal{A}}$ which is seen as a function
131 $P_s : \mathcal{A} \rightarrow \Delta(\mathcal{S})$.

132 We define the measurable space $(\Omega, \mathcal{F})$ to be the canonical space $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$ equipped with the
133 σ -field generated by cylinder sets. Define state-action process $(S_t, A_t)_{t \geq 0}$ by the point evaluation
134 $X_t(\omega) = s_t, A_t(\omega) = a_t$ for all $t \geq 0$ and $\omega = (s_0, a_0, s_1, a_1, \dots) \in \Omega$.

135 An agent may optimize over the class of history-dependent policies, denoted by Π_{HD} , where each
136 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
137 specifies the conditional distribution of the action A_t given the full history, that is, a mapping
138 $\pi_t : (\mathcal{S} \times \mathcal{A})^t \times \mathcal{S} \rightarrow \Delta(\mathcal{A})$. 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 $\pi : \mathcal{S} \rightarrow \mathcal{A}$ [14].

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} [26]. 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 $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$, 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 $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$ 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. [26], S-rectangularity). The uncertainty set $\mathcal{P}$ is S-rectangular if $\mathcal{P} = \times_{s \in \mathcal{S}} \mathcal{P}_s$ 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 $\bar{P}$. 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 [30, 7].

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

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

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

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

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

Definition 2 (DR Bellman Equation). Given S-rectangular $\mathcal{P} = \times_{s \in \mathcal{S}} \mathcal{P}_s$, the DR Bellman equation is the following fixed-point equation for $v : \mathcal{S} \rightarrow \mathbb{R}$

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

It is well known [26] 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 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 [8, 26, 24]. This framework can be extended to more general Markovian or history-dependent adversarial models, while still preserving Markov optimality [24].

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

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

for all $s \in \mathcal{S}$. The optimal DR Bellman operator is $\mathcal{T}^*(v)(s) := \sup_{\pi \in \Pi} \mathcal{T}^{\pi}(v)(s), \forall s \in \mathcal{S}$.

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 $\bar{P}_{s,a}$, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. In particular, given sample size n , we sample i.i.d. $\{S_{s,a}^{(1)}, \dots, S_{s,a}^{(n)}\}$ from $\bar{P}_{s,a}$ and construct the empirical transition probability

$$\bar{P}_{s,a,n}(s') := \frac{1}{n} \sum_{i=1}^n \mathbb{1} \{S_{s,a}^{(i)} = s'\}. \quad (3.6)$$

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

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

Equipped with these definitions, we present our strategy to estimate the optimal value 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 Estimator). Given divergence function f and radius parameter ρ , let $\mathcal{P} = \mathcal{P}(f, \rho)$ and $\mathcal{P}_n = \mathcal{P}_n(f, \rho)$. We define the empirical Bellman estimator $\hat{v}$ to $V_{\mathcal{P}}^*$ as the unique solution to the fixed point equation $\hat{\mathbf{T}}^*(\hat{v}) = \hat{v}$.

The rest of this paper is dedicated to theoretical analyses and numerical validation of the statistical efficiency of estimating $V_{\mathcal{P}}^* = v^*$ using $\hat{v}$. We conclude this section by introducing the following important proposition that provides an upper bound on the l_{∞} estimation error.

Proposition 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

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

with probability 1.

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 $f_{\text{KL}}(t) = f(t) = t \log t$, the corresponding uncertainty set $\mathcal{P}_s(f_{\text{KL}}, \rho)$ is based on the Kullback–Leibler (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 [4].

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 [25].

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

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

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 [19]. Similar negative results hold in the f_k -divergence setting when the parameter k approaches 1 [3], 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.

Lemma 1. With $\mathcal{P} = \mathcal{P}(f_{\text{KL}}, \rho)$ where $f_{\text{KL}}(t) = t \log t$ 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

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

The KL empirical DR Bellman operator $\hat{\mathbf{T}}^\pi$ satisfies (4.1) with $\bar{P}_{s,a}$ replaced by $\bar{P}_{s,a,n}$.

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 $v : \mathcal{S} \rightarrow \mathbb{R}$ and $n \geq 12\mathfrak{p}_\wedge^{-1} \log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)$, with probability at least $1 - \eta$,

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

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

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

Remark 1. Therefore, under the KL-divergence, to achieve an ϵ absolute error of estimating $V_{\mathcal{P}}^* = v^*$ with $\hat{v}$ in l_∞ norm w.h.p., we need a total of $\tilde{O}(|\mathcal{S}||\mathcal{A}|(1 - \gamma)^{-4}\mathfrak{p}_\wedge^{-1}\epsilon^{-2})$ samples from the simulator.

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 [4].

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

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

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

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 $v : \mathcal{S} \rightarrow \mathbb{R}$ and $n \geq 12\mathfrak{p}_\Lambda^{-1} \log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)$, with probability at least $1 - \eta$,

$$\|\hat{\mathbf{T}}^*(v) - \mathcal{T}^*(v)\|_\infty \leq \frac{3 \cdot 2^{k^*} k^* \|R + \gamma v\|_\infty}{\sqrt{n\mathfrak{p}_\Lambda}} \sqrt{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}.$$

The proof of Proposition 3 is provided in Appendix D.1. This, combined with Proposition 1, implies the following error convergence bound, whose proof is deferred to Appendix D.4.

Theorem 2. Under the f_k -divergence uncertainty set with any $\rho \in (0, \infty)$ and $n \geq 12\mathfrak{p}_\Lambda^{-1} \log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)$, with probability at least $1 - \eta$,

$$\|\hat{v} - v^*\|_\infty \leq \frac{3 \cdot 2^{k^*} k^*}{(1 - \gamma)^2 \sqrt{n\mathfrak{p}_\Lambda}} \sqrt{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)}.$$

Remark 2. Therefore, under the f_k -divergence for a fixed k , to achieve an ϵ absolute error of estimating $V_{\mathcal{P}}^* = v^*$ with $\hat{v}$ in l_∞ norm w.h.p., we also need a total of $\tilde{O}(|\mathcal{S}||\mathcal{A}|(1 - \gamma)^{-4}\mathfrak{p}_\Lambda^{-1}\epsilon^{-2})$ samples from the simulator.

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. [7], 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. [30] to illustrate the linear dependence on $|\mathcal{S}||\mathcal{A}|$, which matches the lower bound established in Yang et al. [30].

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 [7]. In each period t , an agent decides the order quantity to maximize cumulative discounted rewards, accounting for holding costs, backlog penalties, and profits.

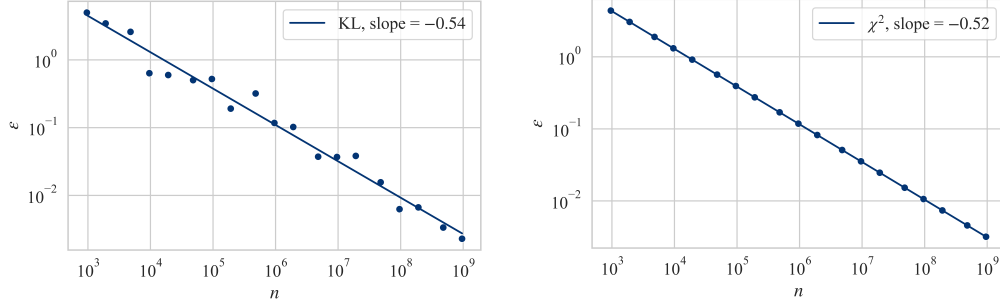
Let I denote the maximum inventory level, B the maximum backlog, and O the maximum order quantity per period. The state space is defined as $\mathcal{S} = \{-B, \dots, 0, \dots, I\}$, the action space is

290 $\mathcal{A} = \{0, \dots, O\}$, where $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ denote the inventory and order item at the beginning of
 291 period t . Demand $D_t \in \{0, \dots, D_{\max}\}$ is an i.i.d. sequence, with distribution $P_D \in \Delta^{D_{\max}+1}$.

292 The MDP dynamics proceed as follows. Due to storage constraints, the effective order size is
 293 $\tilde{A}_t = \min(A_t, I - S_t)$. Then, the next state evolves as $S_{t+1} = \max(S_t + \tilde{A}_t - D_t, -B)$, ensuring
 294 that the backlog does not exceed B . The actual sales in period t are given by $X_t = S_t - S_{t+1} + \tilde{A}_t$.
 295 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,
 296 where p is the sales price, c is the purchase cost, h is the holding cost, and b is the penalty of backlog.

297 For our experiments, we set the parameters as follows: $I = 10$, $B = 5$, $O = 5$, $p = 3$, $c = 2$,
 298 $h = 0.2$, $b = 3$, $\gamma = 0.9$, and use the nominal demand distribution $P_D = [0.1, 0.2, 0.3, 0.3, 0.1]$,
 299 supported on $0, 1, 2, 3, 4$. For each (s, a) , we sample n_0 samples from the nominal transition kernels
 300 to generate the estimated transition probability P_n , and solve the DR-MDP problem with uncertainty
 301 size $\rho = 1/6$ using the algorithm presented in [7].

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



(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.

305 5.2 MDP Instances from the Lower Bound Construction in Yang et al. [30]

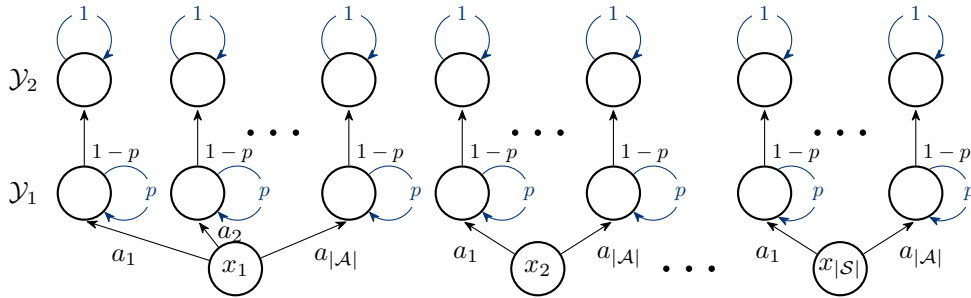


Figure 2: MDP instances from the lower bound construction in Yang et al. [30].

306 In this section, we investigate the relationship between the estimation error and the sizes of the state
 307 space $|\mathcal{S}|$ and action space $|\mathcal{A}|$. We adopt the classic MDP structure introduced in Gheshlaghi Azar
 308 et al. [5] and Yang et al. [30], which comprises three subsets: $\mathcal{S}$, $\mathcal{Y}_1$, and $\mathcal{Y}_2$, as illustrated in Figure 2.

309 Specifically, $\mathcal{S}$ denotes the set of all initial states, each associated with an action set $\mathcal{A}$. When an
 310 action $a_i \in \mathcal{A}$ is taken in state $s \in \mathcal{S}$, the system deterministically transitions (with probability 1)
 311 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
 312 state with nominal probability p , or transitions to the corresponding absorbing state $y_{2,s,a} \in \mathcal{Y}_2$
 313 with nominal probability $1 - p$. All states in $\mathcal{Y}_2$ are absorbing, meaning that once the system enters one of
 314 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 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 $|\mathcal{S}||\mathcal{A}|$.

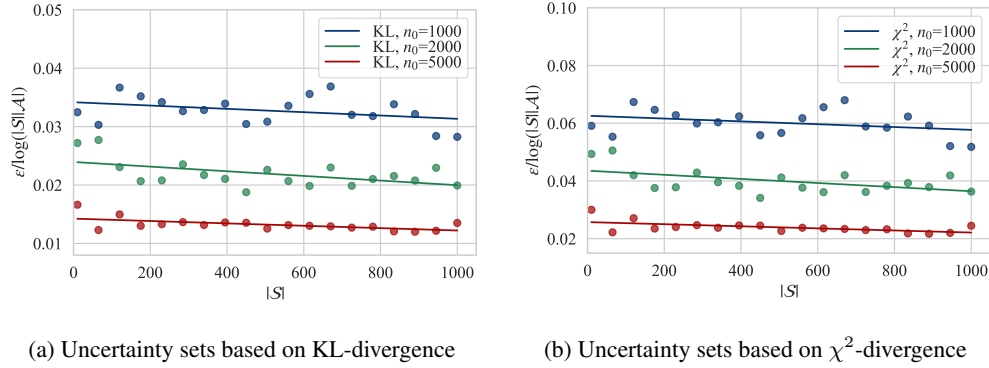


Figure 3: Estimation error versus the number of states $|\mathcal{S}|$ for the MDP instances based on the lower bound construction in Yang et al. [30].

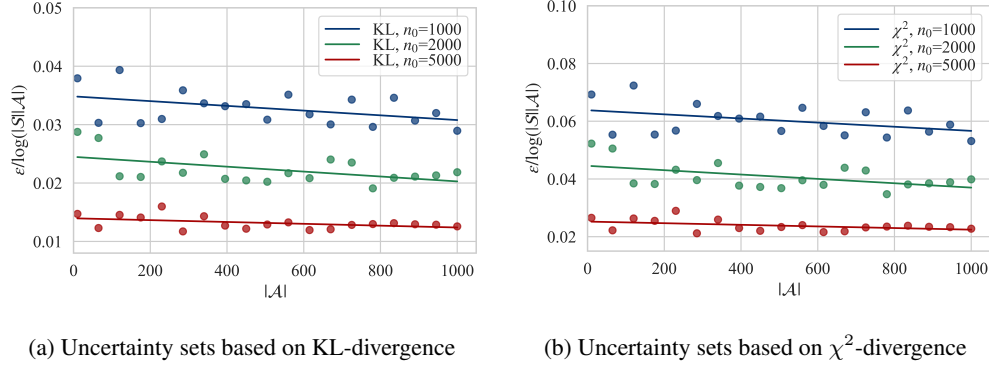


Figure 4: Estimation error versus the number of states $|\mathcal{A}|$ for the MDP instances based on the lower bound construction in Yang et al. [30].

6 Conclusion and Future Work

In this paper, we present near-optimal sample complexity results for divergence-based S-rectangular robust MDPs in the discounted reward setting. Our results are the first to achieve optimal dependence on $|\mathcal{S}|$, $|\mathcal{A}|$, and ε simultaneously. We acknowledge, however, two limitations: the reliance on access to a generative model and the presence of a gap between our upper bound and the minimax lower bound established in Yang [29]. As part of future work, we aim to develop provable theoretical guarantees for other settings, including model-free algorithms and offline reinforcement learning.

References

- [1] 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.
- [2] 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.
- [3] 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].
- [4] John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.
- [9] Yann Le Tallec. *Robust, risk-sensitive, and data-driven control of Markov decision processes*. PhD thesis, Massachusetts Institute of Technology, 2007.
- [10] 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>.
- [11] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70 (2):583–601, 2002.
- [12] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.
- [13] Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model, 2021. URL <https://arxiv.org/abs/2112.01506>.
- [14] 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.
- [15] Alexander Shapiro. Distributionally robust modeling of optimal control. *Operations Research Letters*, 50(5):561–567, 2022.
- [16] 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>.
- [17] 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].

- [18] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model, 2023.
- [19] 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.
- [20] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [21] Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust Q-learning, 2023. URL <https://arxiv.org/abs/2302.13203>.
- [22] Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust Q-learning, 2023.
- [23] 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*, 2024.
- [24] Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. On the Foundation of Distributionally Robust Reinforcement Learning, January 2024. URL <http://arxiv.org/abs/2311.09018>. arXiv:2311.09018 [cs].
- [25] Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample Complexity of Variance-reduced Distributionally Robust Q-learning, September 2024. URL <http://arxiv.org/abs/2305.18420>. arXiv:2305.18420 [cs].
- [26] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.
- [27] Huan Xu and Shie Mannor. Distributionally robust markov decision processes. In *NIPS*, pages 2505–2513, 2010.
- [28] Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning, 2023.
- [29] Insoon Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 2020.
- [30] 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.
- [31] 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.
- [32] 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, 2021.
- [33] 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 2021. URL <https://proceedings.mlr.press/v130/zhou21d.html>.

424 A Proofs of Value Function Error Bounds

425 In this section, we prove Proposition 1. We first show that both the population and the empirical
 426 S-rectangular Bellman operators $\mathcal{T}^*$ and $\hat{\mathbf{T}}^*$ are γ -contractions. This is a well-known fact, see for
 427 example [24]. We include a proof to make the paper self-contained.

428 **Lemma 3.** $\mathcal{T}^*$ and $\hat{\mathbf{T}}^*$ are γ -contraction operators on $(\mathcal{S} \rightarrow \mathbb{R}, \|\cdot\|_\infty)$; i.e. for all $v_1, v_2 : \mathcal{S} \rightarrow \mathbb{R}$,

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

429 *Proof.* Let

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

430 By definition, we have

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

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

$$\begin{aligned}&|\mathcal{T}^*(v_1)(s) - \mathcal{T}^*(v_2)(s)| \\ &\leq \sup_{\pi \in \Pi} \sup_{P \in \mathcal{P}} |f(v_1)(s) - f(v_2)(s)| \\ &= \sup_{\pi \in \Pi} \sup_{P \in \mathcal{P}} \left| \sum_{a,s'} \pi(a|s) P_{s,a}(s') R(s, a, s') + \gamma \sum_{a,s'} \pi(a|s) P_{s,a}(s') v_1(s') \right. \\ &\quad \left. - \sum_{a,s'} \pi(a|s) P_{s,a}(s') R(s, a, s') - \gamma \sum_{a,s'} \pi(a|s) P_{s,a}(s') v_2(s') \right| \\ &= \sup_{\pi \in \Pi} \sup_{P \in \mathcal{P}} \gamma \left| \sum_{a,s'} \pi(a|s) P_{s,a}(s') v_1(s') - \sum_{a,s'} \pi(a|s) P_{s,a}(s') v_2(s') \right| \\ &\leq \gamma \|\pi_s\|_\infty \|P_s\|_\infty \|v_1 - v_2\|_\infty \\ &= \gamma \|v_1 - v_2\|_\infty\end{aligned}$$

432 where $\|P_s\|_\infty = \sup_{\|v\|_\infty=1} \|P_s v\|_\infty$ is the induced operator norm. The above inequality holds for
 433 any $s \in \mathcal{S}$, which lead to

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

434 We replace P_s with $P_{s,n}$, notice that $\|P_{s,n}\|_\infty \leq 1$

$$\begin{aligned}|\hat{\mathbf{T}}^*(v_1)(s) - \hat{\mathbf{T}}^*(v_2)(s)| &\leq \gamma \|\pi_s\|_\infty \|P_{s,n}\|_\infty \|v_1 - v_2\|_\infty \\ &= \gamma \|v_1 - v_2\|_\infty.\end{aligned}$$

435 which lead to

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

436 □

437 A.1 Proof of Proposition 1

438 *Proof.* The proof of Proposition 1 follows a similar argument to that used for the continuous-case
 439 operator in [23].

440 Let $v_0 \equiv 0$ and $v_{k+1} = \hat{\mathbf{T}}^*(v_k)$. $\hat{v}$ is defined as the fix point of $\hat{\mathbf{T}}^* \hat{v} = \hat{\mathbf{T}}^*(\hat{v})$

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

441 By Lemma 3, we have

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

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

$$\|\Delta_k\|_\infty \leq \frac{\gamma^{k-1}}{1-\gamma} + \sum_{j=0}^{k-1} \gamma^j \|V\|_\infty$$

443 by induction: for $k = 1$,

$$\begin{aligned}\|\Delta_1\|_\infty &\leq \|\mathbf{H}(\Delta_0)\|_\infty + \|V\|_\infty \\ &= \|\mathbf{H}(\Delta_0) - \mathbf{H}(0)\|_\infty + \|V\|_\infty \\ &\leq \gamma \|v^*\|_\infty + \|V\|_\infty \\ &\leq \frac{\gamma}{1-\gamma} + \|V\|_\infty.\end{aligned}$$

444 For any k , we have

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

445 Therefore,

$$\|\hat{v} - v^*\|_\infty = \lim_{k \rightarrow \infty} \|\Delta_k\|_\infty \leq \sum_{j=0}^{\infty} \gamma^j \|V\|_\infty = \frac{1}{1-\gamma} \left\| \hat{\mathbf{T}}^*(v^*) - \mathcal{T}^*(v^*) \right\|.$$

446 □

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

448 The proofs for all f -divergence-based uncertainty sets follow a unified framework. We first present
449 Lemma 4, which gives a general dual formulation for any convex f -divergence. For the KL-
450 divergence and the f_k -divergence, we specialise this result by substituting the corresponding conjugate
451 functions f^* . The detailed derivations for the KL-divergence and the f_k -divergence are provided in
452 Appendix B.1 and B.2, respectively.

453 **Lemma 4.** *For any f -divergence uncertainty set, where $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ is a convex function and
454 $f(0) = 1$ and satisfies $f(0) = \lim_{t \downarrow 0} f(t)$, the convex optimization problem*

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

455 can be reformulated as:

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

456 where $f^*(t) = -\inf_{s \geq 0} (f(s) - st)$.

457 *Proof.* We follow the proof of Lemma 8.5 in [30], however, in our case, R is determined by the next
 458 state. We do a change of variables, let $L_{s,a}(s') = \frac{P_{s,a}(s')}{P_{s,a}(s)}$. The original optimization problem can be
 459 reformulated as:

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

460 The Lagrange function of primal problem is

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

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

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

462

□

463 B.1 Proof of Lemma 1

464 *Proof.* Recall that for the KL-divergence, $f(t) = t \log t$, whose conjugate function $f^*(s) = e^{s-1}$.
 465 Substituting f^* into Lemma 4, we obtain the following dual form:

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

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

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

471 Solving for η_a , we obtain

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

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

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

□

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 5 (Duchi and Namkoong [4], Section 2). *Recall that in f_k -divergence,*

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

The conjugate function $f_k^*(s) = \sup_{t \geq 0} (st - f_k(t))$ 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 4, and let $w_{s,a}(S) := \pi(a|s) (R(s, a, S) + \gamma v(S))$, we obtain

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

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 $(\cdot)_+$.

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

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

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

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:

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

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

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

Therefore, we obtain

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

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

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

490

□

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

492 Our techniques in this section refine that in Wang et al. [25]. To follow the constructions in Wang
493 et al. [25], we introduce some notations. Consider $\mu_{s,a} \in \Delta(\mathcal{S})$ and its empirical version $\mu_{s,a,n}$
494 constructed from n i.i.d samples from $\mu_{s,a}$. Define the collection of these measures under state s as
495 $\boldsymbol{\mu}_s := \{\mu_{s,a} : a \in \mathcal{A}\}$. For a function $u : \mathcal{S} \rightarrow \mathbb{R}$ and for each $s \in \mathcal{S}$, we define:

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

496 For the supremum over all states, we define

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

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

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

500 Further, the good event over all states is defined as

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

501 For notation simplicity, we suppress the dependence on the state variable s . Consider a function
502 $u : \mathcal{S} \rightarrow \mathbb{R}$. The dual function under KL-divergence is given by:

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

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

504 We define the deviation between empirical and true measures as

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

505 and their convex interpolation by

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

506 **C.1 Proof of Proposition 2**

507 *Proof.* By definition and $|\sup_X f - \sup_X g| \leq \sup_X |f - g|$, we have

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

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

$$\begin{aligned} \left|\hat{\mathbf{T}}^\pi(v)(s) - \mathcal{T}^\pi(v)(s)\right| &\leq \left|\sup_{\lambda > 0} f(P_{s,n}, R(s, \cdot, \cdot) + \gamma v, \lambda) - \sup_{\lambda > 0} f(P_s, R(s, \cdot, \cdot) + v, \lambda)\right| \\ &\leq \sup_{\lambda > 0} |f(P_{s,n}, R(s, \cdot, \cdot) + \gamma v, \lambda) - f(P_s, R(s, \cdot, \cdot) + v, \lambda)|. \end{aligned} \quad (\text{C.2})$$

509 We analyze the sensitivity of the mapping $\mu \rightarrow f(\mu, u, \lambda)$. For any fixed u, μ and μ_n , define

$$g_n(t, \lambda) = f(\mu_n(t), u, \lambda).$$

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

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

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

512 **Lemma 6.** *For any fixed u and π , $\mu_n \ll \mu$, we have that*

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

513

514 The proof is deferred to Appendix C.2. According to lemma 6, we have

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

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

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

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

519 **Lemma 7.** *For any s' with $\mu(s') > 0$, the measure $\mu_n(t)(s')$ satisfies*

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

520

521 The proof is deferred to Appendix C.3. By using lemma 7, we have $\mu_{a,n}(\tau)(s') \geq (1 - \delta_0)\mu_a(s')$,
 522 therefore,

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

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

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

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

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

526 Recall from (C.2) that

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

527 Replacing μ with P_s and μ_n with $P_{s,n}$ and choose $\delta_0 = \frac{1}{2}$, by union bound, we have

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

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

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

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

529 Under (C.3), we have

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

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

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

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

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

we have

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

This implies Proposition 2. $\square$

C.2 Proof of Lemma 6

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 $e^x - 1 > xe^x$, then we obtain

$$\begin{aligned} &\leq \sum_{a \in \mathcal{A}} \left\|\lambda \frac{e^{d_a(\|u\|_{L^\infty(\mu_a)} - u)} - 1}{e^{d_a(\|u\|_{L^\infty(\mu_a)} - u)/\lambda}}\right\|_{L^\infty(\mu_a)} \left\|\frac{dm_{a,n}}{d\mu_{a,n}(t)}\right\|_{L^\infty(\mu_a)} \\ &\leq \sum_{a \in \mathcal{A}} \|d_a(\|u\|_{L^\infty(\mu_a)} - u)\|_{L^\infty(\mu_a)} \left\|\frac{dm_{a,n}}{d\mu_{a,n}(t)}\right\|_{L^\infty(\mu_a)} \\ &\leq \sum_{a \in \mathcal{A}} 2d_a\|u\|_\infty \left\|\frac{dm_{a,n}}{d\mu_{a,n}(t)}\right\|_{L^\infty(\mu_a)} \\ &\leq 2\|u\|_\infty \left\|\frac{dm_n}{d\mu_n(t)}\right\|_{\infty, \mu} \end{aligned}$$

as claimed. $\square$

542 **C.3 Proof of Lemma 7**

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

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

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

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

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

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

547 □

548 **C.4 Proof of Theorem 1**

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

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

551 with probability $1 - \eta$. □

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

554 **D.1 Proof of Proposition 3**

555 *Proof.* Let

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

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

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

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

560 **Lemma 8.** For any fixed u and π ,

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

561 where $c = 2^{1/(k-1)} k^*$.

562 We decomposed the probability using the event $\Omega_{n,\delta_0}(\mu)$ where the empirical estimates are close to
 563 the population measures. Let $c = 2^{1/(k-1)}k^*$, and by using lemma 8, we obtain

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

564 Again using Lemma 7, we have

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

565 By Chernoff Bound and Bernstein Inequality, we obtain

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

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

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

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

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

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

$$n \geq \frac{12}{\mathfrak{p}_\wedge} \log(4|\mathcal{S}|^2 |\mathcal{A}|/\eta) \quad (\text{D.1})$$

$$t \geq \frac{4c \|R + \gamma v\|_\infty}{3n \mathfrak{p}_\wedge} \log(4|\mathcal{S}|^2 |\mathcal{A}|/\eta) + \frac{2c \|R + \gamma v\|_\infty}{\sqrt{n \mathfrak{p}_\wedge}} \sqrt{2 \log(4|\mathcal{S}|^2 |\mathcal{A}|/\eta)}. \quad (\text{D.2})$$

570 Under (D.1), we have

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

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

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

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

$$t \geq \frac{3 \cdot 2^{k^*} k^* \|R + \gamma v\|_\infty}{\sqrt{n \mathfrak{p}_\wedge}} \sqrt{\log(4|\mathcal{S}|^2|\mathcal{A}|/\eta)},$$

we have

$$P\left(\left\|\hat{\mathbf{T}}^*(v) - \mathcal{T}^*(v)\right\|_\infty > t\right) \leq \eta,$$

which implies the statement of the proposition. $\square$

D.2 Proof of Lemma 8

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

$$\begin{aligned} X_1 &= \left\{ \boldsymbol{\eta} \mid \eta_a \leq d_a \operatorname{essinf}_{\mu_a} u \text{ for all } a \in \mathcal{A} \right\}, \\ X_2 &= \left\{ \boldsymbol{\eta} \mid \eta_a > d_a \operatorname{essinf}_{\mu_a} u \text{ for all } a \in \mathcal{A} \right\}, \\ X_3 &= \mathbb{R}^{|\mathcal{A}|} \setminus \{X_1 \cup X_2\}. \end{aligned}$$

Next we prove that $X_3 = \emptyset$. If $\boldsymbol{\eta}$ is an optimal solution, then it satisfies the conditions described in the following lemma.

Lemma 9. *Let $\boldsymbol{\eta}^*(\boldsymbol{\mu})$ denote the optimal $\boldsymbol{\eta}$ under measure $\boldsymbol{\mu}$, then we have*

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

and when $\boldsymbol{\eta} \in X_2$, we have

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

The proof is deferred to Appendix D.3. Suppose $\boldsymbol{\eta} \in X_3$. Then, there exists some $a' \in \mathcal{A}$ such that $\eta_{a'} \leq d_{a'} \operatorname{essinf}_{\mu_{a'}} u$, implying that $\mu_{a'} [w_{a'}^{1/(k-1)}] = 0$. According to (D.3), this leads to $\mu_a [w_a^{k^*}] = 0$ for all $a \in \mathcal{A}$, which means $\boldsymbol{\eta} \in X_1$, contradicting the initial assumption. Hence, $X_3 = \emptyset$.

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

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

Otherwise, $\boldsymbol{\eta} \in X_2$, for any fixed $\boldsymbol{\mu}, \boldsymbol{\mu}_n$ and u , let

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

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 10 (Envelope theorem, [11], Corollary 3). *Denote V as*

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

Suppose that X is a convex set in a linear space and $f : X \times [0, 1] \rightarrow \mathbb{R}$ is a concave function. Also suppose that $t_0 \in (0, 1)$, and that there is some $\mathbf{x}^ \in X^*(t_0)$ such that $d_t f(\mathbf{x}^*, t_0)$ exists. Then V is differentiable at t_0 and $d_t V(t_0) = \partial_t f(\mathbf{x}^*, t_0)$*

We examine the convexity of X and the concavity of g . X_2 is a convex set since it is defined by linear inequalities for each coordinate. For g , since f serves as the dual objective function and is therefore concave, and concavity is preserved under affine mappings. So given that $\mu_n(t)$ is a linear function of t , g inherits the concavity. Therefore according to Lemma 10, $V(t)$ is differentiable. By mean value theorem, there exists $\tau \in (0, 1)$, for which the following equation holds:

$$\begin{aligned} \left| \sup_{\eta \in X_2} f(\mu_n, u, \eta) - \sup_{\eta \in X_2} f(\mu, u, \eta) \right| &= \left| \sup_{\eta \in X_2} g(\eta, 0) - \sup_{\eta \in X_2} g(\eta, 1) \right| \\ &= \left| \frac{d}{dt} V(t) \Big|_{t=\tau} \right|, \end{aligned}$$

and by envelope theorem, we have

$$\frac{d}{dt} V(t) \Big|_{t=\tau} = \frac{\partial}{\partial t} g(\eta^*, t) \Big|_{t=\tau}.$$

Recall that

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

by using (D.3), we obtain

$$\begin{aligned} \frac{\partial}{\partial t} g(\eta^*, t) &= - \frac{c(k, \rho, |\mathcal{A}|)}{(\sum_{a \in \mathcal{A}} \mu_{a,n}(t) [w_a^{k*}])^{1/k}} \sum_{a \in \mathcal{A}} m_{a,n} [w_a^{k*}] \\ &= - \frac{\sum_{a \in \mathcal{A}} m_{a,n} [w_a^{k*}]}{\mu_{i,n}(t) [w_i^{1/(k-1)}]}. \end{aligned}$$

Therefore,

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

Since the equation above holds for any $i \in \mathcal{A}$, chose $i = a$ for each $a \in \mathcal{A}$, so we can rewrite the equation above as

$$\begin{aligned} \left| \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} f(\mu_n, u, \eta) - \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} f(\mu, u, \eta) \right| &= \left| \sum_{a \in \mathcal{A}} \frac{m_{a,n} [w_a^{k*}]}{\mu_{a,n}(\tau) [w_a^{1/(k-1)}]} \right| \\ &\leq \sum_{a \in \mathcal{A}} \left| \frac{m_{a,n} [w_a^{k*}]}{\mu_{a,n}(\tau) [w_a^{1/(k-1)}]} \right|. \end{aligned}$$

For each term in the summation, we analyze $\eta_a \geq 2d_a \|u\|_{L^\infty(\mu_a)}$ and $d_a \text{essinf}_{\mu_a} u \leq \eta_a < 2d_a \|u\|_{L^\infty(\mu_a)}$ separately. For $\eta_a \geq 2d_a \|u\|_{L^\infty(\mu_a)}$, by mean value theorem, there exists $\xi \in (\eta_a - d_a u, \eta_a)$ satisfies

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

609 Since $\eta_a > 2d_a\|u\|_{L^\infty(\mu_a)}$, $\xi < \eta_a \leq 2(\eta_a - d_a u)$, then we have

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

610 For $d_a \operatorname{essinf}_{\mu_a} u \leq \eta_a < 2d_a\|u\|_{L^\infty(\mu_a)}$, $(\eta_a - d_a u)_+$ is bounded, then we have

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

611 To sum up, we have

$$\begin{aligned}
& \sum_{a \in \mathcal{A}} \left| \frac{m_{a,n} \left[w_a^{k*} \right]}{\mu_{a,n}(\tau) \left[w_a^{1/(k-1)} \right]} \right| \\
& \leq \sum_{a \in \mathcal{A}} \max \left\{ \sup_{\eta_a \geq 2d_a\|u\|_{L^\infty(\mu_a)}} \left| \frac{m_{a,n} \left[w_a^{k*} \right]}{\mu_{a,n}(\tau) \left[w_a^{1/(k-1)} \right]} \right|, \sup_{\substack{d_a \operatorname{essinf}_{\mu_a} u \leq \eta_a \\ \leq 2d_a\|u\|_{L^\infty(\mu_a)}}} \left| \frac{m_{a,n} \left[w_a^{k*} \right]}{\mu_{a,n}(\tau) \left[w_a^{1/(k-1)} \right]} \right| \right\} \\
& = \sum_{a \in \mathcal{A}} \max \left\{ 2^{1/(k-1)} k^*, 2 \right\} d_a \|u\|_{L^\infty(\mu_a)} \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{L^\infty(\mu_a)} \\
& = 2^{1/(k-1)} k^* \|u\|_\infty \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{\infty, \mu}.
\end{aligned}$$

612 Overall, let $c = 2^{1/(k-1)} k^*$, when $\mu \in X_1$,

$$\left| \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} f(\mu_n, u, \eta) - \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} f(\mu, u, \eta) \right| = 0 \leq c \|u\|_\infty \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{\infty, \mu},$$

613 when $\mu \in X_2$,

$$\left| \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} f(\mu_n, u, \eta) - \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} f(\mu, u, \eta) \right| \leq \sum_{a \in \mathcal{A}} \left| \frac{m_{a,n} \left[w_a^{k*} \right]}{\mu_{a,n}(\tau) \left[w_a^{1/(k-1)} \right]} \right| \leq c \|u\|_\infty \left\| \frac{dm_n}{d\mu_n(\tau)} \right\|_{\infty, \mu}.$$

614

□

615 D.3 Proof of Lemma 9

616 Since $f(\boldsymbol{\mu}, u, \boldsymbol{\eta}^*)$ is the objective function of the dual problem and $\boldsymbol{\eta}$ is the dual variable, f is convex
 617 with respect to $\boldsymbol{\eta}$. To optimize f over $\boldsymbol{\eta}$, we set its derivative with respect to η_i to zero, which yields

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

618 which means

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

619 which is (D.3). When $\boldsymbol{\eta} \in X_2$, $\mu_i [w_i^{1/(k-1)}]$ is positive, plug in η_a^* , we obtain

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

620 D.4 Proof of Theorem 2

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

$$\begin{aligned} \|\hat{v} - v^*\|_\infty &\leq \frac{1}{1 - \gamma} \left\| \hat{\mathbf{T}}^*(v^*) - \mathcal{T}^*(v^*) \right\|_\infty \\ &\leq \frac{3 \cdot 2^{k^*} k^* \|R + \gamma v\|_\infty}{(1 - \gamma) \sqrt{n \mathfrak{p}_\wedge}} \sqrt{\log(4|\mathcal{S}|^2 |\mathcal{A}|/\eta)} \\ &\leq \frac{3 \cdot 2^{k^*} k^*}{(1 - \gamma)^2 \sqrt{n \mathfrak{p}_\wedge}} \sqrt{\log(4|\mathcal{S}|^2 |\mathcal{A}|/\eta)} \end{aligned}$$

623 with probability $1 - \eta$. □

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The claims regarding sample complexity and empirical validation in the abstract and introduction are fully supported by the theoretical analyses and experiments presented in the main text.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss the limitations in the Conclusion and Future Work section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We provided the full set of assumptions in Section 3 and Section 4. The complete proofs are provided in the Appendix. All theorems, lemmas, and their assumptions are clearly stated and properly referenced.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We introduced the problem settings (inventory and MDP instances we used) and provided all the hyperparameters in the Numerical Experiment section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code required to reproduce our main results are provided in the supplementary material. Detailed instructions for running the code are included as well.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the hyperparameters in detail. Our experiments do not involve data splitting, optimizer selection, or related settings, as they are not applicable in our setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: All experiments were conducted only once under a single set of hyperparameters. Therefore, we do not report error bars or measures of statistical significance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: All experiments in this paper were run on standard CPU environments, so we did not provide detailed computing resource specifications in the paper. We believe that this information is of limited impact on the reproducibility of our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our paper does not involve human subjects or use any datasets. There are no ethical, societal, or environmental concerns. Therefore, the research fully complies with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work concerns the theoretical analysis of sample complexity and does not involve any particular application, data, or deployment. Therefore, it has no direct societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper does not use any real-world datasets or pretrained models, and therefore does not pose such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Our paper only uses standard, widely adopted open-source libraries (e.g., Numpy, Pyomo, Ipopt) for implementation and experimentation. No external datasets, pre-trained models, or code assets requiring explicit license attribution are used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new datasets, code assets, or models are introduced or released in our paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

936 Justification: Our paper does not involve crowdsourcing nor research with human subjects.

937 Guidelines:

- 938 • The answer NA means that the paper does not involve crowdsourcing nor research with
- 939 human subjects.
- 940 • Depending on the country in which research is conducted, IRB approval (or equivalent)
- 941 may be required for any human subjects research. If you obtained IRB approval, you
- 942 should clearly state this in the paper.
- 943 • We recognize that the procedures for this may vary significantly between institutions
- 944 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
- 945 guidelines for their institution.
- 946 • For initial submissions, do not include any information that would break anonymity (if
- 947 applicable), such as the institution conducting the review.

948 16. Declaration of LLM usage

949 Question: Does the paper describe the usage of LLMs if it is an important, original, or

950 non-standard component of the core methods in this research? Note that if the LLM is used

951 only for writing, editing, or formatting purposes and does not impact the core methodology,

952 scientific rigorousness, or originality of the research, declaration is not required.

953 Answer: [NA]

954 Justification: The core method development in our research does not involve LLMs as any

955 important, original, or non-standard components.

956 Guidelines:

- 957 • The answer NA means that the core method development in this research does not
- 958 involve LLMs as any important, original, or non-standard components.
- 959 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
- 960 for what should or should not be described.