Statistical Learning of Distributionally Robust Stochastic Control in Continuous State Spaces

Shengbo Wang
Stanford University

Nian Si
HKUST

Jose Blanchet
Stanford University

Zhengyuan Zhou
New York University

Abstract

We explore the control of stochastic systems with potentially continuous state and action spaces, characterized by the state dynamics $X_{t+1} = f(X_t, A_t, W_t)$. Here, X, A, and W represent the state, action, and exogenous random noise processes, respectively, with f denoting a known function that describes state transitions. Traditionally, the noise process $\{W_t, t \geq 0\}$ is assumed to be independent and identically distributed, with a distribution that is either fully known or can be consistently estimated. However, the occurrence of distributional shifts, typical in engineering settings, necessitates the consideration of the robustness of the policy. This paper introduces a distributionally robust stochastic control paradigm that accommodates possibly adaptive adversarial perturbation to the noise distribution within a prescribed ambiguity set. We examine two adversary models: current-action-aware and current-action-unaware, leading to different dynamic programming equations. Furthermore, we characterize the optimal finite sample minimax rates for achieving uniform learning of the robust value function across continuum states under both adversary types, considering ambiguity sets defined by f_k-divergence and Wasserstein distance. Finally, we demonstrate the applicability of our framework across various real-world settings.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 Introduction

Stochastic control formulations are extensively utilized in the modeling, design, and optimization of systems influenced by probabilistic dynamics. These formulations play a crucial role across various fields within operations research and management disciplines. Notable applications of stochastic control can be seen in finance (Merton, 1976), communication systems (Yüksel and Başar, 2013), manufacturing and operations management (Tse and Grossglauser, 1997; Porteus, 2002), as well as energy systems (Foschini and Miljanic, 1993). A key aspect common to these applications is the use of a continuous state space, which provides a robust and flexible environment for formulating complex system dynamics. This facilitates the development of realistic dynamic models that accurately reflect the underlying systems, thereby enhancing system management and operational efficiency.

The underlying state dynamics of a large class of stochastic control problems can be described by the following recursion

$$X_{t+1} = f(X_t, A_t, W_t). \tag{1.1}$$

Here, X_t is the state of the system at time t, and $\{W_t : t \geq 1\}$ is assumed to be a sequence of random variables that are independent and identically distributed (i.i.d.), representing the exogenous randomness that underlies the stochasticity of the system. The action A_t taken at time t is based on the information that the controller has accumulated up to that point. Then, a reward $r(X_t, A_t)$ is realized. The goal is to maximize the cumulative infinite horizon α-discounted reward, for some $\alpha \in (0, 1)$. In this setting, it is well-known that Markov policies, which base decisions solely on the current state X_t, are optimal. Furthermore, the dynamic programming principle characterizes this optimality through the Bellman equation, describing the optimal value function and an associated optimal Markov policy.

The Bellman equation can be equivalently expressed using transition probabilities, rather than relying on

the specific function f. This is the standard formulation in the theory of Markov Decision Processes (MDPs). Although stochastic control and MDP formulations are equivalent in terms of modeling expressiveness and optimization, assuming a known form of f, as we have in this paper, presents significant advantages for statistical learning when the distribution of $\{W_t : t \geq 0\}$ is unknown. Specifically, this allows for simultaneous learning across all states when the random variables W_t are observed. Fortunately, as listed above, there is a wide range of learning and dynamic decision-making settings for which the stochastic control formulation is natural. This is especially the case for environments with continuous state spaces, aligning with the objective of this paper. The following overviews of examples will illustrate scenarios where these key features are present. The detailed versions are presented in Section 6.

Example 1 (Portfolio management)**.** We consider managing a portfolio of m assets. $X_t \in \mathbb{R}^m$ denotes the portfolio at time t, where $X_{t,i}$ is the dollar value of asset i at the beginning of time t. We can buy and sell assets at the beginning of each time period. Let $A_t \in \mathbb{R}^m$ be our decision variables at t, representing the dollar values of the trades. The state dynamics is $X_{t+1} = W_t(X_t + A_t)$, where $W_t = \text{diag}(R_t) \in \mathbb{R}^{m\times m}$ is the a diagonal matrix of asset returns $R_t \in \mathbb{R}^m$. Here, $R_{t,i}$ represents the return of the i-th asset from period t to period $t+1$.

Example 2 (Service and manufacturing systems)**.** We consider a simple service and manufacturing system with make-to-order queues. Let X_n denote the waiting time of the n-th job, W_n the inter-arrival time between the $(n+1)$-th and the n-th job, and assume each job requires 1 unit of work. Let A_n be the service rate chosen by the system manager for the n-th job. Then, the system dynamics can be written as $X_{n+1} = (X_n + 1/A_n - W_n)^+$.

However, in practice, the i.i.d. assumptions for the joint distributions of $\{W_t\}$ will often be violated. Challenges such as a correlated noise process, along with the presence of confounders and non-stationarities in the environment, can significantly deteriorate policy performance in real deployment environments. These issues serve as the motivation for adopting a distributionally robust stochastic control (DRSC) formulation. The DRSC framework promotes policy robustness by setting up a dynamic game where the controller selects an action at each time t, while another entity–the adversary–perturbs the distribution of W_t. The introduction of the adversary serves as a strategic device that quantifies the worst-case risk associated with model misspecifications. Although this formulation is natural to dynamic robust decision-making, the learning aspect within DRSC, particularly in continuous state space settings, has not yet been studied. To our knowledge, this paper presents the first optimal sample complexity results for learning in such settings.[1]

Unlike the robust control literature, which often leads to deterministic optimal controls (González-Trejo et al., 2002), our study explores two distinct DRSC formulations where the adversary can be either *current-action-aware* (CAA) or *current-action-unaware* (CAU). The key difference between these models lies in whether the adversary has access to the controller's realized action when deciding how to perturb W_t. The presence of deterministic optimal Markov controls and a corresponding Bellman equation is not guaranteed to hold under these conditions, especially when asymmetric information structures are present (Wang et al., 2023b). Specifically, DRSC problems with CAU adversaries necessitate randomized policies to achieve optimal control. While these issues have been explored within the distributionally robust MDP (DRMDP) context (Wiesemann et al., 2013; Wang et al., 2023b), there has been limited research focusing on differentiating between CAA and CAU formulations in robust control settings, or on their implications for learning. Nevertheless, recognizing these distinctions is crucial from modeling, learning, and optimization perspectives. CAU formulations typically result in less powerful adversaries, leading to less conservative controls, although they may be challenging to learn and optimize. Conversely, the more conservative CAA formulations may be more appropriate models for highly competitive environments. These modeling considerations are exemplified and discussed in Section 6.

Our formulation contribution extends to establishing the existence of dynamic programming principles, expressed as DR Bellman equations, for both CAA and CAU adversaries, which guarantee the optimality of stationary Markov policies in the discounted infinite horizon setting. Although this contribution is fundamental, it is presented in the supplemental materials due to space constraints. Establishing this dynamic programming principle lays the groundwork for the other main focus of our paper: the statistical complexity of learning the DRSC values, which equal the solution of the corresponding Bellman equations, across a continuous state space.

[1] It is important to note that the field of Distributionally Robust Reinforcement Learning (DRRL), which is closely related, is quite active. While discussing the differences, it is crucial to highlight two key distinctions: firstly, the function f is known in our context, and secondly, much of the DRRL literature focuses on discrete state-action spaces.

Table 1: Summary of our results on the statistical complexity of achieving minimax learning of the DRSC value function in the uniform norm. The $\widetilde{\Theta}$ suppress a gap of $\sqrt{\log n}$.

Ambiguity Set	Type	Action	Complexity
Wasserstein	CAA CAU	Continuum Finite	$\Theta\left(n^{-1/2}\right)$
f_k-divergence	CAA CAU	Continuum Finite	$\widetilde{\Theta}\left(n^{-\frac{1}{k'\vee 2}}\right)$

Our results on learning complexities involve two types of admissible adversarial decisions affecting the distribution of $\{W_t : t \geq 0\}$, characterized by Wasserstein distance and f_k-divergence ambiguity sets. These represent the main types of distributional ambiguity in the field of distributionally robust optimization. Wasserstein ambiguity sets effectively capture model errors at the outcome level, while f_k-divergence sets hedge against misspecifications in the likelihood of possible outcomes. In Table 1, we provide a summary of the sample complexity results for each of the four cases studied in the paper. The cases are based on the visibility of the current action to the adversary (CAA vs. CAU) and the type of ambiguity sets (Wasserstein distance vs. f_k-divergence).

1.1 Literature Review

Distributionally robust stochastic control is not a new concept in the literature. Yang (2021) investigate a setting aligned with our current-action-aware formulation, employing a Wasserstein uncertainty set. In contrast, Petersen et al. (2000) consider an uncertainty set based on Kullback–Leibler divergence. For linear systems, where f is linear on X_t,A_t, and W_t, distributionally robust stochastic control has been explored by several authors (Taskesen et al., 2024; Kim and Yang, 2023; Han, 2023; Kotsalis et al., 2021). Nonetheless, existing research predominantly focuses on characterizations of the optimal policies or the development of tractable optimization methods. In our study, we address the statistical complexity associated with the learning problem.

Our work is also closely related to the literature on DRMDPs and distributionally robust reinforcement learning (DRRL). Various formulations are explored in Iyengar (2005); Nilim and El Ghaoui (2005); Le Tallec (2007); Xu and Mannor (2010); Wiesemann et al. (2013); Wang et al. (2023b); Goyal and Grand-Clément (2023); Li and Shapiro (2023). Concurrent with this manuscript, Shapiro and Li (2024) also explores DRSC. However, the focus of the two papers differs significantly: whereas Shapiro and Li (2024) emphasizes formulation aspects of DRSC, our work formulates and investigates statistical properties associated with optimal policy learning within a DRSC framework.

Statistical complexities for associated DRRL problems with finite state and action spaces are subsequently developed in Zhou et al. (2021); Panaganti and Kalathil (2021); Yang et al. (2021); Shi and Chi (2022); Xu et al. (2023); Shi et al. (2023); Blanchet et al. (2024); Liu et al. (2022); Wang et al. (2023a,c); Yang et al. (2023).

However, the picture is very different in the continuous state space setting. If one assumes that only trajectories or a generative model is available, it is known in the literature that achieving efficient (in a minimax sense) RL in a continuous state space without strong restrictions is impossible Chen and Jiang (2019) and this challenge also applies to DRRL. Our formulation of DRSC differs from DRRL as we assume a known state recursion form f driven by an unknown random variable W, whereas in DRRL, the full transition probabilities need to be learned. This difference leads to a significant difference in sample complexities. We derive minimax optimal sample complexities for uniformly estimating the robust value function with parametric convergence rates even in continuous state and action spaces.

Remarks on Paper Organizations. The remainder of this paper is structured as follows: Section 2 outlines the formulations of the DRSC problems and the corresponding dynamic programming principles, with a focus on both CAA and CAU adversaries. Sections 3 and 4 present the upper and lower bounds of the sample complexities, respectively. In Section 5, we present our algorithm design, leveraging function approximations. Finally, Section 6 presents two applications within our framework to demonstrate its effectiveness in modeling real-world problems.

2 Dynamic Programming and Bellman Equation

In this section, we provide a minimal self-consistent introduction to our formulation of the distributionally robust stochastic optimal control problem and its corresponding dynamic programming theory. The fully rigorous construction is provided in Appendix A. Our focus here is on the infinite horizon discounted reward setting. It should be noted that a DRSC formulation for finite horizon systems naturally arises from the same principles.

We consider Polish (i.e. complete separable metric spaces) state, action, and noise measurable spaces $(\mathbb{X}, \mathcal{X}), (\mathbb{A}, \mathcal{A}), (\mathbb{W}, \mathcal{W})$ equipped with the Borel σ-fields generated by open sets. Let $\mathcal{P}(\mathcal{W})$ and $\mathcal{P}(\mathcal{A})$ denote the set of probability measures on the action and noise spaces, endowed with the topology of weak convergence. We also consider a known state dynamic function $f : \mathbb{X} \times \mathbb{A} \times \mathbb{W} \to \mathbb{X}$ given in (1.1), a known reward function $r : \mathbb{X} \times \mathbb{A} \to \mathbb{R}_+$, and a discount factor $\alpha \in (0, 1)$.

Let $\mathcal{Q} \subset \mathcal{P}(\mathcal{A})$ and $\mathcal{P} \subset \mathcal{P}(\mathcal{W})$ be arbitrary Borel measurable subsets. Here, $\mathcal{Q}$ and $\mathcal{P}$ represent the admissible decision sets of the controller and the adversary, respectively. Based on $\mathcal{Q}$ and $\mathcal{P}$, we construct admissible policy classes $\Pi(\mathcal{Q})$ and $\Gamma(\mathcal{P})$ of the controller and adversary, respectively. As we will rigorously develop in the Appendix A, the admissible control policy class $\Pi(\mathcal{Q})$ of the controller will always be a proper subset of the history-dependent $\mathcal{Q}$-constrained policy class:

$$\Pi(\mathcal{Q}) \subset \Pi_{\mathrm{H}}(\mathcal{Q}) := \{\pi = (\pi_0, \pi_1, \cdots) : \pi_t(da|h_t) \in \mathcal{Q}, \\ \forall h_t = (x_0, a_0, \cdots, a_{t-1}, x_t)\}.$$

Intuitively, the controller decides a sequence of the conditional distribution of current action A_t given the history until the last visible state x_t.

Similarly, adversary's admissible policy class $\Gamma(\mathcal{P})$ is a subset

$$\Gamma(\mathcal{P}) \subset \Gamma_{\mathrm{H}} := \{\gamma = (\gamma_0, \gamma_1, \cdots) : \gamma_t(dw|g_t) \in \mathcal{P}, \\ \forall g_t = (x_0, a_0, \cdots, x_t, a_t)\}.$$

The adversarial policy $\pi \in \Gamma(\mathcal{P})$ determines the conditional distribution of W_t given the history until the last visible state action pair x_t, a_t, for each and every $t \geq 0$.

For any given pair of controller and adversarial policy $(\pi, \gamma) \in \Pi(\mathcal{Q}) \times \Gamma(\mathcal{P})$ and an initial state $x \in \mathbb{X}$, the distribution of the state and action process (X, A) is uniquely determined, see (A.2). We denote the expectation under this distribution as $E_x^{\pi,\gamma}$. Then, we define the DRSC value under initial distribution μ and controller's and adversarial policy classes $\Pi(\mathcal{Q})$ and $\Gamma(\mathcal{P})$ as

$$v^*(x, \Pi(\mathcal{Q}), \Gamma(\mathcal{P})) := \\ \sup_{\pi \in \Pi(\mathcal{Q})} \inf_{\gamma \in \Gamma(\mathcal{P})} E_x^{\pi,\gamma} \sum_{t=1}^{\infty} \alpha^t r(X_t, A_t). \tag{2.1}$$

In order for the DRSC value to satisfy a dynamic programming principle, i.e. a Bellman equation, regularity conditions for the reward r and the state transition function f are necessary.

Assumption 1. *Assume that $r : \mathbb{X} \times \mathbb{A} \to \mathbb{R}_+$ is non-negative $r_\vee$-bounded (i.e. $0 \leq r(x, a) \leq r_\vee, \forall x, a$) uniformly continuous and $f : \mathbb{X} \times \mathbb{A} \times \mathbb{W} \to \mathbb{X}$ is uniformly continuous. Furthermore, the controller can take on deterministic decisions; i.e. $\mathcal{Q} \supset \{\delta_{\{a\}} : a \in \mathbb{A}\}$.*

Remark. It is straightforward to generalize all the subsequent results in this paper to the case where the reward at time t is $r(X_t, A_t, W_t)$ which depends on the adversarial noise W_t. Moreover, for the dynamic programming part, the boundedness of reward and uniform continuities can be significantly weakened to the satisfaction of some growth conditions and semi-continuities (González-Trejo et al., 2002). However, to study the minimax statistical complexity, which is one of the main goals of the paper, we need even stronger conditions (e.g. uniform Lipschitz continuity) as in Assumption 2 and 3. Although relaxing the assumptions to be as general as possible while retaining a dynamic programming theory is of great theoretical value, our focus here is to develop a rigorous theoretical framework that allows us to study the statistical complexities of learning a DRSC. Hence, we adopt these more restrictive assumptions that are easy to work with.

Current-Action-Aware Adversary

Observe that, by our construction, a general adversarial decision γ_t at time t can be dependent on the entire history $g_t = (x_0, a_0, \cdots, x_t, a_t)$. In particular, the conditional distribution of W_t given history and different realizations of A_t could be different. In other words, the adversary is aware of the current action, hence the name current-action-aware adversary, and can use this information to harm the performance of the controller. This could be a reasonable model for settings in which ambiguity or non-Markov response arises from the system's reaction to the control inputs. One such application setting of particular practical importance is the portfolio optimization problem, where policy robustness is extremely valuable.

For CAA adversaries, we consider the following distributionally robust Bellman equation.

Definition 1 (Current-Action-Aware Bellman Equation)**.** We define the following fixed point equation as the Bellman equation for current-action-aware adversaries.

$$u(x) = \mathcal{T}(u)(x) := \\ = \sup_{a \in \mathbb{A}} r(x, a) + \alpha \inf_{\psi \in \mathcal{P}} \int_{\mathbb{W}} u(f(x, a, w))\psi(dw) \tag{2.2}$$

We will show in Proposition 2.1 that, under Assumption 1, the CAA Bellman equation (2.2) has a unique bounded solution u^*. Moreover, in the full discussion of the DRSC framework in Appendix A, we will rigorously establish the dynamic programming principles under which the DRSC value in (2.1) is equal to u^*.

Current-Action-Unaware Adversary

Although CAA adversaries give rise to natural distributionally robust control models in various important applications, there are settings for which the noise is considered as exogenous inputs to the system that are independent of the current action of the controller. For example, in the make-to-order manufacturing setting, to achieve robust planning, it might be reasonable to assume the presence of adversarial inter-arrival time given the previous completion times. However, assuming that the inter-arrival times can adversely relate to the service rate could lead to an overly conservative optimal policy, leading to a diminished value in the deployment environment where the interarrival time is independent of the service assignment.

To address this potential issue and increase the versatility of our framework we consider an adversary that cannot observe the current action. Concretely, an adversarial decision $\bar{\gamma}_t$ at time t is said to be *current-action-unaware* (CAU) if for any history $h_t = (x_0, a_0, \cdots, x_t)$ and $(h_t, a) = (x_0, a_0, \cdots, x_t, a)$, we have that $\bar{\gamma}_t(dw|h_t, a) = \bar{\gamma}(dw|h_t, a')$ for all $a, a' \in \mathbb{A}$. In other words, the conditional distribution of W_t given the history is independent of the current action A_t. Note that will use the "overline" notation to signify the CAU setting. We remark that in the DRMDP literature, this behavior is captured by S-rectangular adversaries, see Wiesemann et al. (2013); Wang et al. (2023b) and the reference therein.

For CAU adversaries, we define the following distributionally robust Bellman equation.

Definition 2 (Current-Action-Unaware Bellman Equation)**.** We define the following fixed point equation as the Bellman equation for current-action-unaware adversaries.

$$\bar{u}(x) = \overline{\mathcal{T}}(\bar{u})(x) := \sup_{\phi \in \mathcal{Q}} \inf_{\psi \in \mathcal{P}} \int_{\mathbb{A}\times\mathbb{W}} r(x,a) + \alpha \bar{u}(f(x,a,w))\phi \times \psi(da, dw) \tag{2.3}$$

Again, we will rigorously establish the dynamic programming principles in the full discussion of the DRSC framework in Appendix A.

With these formulation and modeling considerations in mind, we are ready to establish the existence and uniqueness of solutions to the proposed Bellman equation in the following Proposition 2.1. Throughout the paper, we use $\|\cdot\|$ to denote the supremum norm. Let $U_b(\mathbb{X})$ denote the space of uniformly bounded continuous functions on $\mathbb{X}$, which is a Banach space under $\|\cdot\|$. For notation simplicity, we define $\beta := (1-\alpha)^{-1}$.

Proposition 2.1. *Suppose Assumption 1 is in force. The Bellman equations* (2.2) *and* (2.3) *have unique fixed points u^* and $\bar{u}^*$ in $U_b(\mathbb{X})$, respectively. Moreover, $\|u^*\|, \|\bar{u}^*\| \le \beta r_\vee$.*

Combining Proposition 2.1 with the dynamic programming theory in Appendix A.3, we see that learning optimal DRSC value and hence a robust control can be achieved by finding a good approximation to the solutions of (2.2) and (2.3) in CAA and CAU settings, respectively.

3 Upper Bounds on Statistical Complexity

In Definitions 3 and 4, we first introduce Wasserstein distance and f_k-divergence ambiguity sets. We fix a measure $\mu_0 \in \mathcal{P}(\mathcal{W})$ as the center of the ambiguity sets, which is only accessible from samples.

Definition 3 (Wasserstein distance ambiguity sets)**.** Let $c : \mathbb{W} \times \mathbb{W} \to \mathbb{R}_+$ s.t. $c(w,w) = 0$ for all $w \in \mathbb{W}$. The Wasserstein distance for $\mu, \nu \in \mathcal{P}(\mathcal{W})$ with transportation cost function c is defined as

$$W_c(\mu,\nu) := \inf_{\xi \in \Xi(\mu,\nu)} \int_{\mathbb{W}\times\mathbb{W}} c d\xi,$$

where $\Xi(\mu,\nu)$ is the set of probability measures on $\mathcal{W}\times\mathcal{W}$ s.t. $\xi(\cdot, \mathbb{W}) = \mu$ and $\xi(\mathbb{W}, \cdot) = \nu$ for all $\xi \in \Xi(\mu,\nu)$. The Wasserstein distance constrained ambiguity set of adversarial decisions with cost c, transport budget δ and center μ_0 is $\mathcal{P} = \{\mu : W_c(\mu, \mu_0) \le \delta\}$.

Definition 4 (f_k-divergence ambiguity sets)**.** For $k > 1$ and probability measures $\mu \ll \mu_0$ in $\mathcal{P}(\mathcal{W})$, let

$$f_k(t) := \frac{t^k - kt + k - 1}{k(k-1)},$$
$$D_{f_k}(\mu\|\mu_0) := \int_{\mathbb{W}} f_k\left(\frac{d\mu}{d\mu_0}\right) d\mu_0.$$

Then, the f_k-divergence constrained ambiguity set of adversarial decisions with center μ_0 and radius δ is $\mathcal{P} = \{\mu \ll \mu_0 : D_{f_k}(\mu\|\mu_0) \le \delta\}$.

For both the Wasserstein distance and f_k-divergence settings, we let $\widehat{\mathcal{P}}$ do denote the n-sample empirical measure-centered version of the ambiguity sets; i.e. one that replaces μ_0 with $\hat{\mu}$ where $\hat{\mu}(\cdot) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{W_i \in \cdot\}$ for i.i.d. $W_i \sim \mu_0$. Further, we denote the CAA/CAU empirical Bellman operator as $\mathbf{T}/\overline{\mathbf{T}}$, where $\mathcal{P}$ in equation (2.2)/(2.3) is replaced by $\widehat{\mathcal{P}}$, i.e.,

$$\mathbf{T}(u)(x) := \sup_{a\in\mathbb{A}} r(x,a) + \alpha \inf_{\psi\in\widehat{\mathcal{P}}} \int_{\mathbb{W}} u(f(x,a,w))\psi(dw),$$

and

$$\overline{\mathbf{T}}(\bar{u})(x) := \sup_{\phi\in\mathcal{Q}}\inf_{\psi\in\widehat{\mathcal{P}}}\int_{\mathbb{A}\times\mathbb{W}} r(x,a)+\alpha\bar{u}(f(x,a,w))\phi\times\psi(da,dw).$$

Then, the empirical versions of (2.2) and (2.3) are $\mathbf{T}(u)(x)=u(x)$ and $\overline{\mathbf{T}}(\bar{u})(x)=\bar{u}(x)$.

Proposition 3.1. *Let u' and $\hat{u}$ be the solution to the population and empirical versions of (2.2) or (2.3); let $\mathcal{T}'$ and $\mathbf{T}'$ denote the corresponding population and empirical Bellman operators. Then, the estimation error in uniform norm is upper bounded by*

$$\|\hat{u}-u'\|\le\beta\left\|\mathbf{T}'(u')-\mathcal{T}'(u')\right\|.$$

In the following two sections, we establish statistical complexity upper bounds for learning the optimal value uniformly under CAA and CAU adversary models. In both cases, we consider both the Wasserstein distance and f_k-divergence-based adversarial ambiguity set. Our focus here is to obtain a tight convergence rate in n.

3.1 The Current-Action-Aware Case

We begin by clarifying notations and stating the assumptions under which our statistical analysis is carried out. For set $S\subset\mathbb{R}^d$, let $\mathrm{diam}(S):=\sup_{x,y\in S}|x-y|$, where $|\cdot|$ denotes the Euclidean distance. For the CAA adversary case, we assume the following.

Assumption 2. *Assume the following conditions:*

1. *The spaces $\mathbb{X}\subset\mathbb{R}^{d_{\mathbb{X}}},\mathbb{A}\subset\mathbb{R}^{d_{\mathbb{A}}},\mathbb{W}\subset\mathbb{R}^{d_{\mathbb{W}}}$ are equipped with the Euclidean distance. The state and action spaces are bounded: $\mathrm{diam}(\mathbb{A}),\mathrm{diam}(\mathbb{X})<\infty$.*
2. *To simplify notation, we require Assumption 1 to hold with $r_\vee=1$.*
3. *The mapping $(x,a)\to u^*(f(x,a,\cdot))$ uniform ℓ-Lipschitz, i.e.*

$$|u^*(f(x,a,y))-u^*(f(x',a',y))|\le\ell(|x-x'|+|a-a'|).$$

In the following Theorems 1 and 2, u^* and $\hat{u}$ are solutions to (2.2) with corresponding Wasserstein and f_k adversarial ambiguity sets $\mathcal{P}$ and $\widehat{\mathcal{P}}$ centered at μ_0 and $\hat{\mu}$, respectively.

Theorem 1 (Wasserstein distance constrained CAA adversary)**.** *Suppose Assumption 2 is in force. Also, assume that the cost is $c_\vee$-bounded; i.e. $\sup_{w,y\in\mathbb{W}}c(w,y)\le c_\vee$. Then, with $\mathcal{P}=\{\mu:W_c(\mu,\mu_0)\le\delta\}$, we have that*

$$\|\hat{u}-u^*\|\le c\left(\beta D_\vee\sqrt{d_{\mathbb{X}}+d_{\mathbb{A}}}+\beta^{3/2}\sqrt{\log\frac{1}{\eta}}\right)n^{-\frac{1}{2}}$$

w.p. at least $1-\eta$. Here, $c=46\alpha\sqrt{\pi}$ and

$$D_\vee:=\ell(\mathrm{diam}(\mathbb{X})+\mathrm{diam}(\mathbb{A}))+c_\vee\delta^{-1}\beta+1.$$

Remark. Employing the chaining technique, we eliminate an extra $\log n$ factor. This is at a cost of transforming root-log-diameters (c.f. Theorem 2) into linear diameter dependence. Retaining the $\log n$ allows reducing this to root-log-diameters.

For notation simplicity, define $k':=k/(k-1)$,

$$c_k(\delta):=(1+k(k-1)\delta)^{1/k},$$

and $a\vee b=\max\{a,b\}$

Theorem 2 (f_k-divergence constrained CAA adversary)**.** *Suppose Assumption 2 is in force. Then, with $\mathcal{P}=\{\mu\ll\mu_0:D_{f_k}(\mu\|\mu_0)\le\delta\}$, when $n\ge3\vee k$*

$$\|\hat{u}-u^*\|\le30\beta^2n^{-\frac{1}{k'\vee2}}c_k(\delta)^2\left(\frac{c_k(\delta)}{c_k(\delta)-1}\vee2\right)\times\left(\frac{1}{k}+\sqrt{D+\log\frac{1}{\eta}+2(d_{\mathbb{X}}+d_{\mathbb{A}})\log n}\right)$$

w.p. at least $1-\eta$. Here, the parameter

$$D:=d_{\mathbb{X}}\log\left(1+3\ell\mathrm{diam}(\mathbb{X})\right)+d_{\mathbb{A}}\log\left(1+3\ell\mathrm{diam}(\mathbb{A})\right).$$

3.2 The Current-Action-Unaware Case

For the CAU adversary case, we will operate under the following Assumption.

Assumption 3. *Assume the following conditions:*

1. *$\mathbb{X}\subset\mathbb{R}^{d_{\mathbb{X}}},\mathbb{W}\subset\mathbb{R}^{d_{\mathbb{W}}}$ are equipped with the Euclidean distance with $\mathrm{diam}(\mathbb{X})<\infty$.*
2. *The action space $\mathbb{A}$ is finite, equipped with the 0-1 distance; i.e. $d(a,a')=\mathbb{1}\{a=a'\}$.*
3. *Assumption 1 hold with $r_\vee=1$.*
4. *The mapping $(x,a)\to u^*(f(x,a,\cdot))$ uniform ℓ-Lipschitz, i.e.*

$$|u^*(f(x,a,y))-u^*(f(x',a',y))|\le\ell(|x-x'|+\mathbb{1}\{a=a'\}).$$

Remark. As we will discuss in Appendix C.4, our proof will generalize to continuum action spaces under additional covering number requirements, yielding $n^{-1/2}$ convergence rates in those settings. However, in this

paper, for CAU adversaries, we will focus on the cases where $|\mathbb{A}| < \infty$ to get concrete dependencies on the dimensions, diameters, and the size of the action space. However, it is not clear to us if the same rate can be achieved in the CAU setting if, for example, $\mathbb{A} \subset \mathbb{R}^d$ is compact and $\mathcal{Q} = \mathcal{P}(\mathcal{A})$. This is a limitation of this work to be addressed by future research.

In the following Theorems 3 and 4, $\bar{u}^*$ and $\hat{u}$ are solutions to (2.3) with corresponding Wasserstein and f_k adversarial ambiguity sets $\mathcal{P}$ and $\widehat{\mathcal{P}}$ centered at μ_0 and $\hat{\mu}$, respectively.

Theorem 3 (Wasserstein distance constrained CAU adversary)**.** *Suppose Assumption 3 is in force. Also, assume that the cost is $c_\vee$-bounded. Then, with $\mathcal{P} = \{\mu : W_c(\mu, \mu_0) \le \delta\}$,*

$$\|\hat{u} - \bar{u}^*\| \le c\left(\beta \overline{D}_\vee \sqrt{d_{\mathbb{X}} + |\mathbb{A}|} + \beta^{3/2}\sqrt{\log\frac{1}{\eta}}\right) n^{-1/2}$$

w.p. at least $1-\eta$. Here $c = 46\alpha\sqrt{\pi}$ and

$$\overline{D}_\vee := \ell \mathrm{diam}(\mathbb{X}) + 2\beta + c_\vee \delta^{-1}\beta + 1.$$

Theorem 4 (f_k-divergence constrained CAU adversary)**.** *Suppose Assumption 3 is in force. Then, with $\mathcal{P} = \{\mu \ll \mu_0 : D_{f_k}(\mu\|\mu_0) \le \delta\}$, when $n \ge 3 \vee k$*

$$\|\hat{u} - u^*\| \le 30\beta^2 n^{-\frac{1}{k'\vee 2}} c_k(\delta)^2 \left(\frac{c_k(\delta)}{c_k(\delta)-1} \vee 2\right) \times \left(\frac{1}{k} + \sqrt{\overline{D} + \log\frac{1}{\eta} + 2(d_{\mathbb{X}} + |\mathbb{A}|)\log n}\right)$$

w.p. at least $1-\eta$. Here, the parameter

$$\overline{D} := d_{\mathbb{X}} \log\left(1 + 3\ell \mathrm{diam}(\mathbb{X})\right) + |\mathbb{A}| \log\left(1 + 6\beta\right).$$

4 Lower Bounds on Finite Sample Minimax Risks

In this section, we study lower bounds on the finite sample minimax risk associated with learning the optimal value functions for both CAA and CAU adversary models. We demonstrate lower bounds that match the corresponding convergence rate upper bounds specified in the previous section.

Before we establish our lower bounds, we first introduce the finite sample minimax risk considered by this paper. For given value operator $\mathcal{K} : \mathcal{P}(\mathcal{W}) \to C(\mathbb{X})$ and a class of probability measures $\mathcal{U} \subset \mathcal{P}(\mathcal{W})$, we define the n-sample minimax risk over $\mathcal{U}$ of uniformly learning $\mathcal{K}(\mu)$ as

$$\mathfrak{M}_n(\mathcal{U}, \mathcal{K}) := \inf_K \sup_{\mu\in\mathcal{U}} E_{\mu^n} \sup_{x\in\mathbb{X}} |K(W_1, \cdots, W_n)(x) - \mathcal{K}(\mu)(x)|$$

where $\mu^n = \mu \times \cdots \times \mu$ is the n-fold product measure, and the first infimum is taken over all measurable functions $K : \mathbb{W}^n \to C(\mathbb{X})$.

As we will discuss in detail later, the operator $\mathcal{K}$ maps the center μ of the Wasserstein distance and f_k-divergence constrained adversarial ambiguity sets to the solution of the Bellman equations (2.2) and (2.3). According to the dynamic programming principles outlined in Appendix A.3, this solution corresponds to the optimal DRSC value. Therefore, $\mathfrak{M}_n(\mathcal{U}, \mathcal{K})$ represents the error incurred by the optimal learning algorithm (in terms of uniform performance over all centers of the ambiguity sets in $\mathcal{U}$) for the optimal value when the sample size is n. To match the rate in the previous upper bounds, it is sufficient to consider $\mathcal{U}$ as the family of Bernoulli distributions with parameter $p \in [0, 1]$.

Concretely, to establish our lower bound, we consider an instance with infinitely smooth reward and transition function s.t. the minimax risk associated with learning DRSC value function has lower bounds that match the convergence rate of upper bounds in Theorem 1-4.

Specifically, let $\mathbb{X} = \mathbb{A} = \mathbb{W} = [-1, 1]$, $f(x, a, w) = w$, and $r(x, a) = x$. For fixed controller and adversarial ambiguity set, the value operator is the mapping $\mathcal{K}(\mu) = u^*$, where u^* is the solution to (2.2) with μ_0 replaced by μ. Since f and r are independent of a, $u^* = \bar{u}^*$ is also the solution to (2.3).

Lemma 1. *Given adversarial ambiguity set $\mathcal{P}$, the solution to* (2.2) *and* (2.3) *are*

$$u^*(x) = \bar{u}^*(x) = x + \beta \inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} w\psi(dw). \tag{4.1}$$

Using this instance, we have the following lower bounds.

Theorem 5 (Lower bound for W_2-distance constrained Adversary)**.** *Let $\mathcal{P} = \{\mu : W_c(\mu, \mu_0) \le \delta\}$ with $c(x, y) = |x - y|^2$. The minimax risk of learning u^* or $\bar{u}^*$ over any*

$$\mathcal{U} \supset \left\{\mu = p\delta_{\{1\}} + (1-p)\delta_{\{0\}} : p \in [0, 1]\right\}$$

is lower bounded by

$$\mathfrak{M}_n(\mathcal{U}, \mathcal{K}) \ge \frac{\beta}{32} n^{-\frac{1}{2}}.$$

Remark. We state the theorem only in terms of the W_2 distance. This is just for the convenience of calculations. It is not hard to see from the proof that using a Taylor expansion argument, for W_p distances with $c(x, y) = |x - y|^p$, we have the same $n^{-1/2}$ rate, matching that in Theorem 1 and 3. Also, upon investigating the proof, one will find that the minimax risk

of estimating the value function at **one single** x has the same lower bound on the rate.

Theorem 6 (Lower bound for f_k-divergence constrained Adversary)**.** *Let* $\mathcal{P} = \{\mu \ll \mu_0 : D_{f_k}(\mu\|\mu_0) \le \delta\}$. *Then there exist constants* $C_1(k,\delta)$ *and* $C_2(k,\delta)$ *that only depend on* k, δ *s.t. whenever* $n \ge C_1(k,\delta)$*, the minimax risk of learning* u^* *or* $\bar{u}^*$ *over any*

$$\mathcal{U} \supset \left\{\mu = p\delta_{\{1\}} + (1-p)\delta_{\{0\}} : p \in [0,1]\right\}$$

is lower bounded by

$$\mathfrak{M}_n(\mathcal{U}, \mathcal{K}) \ge C_2(k,\delta)\beta n^{-\frac{1}{k'\vee 2}}.$$

Remark. The constants $C_1(k,\delta)$ and $C_2(k,\delta)$ are given in the proof in Appendix D.3. This matches the convergence rate up to a $\sqrt{\log n}$ in Theorem 2 and 4. Again, estimating the value function at one single x has the same lower bound.

5 Algorithm Design

Recall from our theoretical analysis that the solution $\hat{u}$ of the empirical Bellman equations achieves minimax-optimal convergence rates in estimating the true robust optimal values u^* and $\bar{u}$. Thus, our algorithm will focus on approximately solving the empirical Bellman equations: $\mathbf{T}(u) = u$ and $\overline{\mathbf{T}}(u) = u$.

In the subsequent discussion, we focus specifically on the f_k-divergence ambiguity setting, devising actor-critic-style algorithms for DRSC problems with CAA and CAU adversaries.

5.1 The CAA Case

In the CAA setting, we parameterize the value function $u_\theta : \mathbb{X} \to \mathbb{X}$ and the policy $\pi_\eta : \mathbb{X} \to \mathbb{A}$ as neural networks. Leveraging the actor-critic framework, we alternate between Bellman error minimization and policy improvement steps. Specifically, for a given policy-value pair (π_η, u_θ), the empirical Bellman operator is defined as:

$$\mathbf{T}_{\eta,\theta}(x) := r(x, \pi_\eta(x)) + \alpha \inf_{\psi \in \widehat{\mathcal{P}}} \int_{\mathbb{W}} u_\theta(f(x, \pi_\eta(x), w))\psi(dw),$$

where the minimization is performed over the empirical f_k-divergence ambiguity set $\widehat{\mathcal{P}}$ centered around the empirical measure $\hat{\mu}$, as described in previous sections.

Strong Duality and Optimal Lagrange Multiplier: To compute gradients of the empirical Bellman operator, we invoke strong duality (Duchi and Namkoong, 2021), transforming the minimization over probability measures into an equivalent one-dimensional convex problem:

$$\mathbf{T}_{\eta,\theta}(x) = r(x, \pi_\eta(x)) + \alpha \sup_{\lambda \in \mathbb{R}} \Bigg\{ \lambda - c_k(\delta) \left[\int_{\mathbb{W}} (u_\theta(f(x, \pi_\eta(x), w)) - \lambda)_+^{k'} \hat{\mu}(dw) \right]^{1/k'} \Bigg\}.$$

Here, λ is the dual multiplier. Under mild assumptions, the maximization in λ has a unique optimal solution $\lambda^*(\eta, \theta, x)$, which can be efficiently computed using bisection search. We define the intermediate operator for convenience:

$$\widehat{\mathbf{T}}_{\eta,\theta}(\lambda, x) = r(x, \pi_\eta(x)) + \alpha \Bigg\{ \lambda - c_k(\delta) \left[\int_{\mathbb{W}} (u_\theta(f(x, \pi_\eta(x), w)) - \lambda)_+^{k'} \hat{\mu}(dw) \right]^{1/k'} \Bigg\}.$$

Bellman Error Minimization: Given a fixed policy π_η, we minimize the squared L^2 Bellman error:

$$\min_\theta \int_{\mathbb{X}} [u_\theta(x) - \mathbf{T}_{\eta,\theta}(x)]^2 \nu(dx),$$

where ν is a user-specified probability measure supported on the entire state space $\mathbb{X}$. To compute gradients, we utilize the envelope theorem to get that:

$$\nabla_\theta \mathbf{T}_{\eta,\theta}(x) = \nabla_\theta \widehat{\mathbf{T}}_{\eta,\theta}(\lambda^*(\eta, \theta, x), x). \tag{5.1}$$

To minimize Bellman Error we update the θ using first-order algorithms. For illustrative purposes, we formulate the mini-batch stochastic gradient descent in this context. At each iteration, we independently sample states $X_i \sim \nu$, compute the optimal multipliers λ_i^* in parallel, and update the parameters as follows:

$$\theta_{t+1} = \theta_t - \ell_t \frac{1}{n} \sum_{i=1}^n \nabla_\theta [(u_\theta(X_i) - \nabla_\theta \mathbf{T}_{\eta,\theta}(x))^2],$$

where the gradient w.r.t. θ is computed using the chain-rule and (5.1).

Policy Improvement: Given a fixed value function u_θ, the policy π_η is improved by solving the following optimization problem with first-order methods:

$$\max_\eta \int_{\mathbb{X}} \mathbf{T}_{\eta,\theta}(x) \nu(dx)$$

with gradients computed similarly using the envelope theorem. Again, we update parameters via stochastic gradient ascent:

$$\eta_{t+1} = \eta_t + \ell_t' \frac{1}{n} \sum_{i=1}^n \nabla_\eta \widehat{\mathbf{T}}_{\eta,\theta}(\lambda_i^*, X_i).$$

5.2 The CAU Case with Continuous Action

While our theoretical analysis focuses primarily on finite action models in the CAU setting, practical applications often involve continuous action spaces. Hence, we propose an extended algorithm suitable for continuous state-action problems. The CAU case is inherently more complex since deterministic policies are generally suboptimal, necessitating the consideration of randomized policies. To address this, we propose a novel *generative* policy approach. Specifically, we view a policy as a generative model that produces randomized actions given a state $x \in \mathbb{X}$ by utilizing an external source of randomness.

Concretely, we define the policy as a mapping $\pi_\eta : \mathbb{R}^d \times \mathbb{X} \to \mathbb{A}$, where an action is generated according to $\pi_\eta(N, x)$ and $N \sim N(0, I)$ is a standard normal random vector independent of the state. Due to space constraints, a detailed discussion of the algorithm for the CAU case is provided in Appendix E.

6 Applications

In this section, we map two applications into our DRSC framework. Specifically, Section 6.1 outlines a multi-period portfolio optimization problem where market returns may react adversarially to specific actions. Section 6.2 examines a service optimization problem in make-to-order systems, where the distributions of inter-arrival times may be misspecified.

6.1 Portfolio Optimization

We manage a portfolio of m assets where the time is divided into discrete time periods $t = 1, 2, ...$ (not necessarily uniformly spaced in real time). $X_t \in \mathbb{R}^m$ denotes the portfolio (or vector of positions) at time t, where $X_{t,i}$ is the dollar value of asset i at the beginning of time period t: $X_{t,i} > 0$ and $X_{t,i} < 0$ mean a long position and a short position in asset i, respectively.

We can buy and sell assets at the beginning of each time period. Let $A_t \in \mathbb{R}^m, t = 0, 1, \dots$ be our decision variables at t, which is the dollar values of the trades: $A_{t,i} > 0$ or $A_{t,i} < 0$ means buying or selling asset i at the beginning of time period t, respectively. For simplicity, we assume that there is no contribution of capital. However, one can consume the portfolio by having a total sale that is higher than the total purchase; i.e. $\mathbf{1}^\top A_t \leq 0$ where $\mathbf{1} \in \mathbb{R}^n$ is the column vector of all 1's. Therefore, the consumption is $C_t = -\mathbf{1}^\top A_t \geq 0$, which is the cash amount taken out from the portfolio. As such, the self-financing constraint (in the absence of any transaction costs) becomes $\mathbf{1}^T A_t + C_t = 0$. Hence, it suffices to consider the action space $\mathbb{A} := \left\{a \in \mathbb{R}^m : \mathbf{1}^\top a \leq 0\right\}$, and $\mathcal{Q} = \mathcal{P}(\mathcal{A})$, as the consumption $C_t = -\mathbf{1}^T A_t$ is determined by a feasible investment A_t.

At the beginning of the next period, we have $X_{t+1} = W_t(X_t + A_t)$, where $W_t = \text{diag}(R_t) \in \mathbb{R}^{m \times m}$ is the matrix of asset returns, and $R_{t,i}$ is the return of the i-th asset from period t to period $t + 1$.

We assume that the decision maker is endowed with a utility function $U : \mathbb{R}_+ \to \mathbb{R}$, which is concave and non-decreasing. Then, taking action A_t will incur an instantaneous reward utility $r(X_t, A_t) = U(-\mathbf{1}^\top A_t)$. The performance of the policy is measured in terms of the expected infinite horizon discounted total utilities.

As motivated before, in a dynamic portfolio optimization environment, there might be a non-trivial (typically adversarial) market response in reaction to the change in portfolio position induced by current action A_t. Therefore, a CAA adversary could be a reasonable model for robust dynamic decision-making in this context. Thus, given i.i.d. data $\{R_t \in \mathbb{R}^m, t = 1, \cdots n\}$, we can build Wasserstein distance or f_k-divergence ambiguity set $\widehat{\mathcal{P}}$ from the empirical measure $\hat{\mu}$ induced by the data. Then, we estimate the DRSC value function by solving (2.2) with $\widehat{\mathcal{P}}$.

6.2 Service and Manufacturing Systems

Following Example 2, we consider a make-to-order systems. The goal is to minimize discounted sum of Waiting time by dynamically adjusting the service rate. The cost of service rates is denoted as $c(A_n)$, where $c(\cdot)$ is an increasing function. Therefore, the reward function can be written as $r(X_n, A_n) = -X_n - c(A_n)$. And recall the state recursion is $X_{n+1} = (X_n + 1/A_n - W_n)^+$.

Here the randomness came from the inter-arrival times $\{W_n\}$. The standard queueing theory assumes the demand $W_1, W_2, \dots, W_n, \dots$ are i.i.d.. In practice, however, the distribution of W_n is often unknown, and the demand sequence $W_1, W_2, \dots, W_n, \dots$ is non-i.i.d. and non-stationary. For example, an empirical study (Kim and Whitt, 2014) and a publicly available dataset from a US bank call center, "DataMOCCA" (Trofimov et al., 2006), indicate that arrivals at call centers and hospitals exhibit significant time-of-day and day-of-week effects. Therefore, it is important to model this ambiguity using our distributionally robust framework. In this framework, we use adversaries to model all possible inter-arrival (joint) distributions the system manager may face. It is natural to use a CAU adversary, as this inter-arrival time should not depend on the manager's service decision.

Acknowledgements

The material in this paper is partly supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397 and ONR 13983111 (award number N000142412655), 13983263 (award number N000142412672). Support from NSF 2229012, 2312204, 2312205, 2403007, 2419564, and 2025 New York University Center for Global Economy and Business grant is also gratefully acknowledged.

References

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. (2024). Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pages 1042–1051. PMLR.

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

Foschini, G. J. and Miljanic, Z. (1993). A simple distributed autonomous power control algorithm and its convergence. *IEEE transactions on vehicular Technology*, 42(4):641–646.

González-Trejo, J., Hernández-Lerma, O., and Hoyos-Reyes, L. F. (2002). Minimax control of discrete-time stochastic systems. *SIAM Journal on Control and Optimization*, 41(5):1626–1659.

Goyal, V. and Grand-Clément, J. (2023). Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226.

Han, B. (2023). Distributionally robust kalman filtering with volatility uncertainty. *arXiv preprint arXiv:2302.05993*.

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

Kim, K. and Yang, I. (2023). Distributional robustness in minimax linear quadratic control with wasserstein distance. *SIAM Journal on Control and Optimization*, 61(2):458–483.

Kim, S.-H. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480.

Kotsalis, G., Lan, G., and Nemirovski, A. S. (2021). Convex optimization for finite-horizon robust covariance control of linear stochastic systems. *SIAM Journal on Control and Optimization*, 59(1):296–319.

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

Lee, J. and Raginsky, M. (2018). Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 31.

Li, Y. and Shapiro, A. (2023). Rectangularity and duality of distributionally robust markov decision processes.

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. (2022). Distributionally robust Q-learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13623–13643. PMLR.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1):125–144.

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

Panaganti, K. and Kalathil, D. (2021). Sample complexity of robust reinforcement learning with a generative model.

Petersen, I. R., James, M. R., and Dupuis, P. (2000). Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control*, 45(3):398–412.

Porteus, E. L. (2002). *Foundations of stochastic inventory theory*. Stanford University Press.

Shapiro, A. and Li, Y. (2024). Distributionally robust stochastic optimal control.

Shi, L. and Chi, Y. (2022). Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2023). The curious price of distributional robustness in reinforcement learning with a generative model.

Taskesen, B., Iancu, D., Koçyiğit, Ç., and Kuhn, D. (2024). Distributionally robust linear quadratic control. *Advances in Neural Information Processing Systems*, 36.

Trofimov, V., Feigin, P., Mandelbaum, A., Ishay, E., and Nadjharov, E. (2006). Data-mocca: Data model for call center analysis. *Model Description and Introduction to User Interface*, 1.

Tse, D. and Grossglauser, M. (1997). Measurement-based call admission control: Analysis and simulation. In *Proceedings of INFOCOM'97*, volume 3, pages 981–989. IEEE.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023a). A finite sample complexity bound for distributionally robust Q-learning.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023b). On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023c). Sample complexity of variance-reduced distributionally robust Q-learning.

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

Xu, H. and Mannor, S. (2010). Distributionally robust markov decision processes. In *NIPS*, pages 2505–2513.

Xu, Z., Panaganti, K., and Kalathil, D. (2023). Improved sample complexity bounds for distributionally robust reinforcement learning.

Yang, I. (2021). Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 66(8):3863–3870.

Yang, W., Wang, H., Kozuno, T., Jordan, S. M., and Zhang, Z. (2023). Avoiding model estimation in robust markov decision processes with a generative model.

Yang, W., Zhang, L., and Zhang, Z. (2021). Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics.

Yüksel, S. and Başar, T. (2013). *Stochastic networked control systems: Stabilization and optimization under information constraints*. Springer Science & Business Media.

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3331–3339. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes: This paper is self-contained and rigorous.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes: This is one of the main contributions of the paper.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Not Applicable.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes: Theorems are rigorously established.
   (b) Complete proofs of all theoretical results. Yes: All theorems, propositions, and lemmas are proved.
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Not Applicable.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable.

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

(c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.

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

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

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

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

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

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

Supplementary Materials for Statistical Learning of Distributionally Robust Stochastic Control in Continuous State Spaces

A Formulations of Distributionally Robust Stochastic Control

Let $\mathbb{X}, \mathbb{A}, \mathbb{W}$ be Polish spaces and $(\mathbb{X}, \mathcal{X}), (\mathbb{A}, \mathcal{A}), (\mathbb{W}, \mathcal{W})$ equip them with the Borel σ-fields. Let $\mathcal{P}(\mathcal{A})$ and $\mathcal{P}(\mathcal{W})$ be the set of probability measures on $(\mathbb{A}, \mathcal{A})$ and $(\mathbb{W}, \mathcal{W})$, respectively. Endow them with the topology of weak convergence; i.e. $\mu_n \Rightarrow \mu$ if $\int f d\mu_n \to \int f d\mu$ for all bounded continuous f. Then, $\mathcal{P}(\mathcal{A})$ and $\mathcal{P}(\mathcal{W})$ are separable, as $(\mathbb{A}, \mathcal{A})$ and $(\mathbb{W}, \mathcal{W})$ are separable.

We now present our distributionally robust stochastic control formulation. Let $\Omega = \mathbb{X} \times (\mathbb{A} \times \mathbb{W})^{\mathbb{Z}_+}$ and $\mathcal{F}$ is the σ-field generated by cylinder sets. A canonical element $\omega \in \Omega$ is $\omega = (x_0, a_0, w_0, a_1, w_1 \cdots a_t, w_t \cdots)$, $x_0 \in \mathbb{X}$, $w_k \in \mathbb{W}$, and $a_k \in \mathbb{A}$, $\forall k \geq 0$.

Let $W := \{W_t : t \geq 0\}$ and $A := \{A_t : t \geq 0\}$ be the processes of point evaluation of $\{w_t : t \geq 0\}$ and $\{a_t : t \geq 0\}$, respectively; i.e.

$$W_t(\omega) = w_t, \quad A_t(\omega) = a_t.$$

Finally, define the process $X := \{X_t : t \geq 0\}$ by the stochastic recursion $X_0(\omega) = x_0$ and for each $t \geq 0$

$$X_{t+1} = f(X_t, A_t, W_t).$$

We refer to X as the controlled state process, A as the action process, and W as the exogenous noise process. In the classical stochastic control setting, a typical assumption is that the noise process W consists of i.i.d. W_t under any probability measure of interest on $(\Omega, \mathcal{F})$. In our setting, however, the adversary can dynamically perturb the distribution of W_t based on some or all historical information, potentially making it a general stochastic process with arbitrary dependent structure.

A.1 Admissible Policies

In this section, we rigorously formulate the controller and adversarial policies under the DR Stochastic control framework. We formulate the the controller and the adversary policies so that, collectively, they will give rise to a unique probability measure on $(\Omega, \mathcal{F})$. At a high level, for each and every $t \geq 0$, the controller and the adversary choose the conditional distributions of A_t and W_t respectively, given their available information.

Let us define the following notations. For measures μ, ν on $(\mathbb{C}, \mathcal{C})$, we write $\mu(dc) = \nu(dc)$ if $\mu(C) = \nu(C)$ for all $C \in \mathcal{C}$.

For $t \geq 0$, define controller's history

$$\mathbf{H}_t := \{h_t = (x_0, a_0, \cdots, a_{t-1}, x_t) : x_k \in \mathbb{X}, a_k \in \mathbb{A}, \forall k\}.$$

and the adversarial history

$$\mathbf{G}_t := \{g_t = (x_0, a_0, \cdots, x_t, a_t) : x_k \in \mathbb{X}, a_k \in \mathbb{A}, \forall k\}.$$

For convenience, we let $\mathbf{H}_{-1} = \mathbf{G}_{-1} = \varnothing$.

Define the history random elements

$$H_t(\omega) := h_t = (x_0, a_0, \cdots, x_t) \in \mathbf{H}_t \quad \text{and} \quad G_t(\omega) := g_t = (x_0, a_0, \cdots, x_t, a_t) \in \mathbf{G}_t$$

where $x_{k+1} = f(x_k, a_k, w_k)$ for $k = 1, \cdots, t-1$, recursively.

Admissible Controller's Policies

A *decision* of the controller π_t at time t is a (product space) Borel measurable function $\pi_t : \mathbf{H}_t \to \mathcal{P}(\mathcal{A})$. This is seen as the conditional distribution of A_t given history $H_t = h_t$, hence we write $\pi_t(da|h_t)$. A *policy* of the controller $\pi = (\pi_0, \pi_1, \cdots)$ is a sequence of decisions. The largest possible policy class under this framework is the the history-dependent unconstrained controller's policy class:

$$\Pi_{\mathrm{H}} := \{\pi = (\pi_0, \pi_1, \cdots) : \pi_t \in m\{\mathbf{H}_t \to \mathcal{P}(\mathcal{A})\}\}$$

where $m\{\mathbf{H}_t \to \mathcal{P}(\mathcal{A})\}$ denote the set of Borel measureable functions.

To increase the modeling flexibility of our DR stochastic control framework, we consider constraints on the controller in terms of information availability and admissible set of controller's decisions.

We say that a controller's policy $\pi = (\pi_0, \pi_1, \cdots)$ is Markov if for each and every $t \geq 0$,

$$\pi_t(da|g_{t-1}, x_t) = \pi_t(da|g'_{t-1}, x_t)$$

for any $g_{t-1}, g'_{t-1} \in \mathbf{G}_{t-1}$ and $x_t \in \mathbb{X}$; i.e. given x_t, the distribution of the action is independent of the history g_{t-1}. Therefore, through an abuse of notation, we can write $\pi_t(da|x)$ when the decision is Markov. Denote the set of Markov controller's policies by Π_{M}.

Moreover, π is said to be time-homogeneous (or stationary Markov) if

$$\pi_t(da|g_{t-1}, x) = \pi_s(da|g'_{s-1}, x)$$

for every $s, t \geq 0$ and $g_{t-1} \in \mathbf{G}_{t-1}$, $g'_{s-1} \in \mathbf{G}_{s-1}$ and $x \in \mathbb{X}$; i.e. π is Markov and invariant in time. As in the Markov case, we can write $\pi(da|x)$ when the decision is time-homogeneous. We denote the set of time-homogeneous controller's policies by Π_{S}.

We further allow the controller to be constrain to choose its decision $\pi_t(da|h_t) \in \mathcal{Q}$ from a admissible subset $\mathcal{Q} \subset \mathcal{P}(\mathcal{A})$ that is Borel measureable. This can be done under any information structures defined above. We denote such constrained controller with the corresponding information availability as $\Pi_{\mathrm{U}}(\mathcal{Q})$ where $\mathrm{U} = \mathrm{H}, \mathrm{M}, \mathrm{S}$.

Admissible Adversarial Policies

A decision of the adversary γ_t at time t is a measurable function $\gamma_t : \mathbf{G}_t \to \mathcal{P}(\mathcal{W})$, where we write $\gamma_t(dw|g_t)$ and note that it signifies the conditional distribution of W_t given $G_t = g_t$. An adversarial policy $\gamma = (\gamma_0, \gamma_1, \cdots)$ is a sequence of adversarial decisions. This forms the the history-dependent unconstrained adversary's policy class:

$$\Gamma_{\mathrm{H}} := \{\gamma = (\gamma_0, \gamma_1, \cdots) : \gamma_t \in m\{\mathbf{G}_t \to \mathcal{P}(\mathcal{W})\}\}.$$

We define an adversary's policy to be Markov if $\gamma_t(da|g_{t-1}, x_t, a_t) = \gamma_t(da|g'_{t-1}, x_t, a_t)$ for any $g_{t-1}, g'_{t-1} \in \mathbf{G}_{t-1}$ and $x_t \in \mathbb{X}, a_t \in \mathbb{A}$. Further, an adversary's policy is time-homogeneous (or stationary Markov) if $\gamma_t(da|g_{t-1}, x, a) = \gamma_s(da|g'_{s-1}, x, a)$ for every $s, t \geq 0$ and $g_{t-1} \in \mathbf{G}_{t-1}$, $g'_{s-1} \in \mathbf{G}_{s-1}$ and $x \in \mathbb{X}, a \in \mathbb{A}$. As in the controller setting, we write $\gamma_t(dw|x, a)$ and $\gamma(dw|x, a)$ for Markov and time-homogeneous adversary decisions, respectively. Denote the Markov and time-homogeneous adversarial policy classes by Γ_{M} and Γ_{S}, respectively.

As for the controller's case, we allow the adversary to be constrain to choose $\gamma_t(da|g_t) \in \mathcal{P}$ from a admissible subset $\mathcal{P} \subset \mathcal{P}(\mathcal{W})$ that is Borel measureable. We denote such constrained adversarial policy classes with the corresponding information availability as $\Gamma_{\mathrm{U}}(\mathcal{P})$ where $\mathrm{U} = \mathrm{H}, \mathrm{M}, \mathrm{S}$.

Remark. Notice that in this model, the history-dependent controller cannot directly use the realized action w_t of the adversary to make its decision. This should be compared to the stochastic game settings González-Trejo et al. (2002) in which either player observes the action of the other and makes a decision based on such observation. However, this model can include the settings for which both players see the action of each other by considering a new state process $z_t = (x_t, w_{t-1})$ and defining the state space and histories using z_t instead of x_t.

We also note that in general settings for which the modeler decides to construct f so that $W_{k-1} \notin \sigma(X_k)$ for each $k \leq t$. Then, this is as if the adversary cannot use its historical actions $\{W_k : k \leq t-1\}$ to decide the distribution of the current action W_t.

Current-Action-Unaware Adversary

Consider adversarial policy $\gamma = (\gamma_0, \gamma_1, \cdots) \in \Gamma_{\mathrm{H}}$. Because in general, the distribution of W_t depends on the current action a_t through g_t, i.e. $W_t \sim \gamma_t(dw|g_t)$, we say that they are *current-action-aware* (CAA). However, in many settings, the adversary cannot base its decision on the current action a_t. Such adversary is characterized by the following concept of *current-action-unaware* (CAU) decisions.

We say that a adversary's decision γ_t is current-action-unaware if

$$\gamma_t(dw|g_t) = \gamma_t(dw|h_t, a') \tag{A.1}$$

for all $a' \in \mathbb{A}$, where $g_t = (h_t, a_t)$. Then the set of history dependent current-action-unaware adversary with constraint set $\mathcal{P}$ is a subset $\overline{\Gamma}_{\mathrm{H}} \subset \Gamma_{\mathrm{H}}$ defined by

$$\overline{\Gamma}_{\mathrm{H}}(\mathcal{P}) := \{\gamma = (\gamma_0, \gamma_1, \cdots) : \gamma_t \in m\left\{\mathbf{G}_t \to \mathcal{P}\right\}, \gamma_t(dw|g_t) = \gamma_t(dw|h_t, a'), \forall a' \in A\}$$

When γ is independent of the current action, we write $\bar{\gamma}_t(dw|h_t) := \gamma_t(dw|h_t, a)$. Hence, we have $\bar{\gamma} = (\bar{\gamma}_0, \bar{\gamma}_1, \cdots) \in \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})$.

This can be easily generalized to the Markov and time-homogeneous settings by $\overline{\Gamma}_{\mathrm{M}}(\mathcal{P}) := \overline{\Gamma}_{\mathrm{H}}(\mathcal{P}) \cap \Gamma_{\mathrm{M}}(\mathcal{P})$ and $\overline{\Gamma}_{\mathrm{S}}(\mathcal{P}) := \overline{\Gamma}_{\mathrm{H}}(\mathcal{P}) \cap \Gamma_{\mathrm{S}}(\mathcal{P})$, consisting of Markov and time-homogeneous policies for which the decision at any time is current-action-unaware as defined in (A.1).

A.2 The Distributionally Robust Stochastic Control Problem

Given an initial distribution μ_0 on $\mathcal{P}(\mathcal{X})$ and a pair of controller's and adversary's policy (π, γ), define a probability measure $P_\mu^{\pi,\gamma}$ on Ω as follows. For cylinder sets of the form

$$C_t := B_0 \times Y_0 \times \cdots \times B_t \times Y_t \times \mathbb{A} \times \mathbb{W} \times \mathbb{A} \times \mathbb{W} \cdots$$

for some $B_k \in \mathcal{A}$ and $Y_k \in \mathcal{W}$ for each $k \le t$, define

$$P_\mu^{\pi,\gamma}(C_t) := \int_{\mathbb{X}} \int_{B_0} \int_{Y_0} \int_{B_1} \cdots \int_{Y_t} \gamma_t(dw_t|g_t) \cdots \pi_1(da_1|h_1)\gamma_0(dw_0|g_0)\pi_0(da_0|h_0)\mu_0(dx_0). \tag{A.2}$$

This uniquely extends to a probability measure on $(\Omega, \mathcal{F})$. Let $E_\mu^{\pi,\gamma}$ denote the expectation under $P_\mu^{\pi,\gamma}$.

The distributionally robust stochastic control (DRSC) paradigm under this formulation where an adversary perturbs the exogenous driving randomness aims to find the infinite horizon discounted maxmin value function

$$v^*(\mu, \Pi, \Gamma) := \sup_{\pi \in \Pi} \inf_{\gamma \in \Gamma} E_\mu^{\pi,\gamma} \sum_{t=1}^{\infty} \alpha^t r(X_t, A_t) \tag{A.3}$$

subject to $X_{k+1} = f(X_k, A_k, W_k)$, $\forall k$. Here, the admissible policy classes Π, Γ are $\Pi = \Pi_{\mathrm{U}}(\mathcal{Q})$ and $\Gamma = \Gamma_{\mathrm{U}}(\mathcal{P}), \overline{\Gamma}_{\mathrm{U}}(\mathcal{P})$ where $\mathrm{U} = \mathrm{H}, \mathrm{M}, \mathrm{S}$. This is the rigorous version of (2.1).

For simplicity, we write $v^*(x, \Pi, \Gamma) := v^*(\delta_{\{x\}}, \Pi, \Gamma)$, and $v^*(\Pi, \Gamma)$ can be seen as a function $x \to v^*(x, \Pi, \Gamma)$.

A.3 Dynamic Programming

In this section, we show that the solutions to the distributional robust Bellman equations (2.2) and (2.3) will correspond to the DRSC value (A.3).

Theorem 7 (Dynamic Programming for CAA Adversaries)**.** *Suppose Assumption 1 is in force, then* $u^* = v^*(\Pi(\mathcal{Q}), \Gamma(\mathcal{P}))$ *for each and every one of the 9 pairings* $\Pi(\mathcal{Q}) = \Pi_{\mathrm{H}}(\mathcal{Q}), \Pi_{\mathrm{M}}(\mathcal{Q}), \Pi_{\mathrm{S}}(\mathcal{Q})$ *and* $\Gamma(\mathcal{P}) = \Gamma_{\mathrm{H}}(\mathcal{P}), \Gamma_{\mathrm{M}}(\mathcal{P}), \Gamma_{\mathrm{S}}(\mathcal{P})$.

Theorem 8 (Dynamic Programming for CAU Adversaries)**.** *Suppose Assumption 1 is in force, then there is a unique bounded continuous solution* $\bar{u}^*$ *to* (2.3)*. Moreover,* $\bar{u}^*$ *is the optimal DRSC values*

$$\begin{aligned}
\bar{u}^* &= v^*(\Pi_{\mathrm{H}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) \\
&= v^*(\Pi_{\mathrm{M}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) && = v^*(\Pi_{\mathrm{M}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{M}}(\mathcal{P})) \\
&= v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) && = v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{M}}(\mathcal{P})) && = v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{S}}(\mathcal{P})).
\end{aligned}$$

Remark. The equality $v^*(\Pi_{\mathrm{H}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) = v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P}))$ and $v^*(\Pi_{\mathrm{M}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{M}}(\mathcal{P})) = v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{M}}(\mathcal{P}))$ implies time-homogeneous (or stationary Markov) policies are optimal for history-dependent and Markov adversary.

A history-dependent version of Theorem 7 is established in González-Trejo et al. (2002). In this paper, we will prove the more technically interesting Theorem 8. The proof of the rest of the Theorem 7 can be easily achieved by adapting the same proof techniques to deal with continuous state spaces in this paper to that of Theorem 2 from Wang et al. (2023b).

B Proofs for Section 2 and A.3

B.1 Proof of Proposition 2.1

We prove Proposition 2.1 by applying the Banach fixed-point to the mapping $\mathcal{T}$ and $\overline{\mathcal{T}}$.

Lemma 2. *Under Assumption 1, $\mathcal{T}$ and $\overline{\mathcal{T}}$ are α-contractions on $(U_b(\mathbb{X}), \|\cdot\|)$; i.e. $\mathcal{T}' : U_b(\mathbb{X}) \to U_b(\mathbb{X})$ satisfies*

$$\|\mathcal{T}'(u_1) - \mathcal{T}'(u_2)\| \leq \alpha \|u_1 - u_2\|$$

for all $u_1, u_2 \in U_b(\mathbb{X})$, where $\mathcal{T}' = \mathcal{T}, \overline{\mathcal{T}}$.

Therefore, there exists unique fixed-points u^* for (2.2) and $\bar{u}^*$ (2.3).

Moreover, for $\mathcal{T}' = \mathcal{T}, \overline{\mathcal{T}}$ and $u' = u^*, \bar{u}^*$, we have that

$$\|u'\| = \|\mathcal{T}'(u')\| \leq \|r\| + \alpha \|u'\| = r_\vee + \alpha \|u'\| .$$

Hence, $\|u'\| \leq \beta \|u'\|$.

B.1.1 Proof of Lemma 2

We will establish the result for $\overline{\mathcal{T}}$, the statement for $\mathcal{T}$ follows from the same arguments. First, we check that for $u \in U_b(\mathbb{X})$, $\overline{\mathcal{T}}(u) \in U_b(\mathbb{X})$. Observe that by uniform continuity, for $x, z \in \mathbb{X}$ s.t. $d(x, z) \leq \epsilon$ there are $\delta, \delta', \delta'' > 0$ s.t.

$$\begin{aligned}
&\left|\overline{\mathcal{T}}(u)(z) - \overline{\mathcal{T}}(u)(x)\right| \\
&\leq \sup_{\phi \in \mathcal{Q}} \left| \inf_{\psi \in \mathcal{P}} \int_{\mathbb{A}\times\mathbb{W}} r(x,a) + \alpha u(f(x,a,w))\phi \times \psi(da, dw) + \sup_{\psi \in \mathcal{P}} \int_{\mathbb{A}\times\mathbb{W}} -r(z,a) - \alpha u(f(z,a,w))\phi \times \psi(da, dw) \right| \\
&\overset{(i)}{\leq} \sup_{\phi \in \mathcal{Q}} \sup_{\psi \in \mathcal{P}} \int_{\mathbb{A}\times\mathbb{W}} |r(x,a) - r(z,a)| + \alpha |u(f(x,a,w)) - \alpha u(f(z,a,w))| \phi \times \psi(da, dw) \\
&\leq \delta + \sup_{a \in \mathbb{A}, w \in \mathbb{W}} \sup_{y \in \bar{B}_{f(x,a,w)}(\delta')} \alpha |u(f(x,a,w)) - u(y)| \\
&\leq \delta + \delta''
\end{aligned}$$

uniformly in x, where (i) follows from $|\inf f_1 + \sup f_2| \leq \max\{|\sup(f_1 + f_2)|, |\inf(f_1 + f_2)|\} \leq \sup |f_1 + f_2|$ and $\bar{B}_{f(x,a,w)}(\delta') := \{y \in \mathbb{X} : d(f(x,a,w), y) \leq \delta'\}$. Hence, $\overline{\mathcal{T}}(u) \in U_b(\mathbb{X})$.

Next, we show that it is indeed a α-contraction. Consider for $u_1, u_2 \in U_b(\mathbb{X})$, by the same argument, one has

$$\begin{aligned}
\left\|\overline{\mathcal{T}}(u_1) - \overline{\mathcal{T}}(u_2)\right\| &\leq \sup_{x \in \mathbb{X}, \phi \in \mathcal{Q}, \psi \in \mathcal{P}} \int_{\mathbb{A}\times\mathbb{W}} \alpha |u_1(f(x,a,w)) - u_2(f(x,a,w))| \phi \times \psi(da, dw) \\
&\leq \sup_{x \in \mathbb{X}, a \in \mathbb{A}, w \in \mathbb{W}} \alpha |u_1(f(x,a,w)) - u_2(f(x,a,w))| \\
&\leq \alpha \|u_1 - u_2\| .
\end{aligned}$$

This completes the proof.

B.2 Proof of Theorem 8

We decompose our proof of Theorem 8 to two main Propositions as follows.

Proposition B.1. *Under the assumptions of Theorem 8, for any* $\pi \in \Pi_{\mathrm{U}}(\mathcal{Q})$,

$$\inf_{\bar{\gamma}\in\overline{\Gamma}_{\mathrm{U}}(\mathcal{P})} v(x,\pi,\bar{\gamma}) \le \bar{u}^*(x),$$

where $\mathrm{U} = \mathrm{H}, \mathrm{M}, \mathrm{S}$.

In particular, Proposition B.1 implies that $\bar{u}^* \ge v^*(\Pi_{\mathrm{H}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P}))$, $\bar{u}^* \ge v^*(\Pi_{\mathrm{M}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{M}}(\mathcal{P}))$, and $\bar{u}^* \ge v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{S}}(\mathcal{P}))$.

Proposition B.2. *Under the assumptions of Theorem 8,*

$$v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) \ge \bar{u}^*(x)$$

Therefore, we have that by the inclusion relationship $\Pi_{\mathrm{H}}(\mathcal{Q}) \supset \Pi_{\mathrm{M}}(\mathcal{Q}) \supset \Pi_{\mathrm{S}}(\mathcal{Q})$, we have

$$\bar{u}^* \ge v^*(\Pi_{\mathrm{H}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) \ge v^*(\Pi_{\mathrm{M}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) \ge v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) \ge \bar{u}^*.$$

So all the quantities above are equal. Similarly,

$$\begin{aligned} u^* = v^*(\Pi_{\mathrm{M}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) &\le v^*(\Pi_{\mathrm{M}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{M}}(\mathcal{P})) \le u^* \\ u^* = v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) &= v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{M}}(\mathcal{P})) \qquad = v^*(\Pi_{\mathrm{S}}(\mathcal{Q}), \overline{\Gamma}_{\mathrm{S}}(\mathcal{P})) \le u^*. \end{aligned}$$

This proves Theorem 8.

B.2.1 Proof of Auxiliary Results for Theorem 8

B.2.2 Proof of Proposition B.1

Fix an arbitrary $\pi = (\pi_0, \pi_1, \cdots) \in \Pi_{\mathrm{U}}(\mathcal{Q})$. It suffice to show that for any $\epsilon > 0$ there exists $\bar{\gamma} \in \overline{\Gamma}_{\mathrm{U}}(\mathcal{P})$ s.t.

$$v(x,\pi,\bar{\gamma}) \le \bar{u}^*(x) + \epsilon. \tag{B.1}$$

Recall from 2.1 that $\|\bar{u}^*\| \le \beta r_\vee$. Define and denote the T-step truncated value with terminal reward $\bar{u}^*$ by

$$v_T(x,\pi,\gamma) := E_x^{\pi,\gamma}\left[\sum_{t=0}^{T_\epsilon - 1} \alpha^t r(X_t, A_t) + \alpha^{T_\epsilon}\bar{u}^*(X_{T_\epsilon})\right]. \tag{B.2}$$

Also, define $T_\eta = \lceil \beta \log(2r_\vee\beta/\eta)\rceil$ where $\beta = \frac{1}{1-\alpha}$. Then,

$$\alpha^{T_\eta} \le \left(1 - \frac{1}{\beta}\right)^{\beta\log(2r_\vee\beta/\eta)} \le \exp(-\log(2r_\vee\beta/\eta)) = \frac{\eta}{2r_\vee\beta}.$$

Thus, consider, for any $\pi \in \Pi_{\mathrm{U}}(\mathcal{Q})$, $\gamma \in \Gamma_{\mathrm{U}}(\mathcal{P})$, and $x \in \mathbb{X}$, we have

$$\begin{aligned} \left|v(x,\pi,\gamma) - v_{T_\eta}(x,\pi,\gamma)\right| &= \left|E_x^{\pi,\gamma}\left[\sum_{t=0}^{\infty}\alpha^t r(X_t,A_t)\right] - E_x^{\pi,\gamma}\left[\sum_{t=0}^{T_\eta-1}\alpha^t r(X_t,A_t) + \alpha^{T_\eta}\bar{u}^*(X_{T_\eta})\right]\right| \\ &= \left|E_x^{\pi,\gamma}\left[\sum_{t=T_\eta}^{\infty}\alpha^t r(X_t,A_t) - \alpha^{T_\eta}\bar{u}^*(X_{T_\eta})\right]\right| \\ &\le \alpha^{T_\eta}\left(\left|E_x^{\pi,\gamma}\sum_{t=T_\eta}^{\infty}\alpha^{t-T_\eta} r(X_t,A_t)\right| + \left|E_x^{\pi,\gamma}\bar{u}^*(X_{T_\eta})\right|\right) \\ &\le \alpha^{T_\eta}\frac{2r_\vee}{1-\alpha} \\ &\le \eta. \end{aligned} \tag{B.3}$$

Next, we focus on the history-dependent case U = H. Since $\bar{u}^*$ is the solution, we must have that for any $h_t \in \mathbf{H}_t$,

$$\inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\psi(dw) \le \bar{u}^*(x_t)$$

Fix any $\delta > 0$. Let $\Psi_t^\delta(h_t)$ be a set of $\psi \in \mathcal{P}$ s.t.

$$\int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\psi(dw) \le \bar{u}^*(x_t)+\delta. \tag{B.4}$$

We want to show that there is a $\mathbf{H}_t \to \mathcal{P}(\mathcal{W})$ measurable selection $\bar{\gamma}_t^\delta(dw|h_t) \in \Psi_t^\delta(h_t)$.

We note that

$$\left\{\psi\in\mathcal{P} : \int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\psi(dw) \le \bar{u}^*(x_t)+\delta\right\}$$

is a closed set. It is clearly nonempty, as $\bar{u}^*$ satisfies (2.3). Now, by the Kuratowski-Ryll-Nardzewski measurable selection theorem, we show that for any open set $\mathcal{D}$ of $\mathcal{P}$, we have that

$$\left\{h_t\in\mathbf{H}_t : \varnothing \neq \left\{\psi\in\mathcal{P} : \int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\psi(dw) \le \bar{u}^*(x_t)+\delta\right\}\cap\mathcal{D}\right\} \in \mathcal{B}(\mathbf{H}_t).$$

Note that,

$$\begin{aligned}
&\varnothing \neq \left\{\psi\in\mathcal{P} : \int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\psi(dw) \le \bar{u}^*(x_t)+\delta\right\}\cap\mathcal{D}\\
\iff &\exists\psi\in\mathcal{D} : \int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\psi(dw) \le \bar{u}^*(x_t)+\delta\\
\iff &\inf_{\psi\in\mathcal{D}} \int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\psi(dw) - \bar{u}^*(x_t)-\delta \le 0
\end{aligned}$$

Recall that $\mathcal{P}(\mathcal{W})$ is endowed with the Lévy–Prokhorov metric. Since $\mathbb{W}$ is separable, so is $\mathcal{P}(\mathcal{W})$. Observe that by bounded convergence and the continuity of f and $\bar{u}^*$, for $w_k \to w$ on $\mathbb{W}$, we have that

$$\lim_{k\to\infty}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w_k))\pi_t(da|h_t) = \int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)$$

i.e. $w \to \int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w_k))\pi_t(da|h_t)$ is bounded continuous. Hence,

$$\psi \to \int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\psi(dw)$$

is continuous. Therefore, we have that

$$c(h_t) := \inf_{\psi\in\mathcal{D}} \int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\psi(dw) - \bar{u}^*(x_t)-\delta$$

is $\mathcal{B}(\mathbf{H}_t) \to \mathbb{R}$ measurable, by the measurability of π_t and that the infimum can be taken over a dense subset. Hence, the sub-level set $c(h_t) \le 0$ is $\mathcal{B}(\mathbf{H}_t)$ measurable.

Therefore, the measurable selection theorem applies and we conclude that there exists $\bar{\gamma}_t^\delta : \mathbf{H}_t \to \mathcal{P}$ measurable s.t.

$$\int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a)+\alpha\bar{u}^*(f(x_t,a,w))\pi_t(da|h_t)\bar{\gamma}_t^\delta(dw|h_t) \le \bar{u}^*(x_t)+\delta.$$

Since this can be done for each t, we can construct $\bar{\gamma}^\delta \in \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})$ s.t. the above inequality holds for each and every t.

Now, we first consider

$$\begin{aligned}
E_x^{\pi,\bar{\gamma}^\delta}[r(X_t,A_t)+\alpha\bar{u}^*(X_{t+1})] &= E_x^{\pi,\bar{\gamma}^\delta}[r(X_t,A_t)+\alpha\bar{u}^*(f(X_t,A_t,W_t))]\\
&= E_x^{\pi,\bar{\gamma}^\delta}E_x^{\pi,\bar{\gamma}^\delta}[r(X_t,A_t)+\alpha\bar{u}^*(f(X_t,A_t,W_t))|H_t]\\
&= E_x^{\pi,\bar{\gamma}^\delta}\int_{\mathbb{W}}\int_{\mathbb{A}} r(X_t,a)+\alpha\bar{u}^*(f(X_t,a,w))\pi_t(da|H_t)\bar{\gamma}_t^\delta(dw|H_t)\\
&\le E_x^{\pi,\bar{\gamma}^\delta}\bar{u}^*(X_t)+\delta.
\end{aligned}$$

Recall the definition (B.2), we have that for any $T \geq 1$,

$$\begin{aligned} v_T(x,\pi,\bar{\gamma}^\delta) &= E_x^{\pi,\bar{\gamma}^\delta}\left[\sum_{t=0}^{T-1}\alpha^t r(X_t,A_t) + \alpha^T \bar{u}^*(X_T)\right] \\ &= E_x^{\pi,\bar{\gamma}^\delta}\left[\sum_{t=0}^{T-2}\alpha^t r(X_t,A_t)\right] + \alpha^{T-1}E_x^{\pi,\bar{\gamma}^\delta}\left[r(X_{T-1},A_{T-1}) + \alpha\bar{u}^*(X_T)\right] \\ &\leq E_x^{\pi,\bar{\gamma}^\delta}\left[\sum_{t=0}^{T-2}\alpha^t r(X_t,A_t)\right] + \alpha^{T-1}E_x^{\pi,\bar{\gamma}^\delta}\left[\bar{u}^*(X_{T-1})\right] + \alpha^{T-1}\delta \\ &\leq v_{T-1}(x,\pi,\bar{\gamma}^\delta) + \alpha^{T-1}\delta. \end{aligned} \tag{B.5}$$

Therefore, by induction on T, we conclude that

$$\begin{aligned} v_T(x,\pi,\bar{\gamma}^\delta) &\leq v_1(x,\pi,\bar{\gamma}^\delta) + \delta\sum_{t=1}^{T-1}\alpha^t \\ &\leq E_x^{\pi,\bar{\gamma}^\delta}[r(X_0,A_0) + \alpha\bar{u}^*(X_1)] - \delta + \beta\delta \\ &\leq \bar{u}^*(x) + \beta\delta. \end{aligned}$$

Since δ is arbitrary, choosing $\delta = \epsilon/(2\beta)$ and $T = T_{\epsilon/2}$, we conclude that by (B.3),

$$\begin{aligned} v(x,\pi,\bar{\gamma}^\delta) &\leq v_{T_{\epsilon/2}}(x,\pi,\bar{\gamma}^\delta) + \frac{\epsilon}{2} \\ &\leq \bar{u}^*(x) + \beta\delta + \frac{\epsilon}{2} \\ &\leq \bar{u}^*(x) + \epsilon; \end{aligned}$$

i.e. inequality (B.1) holds with adversarial policy $\bar{\gamma}^\delta$. This completes the proof for the case U = H.

For cases U = M, S, the proof remains the same except we choose the adversary to be Markov or time-homogeneous, in the presence of a Markov or time-homogeneous controller, respectively. For instance, in the time-homogeneous case, given any $\pi(da|x)$ we choose a policy $\bar{\gamma}^\delta \in \overline{\Gamma}_{\mathrm{S}}(\mathcal{P})$ s.t.

$$\int_{\mathbb{W}}\int_{\mathbb{A}} r(x_t,a) + \alpha\bar{u}^*(f(x_t,a,w))\pi(da|x_t)\bar{\gamma}^\delta(dw|x_t) \leq \bar{u}^*(x_t) + \delta.$$

for every x_t. A measurable choice is always possible because $\bar{\gamma}^\delta(da|x_t)$ has the same information dependence on x_t as $\pi(da|x_t)$.

B.2.3 Proof of Proposition B.2

Since $\bar{u}^*$ is the solution to (2.3), by the same measurable selection argument and separability, for any fixed $\delta > 0$, there exists measurable $\pi^\delta(da|x)$ s.t. $\pi^\delta(da|x) \in \mathcal{Q}$ for all $x \in \mathbb{X}$ and

$$\bar{u}^*(s) \leq \inf_{\psi\in\mathcal{P}}\int_{\mathbb{A}}\int_{\mathbb{W}} r(x,a) + \alpha\bar{u}^*(f(x,a,w))\pi^\delta(da|x)\psi(dw) + \delta. \tag{B.6}$$

Let $\pi^\delta = (\pi^\delta,\pi^\delta,\cdots) \in \Pi_{\mathrm{S}}(\mathcal{Q})$. We consider for any $\bar{\gamma} = (\bar{\gamma}_0,\bar{\gamma}_1,\cdots) \in \overline{\Gamma}_{\mathrm{H}}(\mathcal{P})$,

$$\begin{aligned} E_\mu^{\pi^\delta,\bar{\gamma}}[u^*(X_{t+1})] &= E_\mu^{\pi^\delta,\bar{\gamma}}E_\mu^{\pi^\delta,\bar{\gamma}}[\bar{u}^*(f(X_t,A_t,W_t))|G_t] \\ &= E_\mu^{\pi^\delta,\bar{\gamma}}\int_{\mathbb{W}}\bar{u}^*(f(X_t,A_t,w))\bar{\gamma}_t(dw|G_t) \\ &= E_\mu^{\pi^\delta,\bar{\gamma}}\int_{\mathbb{A}}\int_{\mathbb{W}}\bar{u}^*(f(X_t,a,w))\bar{\gamma}_t(dw|H_t,a)\pi^\delta(da|X_t) \\ &= E_\mu^{\pi^\delta,\bar{\gamma}}\int_{\mathbb{A}}\int_{\mathbb{W}}\bar{u}^*(f(X_t,a,w))\bar{\gamma}_t(dw|G_{t-1},X_t,a)\pi^\delta(da|X_t) \end{aligned}$$

Given G_{t-1} and X_t,

$$\int_{\mathbb{A}}\int_{\mathbb{W}} \bar{u}^*(f(X_t,a,w))\bar{\gamma}_t(dw|G_{t-1},X_t,a)\pi^\delta(da|X_t) \geq \inf_{\psi\in\mathcal{P}}\int_{\mathbb{A}}\int_{\mathbb{W}} \bar{u}^*(f(X_t,a,w))\psi(dw)\pi^\delta(da|X_t)$$

Therefore, by (B.6)

$$\begin{aligned}
&E_\mu^{\pi^\delta,\bar{\gamma}}\left[r(X_t,A_t)+\alpha u^*(X_{t+1})\right]\\
&\geq E_\mu^{\pi^\delta,\bar{\gamma}}\left[\int_{\mathbb{A}} r(X_t,a)\pi^\delta(da|X_t)+\alpha\inf_{\psi\in\mathcal{P}}\int_{\mathbb{A}}\int_{\mathbb{W}}\bar{u}^*(f(X_t,a,w))\psi(dw)\pi^\delta(da|X_t)\right]\\
&= E_\mu^{\pi^\delta,\bar{\gamma}}\left[\inf_{\psi\in\mathcal{P}}\int_{\mathbb{A}}\int_{\mathbb{W}} r(X_t,a)+\alpha\bar{u}^*(f(X_t,a,w))\psi(dw)\pi^\delta(da|X_t)\right]\\
&\geq E_\mu^{\pi^\delta,\bar{\gamma}}\bar{u}^*(X_t)-\delta.
\end{aligned}$$

By the same technique as in (B.5), we see that for all T,

$$v_T(x,\pi^\delta,\bar{\gamma}) \geq \bar{u}^*(x)-\beta\delta$$

uniformly in x.

Since $\pi\in\Pi_{\mathrm{S}}(\mathcal{Q})$,

$$\begin{aligned}
v^*(\Pi_{\mathrm{S}}(\mathcal{Q}),\overline{\Gamma}_{\mathrm{H}}(\mathcal{P})) &\geq \inf_{\bar{\gamma}\in\overline{\Gamma}_{\mathrm{H}}(\mathcal{P})} v(\pi^\delta,\bar{\gamma})\\
&\geq \inf_{\bar{\gamma}\in\overline{\Gamma}_{\mathrm{H}}(\mathcal{P})} v_{T_\epsilon}(\pi^\delta,\bar{\gamma})-\epsilon\\
&\geq \bar{u}^*-\beta\delta-\epsilon
\end{aligned}$$

by (B.3). Since ϵ and δ are arbitrary, we complete the proof.

C Proofs for Section 3

C.1 Proof of Proposition 3.1

Define a sequence of functions $u_0\equiv 0$ and $u_{k+1}=\mathbf{T}'(u_k)$. By the Banach fixed point theorem, $u_k\to\hat{u}$ in uniform norm. Since $u'=\mathcal{T}'(u')$, the error

$$\begin{aligned}
\Delta_{k+1} &:= u_{k+1}-u'\\
&= \mathbf{T}'(u_k)-\mathbf{T}'(u')+\mathbf{T}'(u')-\mathcal{T}'(u')\\
&= [\mathbf{T}'(u'+\Delta_k)-\mathbf{T}'(u')]+[\mathbf{T}'(u')-\mathcal{T}'(u')]\\
&=: \mathbf{H}(\Delta_k)+U.
\end{aligned}$$

By Proposition 2.1, it is easy to see that $\mathbf{H}$ is also a α-contraction on $U_b(\mathbb{X})$, with 0 as its unique fixed point.

We claim that for $k\geq 1$,

$$\|\Delta_k\| \leq \beta\alpha^{k-1}+\sum_{j=0}^{k-1}\alpha^j\|U\|.$$

We check this by induction: for $k=1$,

$$\begin{aligned}
\|\Delta_1\| &\leq \|\mathbf{H}(\Delta_0)\|+\|U\|\\
&= \|\mathbf{H}(\Delta_0)-\mathbf{H}(0)\|+\|U\|\\
&\leq \alpha\|u'\|+\|U\|\\
&\leq \alpha\beta+\|U\|.
\end{aligned}$$

For the induction step, we have that

$$\begin{aligned}\|\Delta_{k+1}\| &\le \|\mathbf{H}(\Delta_k) - \mathbf{H}(0)\| + \|U\| \\ &\le \alpha \|\Delta_k\| + \|U\| \\ &\le \beta\alpha^k + \sum_{j=0}^{k} \alpha^j \|U\|.\end{aligned}$$

where the last inequality follows from the induction assumption.

Therefore,

$$\|\hat{u} - u'\| = \lim_{k\to\infty} \|\Delta_k\| \le \sum_{j=0}^{\infty} \alpha^j \|U\| = \beta \|\mathbf{T}'(u') - \mathcal{T}'(u')\|.$$

C.2 Proof of Theorem 1

We remark that our proof techniques have similarities with that in Lee and Raginsky (2018). By Proposition 3.1, to achieve an upper bound on the uniform learning error, it suffices to prove an upper bound for $\|\mathbf{T}(u^*) - \mathcal{T}(u^*)\|$.

To do this, we first rewrite the Bellman operator using its dual form. By strong duality (Blanchet and Murthy, 2019), for $\mathcal{P} = \{\mu \in \mathcal{P}(\mathcal{X}) : W(\mu, \mu_0) \le \delta\}$

$$\inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} u^*(f(z,w))\psi(dw) = \sup_{\lambda\ge 0} -\lambda\delta + \int_{\mathbb{X}} \inf_{y\in\mathbb{W}} \left[u^*(f(z,y)) + \lambda c(w,y)\right] \mu_0(dw).$$

Notice that since

$$\int_{\mathbb{X}} \inf_{y\in\mathbb{W}} \left[u^*(f(z,y)) + \lambda c(w,y)\right] \mu_0(dw) \le \int_{\mathbb{X}} u^*(f(z,w))\mu_0(dw) \le \|u\|$$

and

$$\inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} u^*(f(z,w))\psi(dw) \ge 0,$$

it suffices to maximize λ over $\Lambda := [0, \delta^{-1}\|u^*\|]$.

Therefore, we have that

$$\begin{aligned}&\|\mathbf{T}(u^*) - \mathcal{T}(u^*)\| \\ &\le \alpha \sup_{z\in\mathbb{X}\times\mathbb{A}} \left| \sup_{\lambda\in\Lambda} \left[\int_{\mathbb{X}} \inf_{y\in\mathbb{W}} \left[u^*(f(z,y)) + \lambda c(w,y)\right] \mu_0(dw) - \lambda\delta \right] - \sup_{\lambda\in\Lambda} \left[\int_{\mathbb{X}} \inf_{y\in\mathbb{W}} \left[u^*(f(z,y)) + \lambda c(w,y)\right] \hat{\mu}(dw) - \lambda\delta \right] \right| \\ &\le \alpha \sup_{z\in\mathbb{X}\times\mathbb{A}} \sup_{\lambda\in\Lambda} \left| \int_{\mathbb{X}} \inf_{y\in\mathbb{W}} \left[u^*(f(z,y)) + \lambda c(w,y)\right] (\mu_0 - \hat{\mu})(dw) \right| \\ &= \alpha \sup_{\theta\in\Theta} |(\mu_0 - \hat{\mu})[g_\theta]|\end{aligned}$$

where $\Theta = \{\theta = (z,\lambda) : z \in \mathbb{X}\times\mathbb{A}, \lambda \in \Lambda\}$ and

$$g_\theta = \inf_{y\in\mathbb{W}} \left[u^*(f(z,y)) + \lambda c(\cdot, y)\right].$$

Therefore, the estimation error is bounded by a supremum of empirical process.

To bound this, we use the Rademacher process and a chaining argument. Specifically, for fixed sequence $\mathbf{w} := \{w_i \in \mathbb{W} : i = 1, 2, \cdots, n\}$, we define the Rademacher process indexed by $g_\theta \in \mathcal{G}$ as

$$R_n(\mathbf{w}, g_\theta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i g_\theta(w_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \inf_{w\in\mathbb{W}} \left[u^*(f(z,y)) + \lambda c(w_i, y)\right]. \tag{C.1}$$

The empirical and population Rademacher complexities of the function class $\mathcal{G} := \{g_\theta : \theta \in \Theta\}$

$$\mathcal{R}_n(\mathbf{w}, \mathcal{G}) := E_\epsilon \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} R_n(\mathbf{w}, g), \quad \text{and} \quad \mathcal{R}_n(\mathcal{G}) := E_{\mu_0^n} \mathcal{R}_n(\mathbf{W}, \mathcal{G}) \tag{C.2}$$

where $\mu_0^n = \mu_0 \times \cdots \times \mu_0$ the n-fold product measure, and $W = (W_1, \cdots, W_n)$.

From empirical process theory, see e.g. Wainwright (2019, Theorem 4.10), w.p. at least $1-\eta$,

$$\sup_{g\in\mathcal{G}} |(\mu_n - \mu_0)[g]| \le 2\mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{2g_\vee}{n}\log\left(\frac{1}{\eta}\right)}. \tag{C.3}$$

where $g_\vee := \sup_{g\in\mathcal{G}} \|g\| \le \|u^*\| \le \beta$. Thus, we proceed to bound $\mathcal{R}_n(\mathbf{w},\mathcal{G})$ and hence $\mathcal{R}_n(\mathcal{G})$. We achieve this by using subgaussian processes and entropy integrals.

Specifically, we consider the moment generating function of the Rademacher process (C.1). For ξ in some neighborhood of the origin

$$\begin{aligned}
&E_\epsilon \exp[\xi(R_n(\mathbf{w},g_\theta) - R_n(\mathbf{w},g_{\theta'}))] \\
&= E_\epsilon \exp\left(\frac{\xi}{\sqrt{n}}\sum_{i=1}^n \epsilon_i \left[\inf_{y\in\mathbb{W}}[u^*(f(z,y)) + \lambda c(w_i,y)] - \inf_{y\in\mathbb{W}}[u^*(f(z',y)) + \lambda' c(w_i,y)]\right]\right) \\
&\le \exp\left(\frac{\xi^2}{2n}\sum_{i=1}^n \left[\inf_{y\in\mathbb{W}}[u^*(f(z,y)) + \lambda c(w_i,y)] - \inf_{y\in\mathbb{W}}[u^*(f(z',y)) + \lambda' c(w_i,y)]\right]^2\right) \\
&\le \exp\left(\frac{\xi^2}{2n}\sum_{i=1}^n \sup_{y\in\mathbb{W}} |u^*(f(z,y)) - u^*(f(z',y)) + (\lambda'-\lambda)c(w_i,y)|^2\right) \\
&\overset{(i)}{\le} \exp\left(\frac{\xi^2}{2}\left(\sup_{y\in\mathbb{W}} |u^*(f(z,y)) - u^*(f(z',y))| + |\lambda-\lambda'|\, c_\vee\right)^2\right) \\
&\overset{(ii)}{\le} \exp\left(\frac{\xi^2}{2}\left(\ell(|x-x'| + |a-a'|) + |\lambda-\lambda'|\, c_\vee\right)^2\right)
\end{aligned}$$

where (i) uses the transport cost being bounded by $c_\vee$ and (ii) follows from the uniform Lipschitzness in Assumption 2. Therefore, defining

$$\rho(\theta,\theta') := \ell(|x-x'| + |a-a'|) + c_\vee|\lambda-\lambda'|,$$

which is a distance on Θ, we obtain that

$$E_\epsilon \exp[\xi(R_n(\mathbf{w},g_\theta) - R_n(\mathbf{w},g_{\theta'}))] \le \exp\left(\frac{\xi^2}{2}\rho^2(g_\theta,g_{\theta'})\right).$$

This shows that the stochastic process $\{R_n(\mathbf{w},g_\theta) : \theta\in\Theta\}$ is subgaussian w.r.t. ρ.

Therefore, using Dudley's entropy integral for subgaussian processes (Wainwright, 2019, Chapter 5), the empirical Rademacher complexity in (C.2) can be bounded by

$$\begin{aligned}
\mathcal{R}_n(\mathbf{w},\mathcal{G}) &= E_\epsilon \sup_{\theta\in\Theta} \frac{1}{\sqrt{n}} R_n(\mathbf{w},g_\theta) \\
&\le \frac{32}{\sqrt{n}}\int_0^{D_\vee} \sqrt{\log \mathcal{N}(\epsilon;\Theta,\rho)}\, d\epsilon
\end{aligned} \tag{C.4}$$

w.p.1., where $\mathcal{N}(\epsilon;\Theta,\rho)$ is the ϵ covering number of Θ in distance ρ and

$$\begin{aligned}
D_\vee &:= \ell(\mathrm{diam}(\mathbb{X}) + \mathrm{diam}(\mathbb{A})) + c_\vee\delta^{-1}\beta + 1 \\
&\ge \ell(\mathrm{diam}(\mathbb{X}) + \mathrm{diam}(\mathbb{A})) + c_\vee \mathrm{diam}(\Lambda) + 1 \\
&\ge \sup_{\theta,\theta'\in\Theta} \rho(\theta,\theta')
\end{aligned}$$

is an upper bound on the diameter of Θ in terms of ρ.

Note that as the r.h.s. of (C.4) is deterministic, we take expectation over $\mathbf{W}\sim\mu_0^n$ to conclude that the population Rademacher complexity

$$\mathcal{R}_n(\mathcal{G}) = E_{\mu_0^n}\mathcal{R}_n(\mathbf{W},\mathcal{G}) \le \frac{32}{\sqrt{n}}\int_0^{D_\vee} \sqrt{\log \mathcal{N}(\epsilon;\Theta,\rho)}\, d\epsilon$$

satisfying the same bound. Moreover, the covering number

$$\begin{aligned}\mathcal{N}(\epsilon;\Theta,\rho) &= \mathcal{N}(\epsilon;\mathbb{X}\times\mathbb{A},\ell|\cdot|)\cdot\mathcal{N}(\epsilon;\Lambda,c_\vee|\cdot|)\\ &= \mathcal{N}(\epsilon;\mathbb{X},\ell|\cdot|)\cdot\mathcal{N}(\epsilon;\mathbb{A},\ell|\cdot|)\cdot\mathcal{N}(\epsilon;\Lambda,c_\vee|\cdot|)\\ &\le \left(1+\frac{\ell\mathrm{diam}(\mathbb{X})}{\epsilon}\right)^{d_\mathbb{X}}\left(1+\frac{\ell\mathrm{diam}(\mathbb{A})}{\epsilon}\right)^{d_\mathbb{A}}\left(1+\frac{c_\vee\mathrm{diam}(\Lambda)}{\epsilon}\right).\end{aligned}$$

Therefore, the entropy integral

$$\begin{aligned}\int_0^{D_\vee}\sqrt{\log\mathcal{N}(\epsilon;\Theta,\rho)}d\epsilon &\le \int_0^{D_\vee}\sqrt{(d_\mathbb{X}+d_\mathbb{A}+1)\log\left(1+\max\left\{\frac{\ell\mathrm{diam}(\mathbb{A})}{\epsilon},\frac{\ell\mathrm{diam}(\mathbb{X})}{\epsilon},\frac{c_\vee\mathrm{diam}(\Lambda)}{\epsilon}\right\}\right)}d\epsilon\\ &\le \int_0^{D_\vee}\sqrt{(d_\mathbb{X}+d_\mathbb{A}+1)\log(D_\vee/\epsilon)}d\epsilon\\ &= \frac{\sqrt{\pi}}{2}D_\vee\sqrt{d_\mathbb{X}+d_\mathbb{A}+1}\end{aligned}$$

We conclude that by Proposition 3.1 and (C.3), the estimation error

$$\begin{aligned}\|\hat{u}-u^*\| &\le \beta\left\|\mathbf{T}(u^*)(x)-\mathcal{T}(u^*)(x)\right\|\\ &\le \alpha\beta\sup_{\theta\in\Theta}|(\mu_n-\mu_0)[g_\theta]|\\ &\overset{(i)}{\le} 2\alpha\beta\mathcal{R}_n(\mathcal{G})+\alpha\beta\sqrt{\frac{2g_\vee}{n}\log\left(\frac{1}{\eta}\right)}\\ &\le \left(32\sqrt{\pi}\alpha\beta D_\vee\sqrt{d_\mathbb{X}+d_\mathbb{A}+1}+\sqrt{2}\alpha\beta^{3/2}\sqrt{\log\left(\frac{1}{\eta}\right)}\right)n^{-\frac{1}{2}}\end{aligned}$$

where (i) holds w.p. at least $1-\eta$. This bound implies the theorem as we note that $d_\mathbb{X}\ge 1$.

C.3 Proof of Theorem 2

As in the previous proof, we bound

$$\|\mathbf{T}(u^*)-\mathcal{T}(u^*)\|\le\alpha\sup_{z\in\mathbb{X}\times\mathbb{A}}\left|\inf_{\psi\in\mathcal{P}}\int_\mathbb{W}u^*(f(z,w))\psi(dw)-\inf_{\psi\in\widehat{\mathcal{P}}}\int_\mathbb{W}u^*(f(z,w))\psi(dw)\right|$$

Define the function class $\mathcal{F}:=\{u^*(f(z,\cdot)):z\in\mathbb{X}\times\mathbb{A})\}$. By Duchi and Namkoong (2021, Corollary 1), the r.h.s. satisfies w.p. at least $1-2\mathcal{N}(\epsilon/3;\mathcal{F},\|\cdot\|)e^{-t}$,

$$\sup_{z\in\mathbb{X}\times\mathbb{A}}\left|\inf_{\psi\in\mathcal{P}}\int_\mathbb{W}u^*(f(x,a,w))\psi(dw)-\inf_{\psi\in\widehat{\mathcal{P}}}\int_\mathbb{W}u^*(f(x,a,w))\psi(dw)\right|\le 30\beta\epsilon$$

where

$$\epsilon=n^{-\frac{1}{k'\vee 2}}c_k(\delta)^2\left(\frac{c_k(\delta)}{c_k(\delta)-1}\vee 2\right)\left(\frac{1}{k}+\sqrt{t+2\log n}\right).$$

Therefore, choosing $t=\log(2\mathcal{N}(\epsilon/3;\mathcal{F},\|\cdot\|)/\eta)$, we have that w.p. at least $1-\eta$

$$\|\mathbf{T}(u^*)-\mathcal{T}(u^*)\|\le 30\alpha\beta n^{-\frac{1}{k'\vee 2}}c_k(\delta)^2\left(\frac{c_k(\delta)}{c_k(\delta)-1}\vee 2\right)\left(\frac{1}{k}+\sqrt{\log(2\mathcal{N}(\epsilon/3;\mathcal{F},\|\cdot\|)+\log\frac{1}{\eta}+2\log n}\right).$$

By the uniform Lipschitz assumption of $z\to u^*(f(z,\cdot))$ and that $\epsilon\ge n^{-1/2}$ for all $t\ge 1$, we have that by van der Vaart and Wellner (1996, Chapter 2.7.4)

$$\begin{aligned}\log\mathcal{N}(\epsilon/3;\mathcal{F},\|\cdot\|) &\le \log\mathcal{N}\left(\frac{\epsilon\wedge 1}{3\ell};\mathbb{X}\times\mathbb{A},|\cdot|\right)\\ &\le d_\mathbb{X}\log\left(1+\frac{3\ell\mathrm{diam}(\mathbb{X})}{\epsilon\wedge 1}\right)+d_\mathbb{A}\log\left(1+\frac{3\ell\mathrm{diam}(\mathbb{A})}{\epsilon\wedge 1}\right)\\ &\le d_\mathbb{X}\log\left(1+3\ell\mathrm{diam}(\mathbb{X})\right)+d_\mathbb{A}\log\left(1+3\ell\mathrm{diam}(\mathbb{A})\right)+\frac{1}{2}(d_\mathbb{X}+d_\mathbb{A})\log n.\end{aligned}$$

Therefore, defining $D = d_{\mathbb{X}} \log(1 + 3\ell \mathrm{diam}(\mathbb{X})) + d_{\mathbb{A}} \log(1 + 3\ell \mathrm{diam}(\mathbb{A}))$, we conclude that by Proposition 3.1, the estimation error

$$\begin{aligned}\|\hat{u} - u^*\| &\le \beta \|\mathbf{T}(u^*)(x) - \mathcal{T}(u^*)(x)\| \\ &\overset{(i)}{\le} 30\beta^2 n^{-\frac{1}{k' \vee 2}} c_k(\delta)^2 \left(\frac{c_k(\delta)}{c_k(\delta) - 1} \vee 2\right) \left(\frac{1}{k} + \sqrt{D + \log\frac{1}{\eta} + 2(d_{\mathbb{X}} + d_{\mathbb{A}}) \log n}\right)\end{aligned}$$

where (i) holds w.p. at least $1 - \eta$.

C.4 Proof of Theorem 3

In this proof, we first consider general Polish action space and then specialize to finite action space to achieve $n^{-1/2}$ rate. Through the proof, we identify possible structures of the controller's decision space $\mathcal{Q}$ so that

We employ the same proof strategy as that of Theorem 1 in Appendix C.2. By the strong duality, positivity, and Bellman equation (2.3), we have that

$$\begin{aligned}\bar{u}^*(x) &= \sup_{\phi \in \mathcal{Q}} \int_{\mathbb{A}} r(x,a)\phi(da) + \alpha \inf_{\psi \in \mathcal{P}} \int_{\mathbb{W}} \int_{\mathbb{A}} \bar{u}^*(f(x,a,w))\phi(da)\psi(dw) \\ &= \sup_{\phi \in \mathcal{Q}} \int_{\mathbb{A}} r(x,a)\phi(da) + \sup_{\lambda \ge 0} -\lambda\delta + \int_{\mathbb{W}} \inf_{y \in \mathbb{W}} \left[\int_{\mathbb{A}} \bar{u}^*(f(x,a,y))\phi(da) + \lambda c(w,y)\right] \mu_0(dw)\end{aligned}$$

By the same argument, the supremum is achieved within $\Lambda := [0, \delta^{-1} \|\bar{u}^*\|]$. So, we have that

$$\begin{aligned}\left\|\overline{\mathbf{T}}(\bar{u}^*)(x) - \mathcal{T}(\bar{u}^*)(x)\right\| &\le \alpha \sup_{x \in \mathbb{X}, \phi \in \mathcal{Q}} \left| \inf_{\psi \in \widehat{\mathcal{P}}} \int_{\mathbb{W}} \int_{\mathbb{A}} \bar{u}^*(f(x,a,w))\phi(da)\psi(dw) - \inf_{\psi \in \mathcal{P}} \int_{\mathbb{W}} \int_{\mathbb{A}} \bar{u}^*(f(x,a,w))\phi(da)\psi(dw) \right| \\ &\le \alpha \sup_{x \in \mathbb{X}, \phi \in \mathcal{Q}, \lambda \in \Lambda} \left| \int_{\mathbb{W}} \inf_{y \in \mathbb{W}} \left[\int_{\mathbb{A}} \bar{u}^*(f(x,a,y))\phi(da) + \lambda c(w,y)\right] (\mu_0 - \hat{\mu})(dw) \right| \\ &=: \sup_{g \in \mathcal{G}} |(\mu_0 - \hat{\mu})[g]|\end{aligned}$$

Here, the parametric function class $\mathcal{G}$ is characterized by $(x, \psi, \lambda) \in \Theta = \mathbb{X} \times \mathcal{Q} \times \Lambda$ and

$$\mathcal{G} := \left\{ w \to \inf_{y \in \mathbb{W}} \left[\int_{\mathbb{A}} \bar{u}^*(f(x,a,y))\phi(da) + \lambda c(w,y)\right] : (x, \psi, \lambda) \in \Theta \right\}.$$

To bound the previous empirical process supremum, we still employ the Rademacher complexity bound as in (C.3). In this case, for $g \in \mathcal{G}$ and sequence $\mathbf{w} := \{w_i \in \mathbb{W} : i = 1, 2, \cdots, n\}$, the Rademacher process is

$$R_n(\mathbf{w}, g) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i g(w_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i \inf_{y \in \mathbb{W}} \left[\int_{\mathbb{A}} \bar{u}^*(f(x,a,y))\phi(da) + \lambda c(w_i,y)\right],$$

compare to (C.1), and the empirical and population complexities are defined as in (C.2) accordingly. We then consider the moment generating function: for ξ in some neighborhood of the origin

$$\begin{aligned}&E_\epsilon \exp[\xi(R_n(\mathbf{w}, g_\theta) - R_n(\mathbf{w}, g_{\theta'}))] \\ &= E_\epsilon \exp\left(\frac{\xi}{\sqrt{n}} \sum_{i=1}^n \epsilon_i [g_\theta(w_i) - g_{\theta'}(w_i)]\right) \\ &\le \exp\left(\frac{\xi^2}{2n} \sum_{i=1}^n [g_\theta(w_i) - g_{\theta'}(w_i)]^2\right) \\ &\le \exp\left(\frac{\xi^2}{2n} \sum_{i=1}^n \sup_{y \in \mathbb{W}} \left| \int_{\mathbb{A}} \bar{u}^*(f(x,a,w))\phi(da) - \int_{\mathbb{A}} \bar{u}^*(f(x',a,w))\phi'(da) + (\lambda - \lambda')c(w_i,y) \right|^2\right) \\ &\le \exp\left(\frac{\xi^2}{2} \left[\sup_{y \in \mathbb{W}} \left| \int_{\mathbb{A}} \bar{u}^*(f(x,a,y))\phi(da) - \int_{\mathbb{A}} \bar{u}^*(f(x',a,y))\phi'(da) \right| + c_\vee |\lambda - \lambda'| \right]^2\right)\end{aligned}$$

Consider

$$\begin{aligned}&\left|\int_{\mathbb{A}} \bar{u}^*(f(x,a,y))\phi(da) - \int_{\mathbb{A}} \bar{u}^*(f(x',a,y))\phi'(da)\right| \\ &\le \left|\int_{\mathbb{A}} \bar{u}^*(f(x,a,w)) - u^*(f(x',a,w))\phi(da)\right| + \left|\int_{\mathbb{A}} \bar{u}^*(f(x',a,w))[\phi-\phi'](da)\right| \\ &\le \ell|x-x'| + \min\left\{\beta\left\|\phi-\phi'\right\|_{\mathrm{TV}}, \ell W_1(\phi,\phi')\right\}\end{aligned} \tag{C.5}$$

Remark. As we will easily see from the reset of the proof, if $\mathcal{Q}$ is set of measures with ϵ covers of cardinality $O(\epsilon^{-d})$ in either W_1 or total variation distance, for example $\mathcal{Q}$ is a set of smoothly parameterized set of measures or $|\mathbb{A}| < \infty$, then the entropy integral will be finite, yielding a $n^{-1/2}$ convergence rate. However, in the following development, we will focus on the case where $|\mathbb{A}| < \infty$ to get concrete dependencies on the dimensions, diameters, and the size of the action space.

With $|\mathbb{A}| < \infty$, we conclude that

$$E_\epsilon \exp[\xi(R_n(\mathbf{u}, g_\theta) - R_n(\mathbf{u}, g_{\theta'}))] \le \exp\left(\frac{\xi^2}{2}\rho(\theta,\theta')\right)$$

where

$$\rho(\theta,\theta') := \ell|x-x'| + \beta\left\|\phi-\phi'\right\|_{\mathrm{TV}} + c_\vee|\lambda-\lambda'|$$

which is a distance on $\mathbb{X}\times\mathcal{Q}\times\Lambda$. This shows that the process $\{R_n(\mathbf{u}, g_\theta) : \theta\in\Theta\}$ is subgaussian w.r.t. ρ.

Therefore, using Dudley's entropy integral (Wainwright, 2019, Chapter 5), the empirical Rademacher complexity can be bounded

$$\mathcal{R}_n(\mathbf{w},\mathcal{G}) \le \frac{32}{\sqrt{n}}\int_0^{D_\vee}\sqrt{\log\mathcal{N}(\epsilon;\Theta,\rho)}d\epsilon$$

w.p.1., where

$$\overline{D}_\vee := \ell\mathrm{diam}(\mathbb{X}) + 2\beta + c_\vee\delta^{-1}\beta + 1 \ge \sup_{g,g'\in\mathcal{G}}\rho(g,g').$$

In particular, as the r.h.s. is deterministic, we take expectation over $\mathbf{W} = \{W_i : i = 1,\cdots,n\} \sim \mu_0^n$ to conclude that the population Rademacher complexity

$$\mathcal{R}_n(\mathcal{G}) \le \frac{32}{\sqrt{n}}\int_0^{D_\vee}\sqrt{\log\mathcal{N}(\epsilon;\Theta,\rho)}d\epsilon$$

satisfying the same bound. Moreover, the covering number

$$\begin{aligned}\mathcal{N}(\epsilon;\Theta,\rho) &\le \mathcal{N}(\epsilon;\mathbb{X},\ell|\cdot|)\cdot\mathcal{N}(\epsilon;\mathcal{Q},\beta\left\|\cdot\right\|_{\mathrm{TV}})\cdot\mathcal{N}(\epsilon;\Lambda,c_\vee|\cdot|) \\ &\le \mathcal{N}(\epsilon;\mathbb{X},\ell|\cdot|)\cdot\mathcal{N}(\epsilon;B_1^{|\mathbb{A}|},\beta\left\|\cdot\right\|_1)\cdot\mathcal{N}(\epsilon;\Lambda,c_\vee|\cdot|) \\ &\le \left(1+\frac{\ell\mathrm{diam}(\mathbb{X})}{\epsilon}\right)^{d_{\mathbb{X}}}\left(1+\frac{2\beta}{\epsilon}\right)^{|\mathbb{A}|}\left(\frac{c_\vee\mathrm{diam}(\Lambda)}{\epsilon}+1\right).\end{aligned} \tag{C.6}$$

where $B_1^{|\mathbb{A}|}$ is the $|\mathbb{A}|$ dimensional ℓ_1-ball of radius 1, and its covering number bound follows from Wainwright (2019, Example 5.8). Therefore, the entropy integral

$$\begin{aligned}\int_0^{\overline{D}_\vee}\sqrt{\log\mathcal{N}(\epsilon;\Theta,\rho)}d\epsilon &\le \int_0^{\overline{D}_\vee}\sqrt{(d_{\mathbb{X}}+|\mathbb{A}|+1)\log(\overline{D}_\vee/\epsilon)}d\epsilon \\ &= \frac{\sqrt{\pi}}{2}\overline{D}_\vee\sqrt{d_{\mathbb{X}}+|\mathbb{A}|+1}\end{aligned}$$

Therefore, we conclude that by Proposition 3.1 and the Rademacher complexity bound (C.3), the estimation error

$$\begin{aligned}\|\hat{u}-\bar{u}^*\| &\le \beta\left\|\overline{\mathbf{T}}(\bar{u}^*) - \overline{\mathcal{T}}(\bar{u}^*)\right\| \\ &\overset{(i)}{\le} 2\alpha\beta\mathcal{R}_n(\mathcal{G}) + \alpha\beta\sqrt{\frac{2g_\vee}{n}\log\frac{1}{\eta}} \\ &\le \left(32\sqrt{\pi}\alpha\beta\overline{D}_\vee\sqrt{d_{\mathbb{X}}+|\mathbb{A}|+1} + \sqrt{2}\alpha\beta^{3/2}\sqrt{2\log\frac{1}{\eta}}\right)n^{-1/2}\end{aligned}$$

where (i) holds w.p. at least $1-\eta$. This bound implies the theorem as we note that $d_{\mathbb{X}} \ge 1$.

C.5 Proof of Theorem 4

Again, we have that

$$\left\|\overline{\mathbf{T}}(\bar{u}^*)(x) - \mathcal{T}(\bar{u}^*)(x)\right\| \le \alpha \sup_{x\in\mathbb{X},\phi\in\mathcal{Q}} \left| \inf_{\psi\in\widehat{\mathcal{P}}} \int_{\mathbb{W}} \int_{\mathbb{A}} \bar{u}^*(f(x,a,w))\phi(da)\psi(dw) - \inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} \int_{\mathbb{A}} \bar{u}^*(f(x,a,w))\phi(da)\psi(dw) \right|$$
$$= \alpha \sup_{\theta\in\Theta} \left| \inf_{\psi\in\widehat{\mathcal{P}}} \int_{\mathbb{W}} g_\theta(w)\psi(dw) - \inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} g_\theta(w)\psi(dw) \right|$$

where $\Theta = \mathbb{X}\times\mathcal{Q}$ and for $\theta = (x,\phi)$,

$$g_\theta(w) := \int_{\mathbb{A}} \bar{u}^*(f(x,a,w))\phi(da).$$

By Duchi and Namkoong (2021, Corollary 1), for $n \ge k\vee 3$, the r.h.s. satisfies w.p. at least $1-2\mathcal{N}(\epsilon/3;\mathcal{G},\|\cdot\|)e^{-t}$,

$$\sup_{g\in\mathcal{G}} \left| \inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} g(w)\psi(dw) - \inf_{\psi\in\widehat{\mathcal{P}}} g(w)\psi(dw) \right| \le 30\beta\epsilon$$

where $\mathcal{G} := \{g_\theta : \theta\in\Theta\}$ and

$$\epsilon = n^{-\frac{1}{k'\vee 2}} c_k(\delta)^2 \left(\frac{c_k(\delta)}{c_k(\delta)-1}\vee 2\right)\left(\frac{1}{k} + \sqrt{t+2\log n}\right).$$

Therefore, choosing $t = \log(2\mathcal{N}(\epsilon/3;\mathcal{G},\|\cdot\|)/\eta)$, we have that w.p. at least $1-\eta$

$$\left\|\overline{\mathbf{T}}(\bar{u}^*) - \overline{\mathcal{T}}(\bar{u}^*)\right\| \le 30\alpha\beta n^{-\frac{1}{k'\vee 2}} c_k(\delta)^2 \left(\frac{c_k(\delta)}{c_k(\delta)-1}\vee 2\right)\left(\frac{1}{k} + \sqrt{\log(2\mathcal{N}(\epsilon/3;\mathcal{G},\|\cdot\|) + \log\frac{1}{\eta} + 2\log n}\right).$$

To bound the covering number, we recall (C.5). Again, we can generalize to continuum settings. However, we focus on the finite action setting in this paper. In this case, we have that by (C.5), $\theta\to g_\theta(w)$ is uniformly 1-Lipschitz in the distance

$$d((x,\phi),(x',\phi')) = \ell|x-x'| + \beta\left\|\phi-\phi'\right\|_{\mathrm{TV}}.$$

This and the Lipschitz covering number bound (van der Vaart and Wellner, 1996, Chapter 2.7.4) implies that

$$\log\mathcal{N}(\epsilon/3;\Theta,\rho) \le d_{\mathbb{X}}\log\left(1+\frac{3\ell\mathrm{diam}(\mathbb{X})}{\epsilon\wedge 1}\right) + |\mathbb{A}|\log\left(1+\frac{6\beta}{\epsilon\wedge 1}\right)$$
$$\le d_{\mathbb{X}}\log\left(1+3\ell\mathrm{diam}(\mathbb{X})\right) + |\mathbb{A}|\log\left(1+6\beta\right) + \frac{1}{2}\left(d_{\mathbb{X}}+|A|\right)\log n$$

where we handle the covering number of probability measures on $(\mathbb{A},\|\cdot\|_{\mathrm{TV}})$ the same way as in (C.6) and the last inequality uses $\epsilon \ge n^{-1/2}$.

Therefore, defining $\overline{D} := d_{\mathbb{X}}\log\left(1+3\ell\mathrm{diam}(\mathbb{X})\right) + |\mathbb{A}|\log\left(1+6\beta\right)$, we conclude that

$$\|\hat{u}-u^*\| \le 30\beta^2 n^{-\frac{1}{k'\vee 2}} c_k(\delta)^2 \left(\frac{c_k(\delta)}{c_k(\delta)-1}\vee 2\right)\left(\frac{1}{k} + \sqrt{\overline{D} + \log\frac{1}{\eta} + 2(d_{\mathbb{X}}+|\mathbb{A}|)\log n}\right)$$

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

D Proofs for Section 4

D.1 Proof of Lemma 1

Since f and r doesn't depend on a, we have that

$$\bar{u}^*(x) = u^*(x) = x + \alpha \inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} u^*(w)\psi(dw)$$

We guess that $u(x) = x + c$ is the unique solution, and define $c' = \inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} w\psi(dw)$. Then, we have

$$x + \alpha \inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} u(w)\psi(dw) = x + \alpha c + \alpha \inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} w\psi(dw) = x + \alpha(c + c')$$

This shows that if we choose $c = \beta c'$, then u is the unique solution.

Now we move on to show the lower bounds. For fixed $x \in \mathbb{X}$, we define a local version of the minimax risk

$$\mathfrak{M}_n(\mathcal{U}, \mathcal{K}, x) = \inf_K \sup_{\mu\in\mathcal{U}} E_{\mu^n} |K(W_1, \cdots, W_n)(x) - \mathcal{K}(\mu)(x)| \le \mathfrak{M}_n(\mathcal{U}, \mathcal{K})$$

which trivially lower bounds the uniform version. We will prove Theorem 5 and 6 by showing the same lower bound for this local risk.

D.2 Proof of Theorem 5

We apply Le Cam's technique to prove the lower bound. Recall the instance in Lemma 1. Fix $x \in [0, 1]$, for any $\eta > 0$ and $\mu_0, \mu_1 \in \mathcal{U}$ s.t. whenever $|\mathcal{K}(\mu_0)(x) - \mathcal{K}(\mu_1)(x)| \ge 2\eta$, we have

$$\mathfrak{M}_n(\mathcal{U}, \mathcal{K}, x) \ge \frac{\eta}{2}(1 - \|\mu_1^n - \mu_0^n\|_{\mathrm{TV}}).$$

We consider $\mu_0 = p_0\delta_{\{1\}} + (1 - p_0)\delta_{\{0\}}$ with $p_0 \le \frac{1}{2}$. Then

$$\begin{aligned}
\mathcal{K}(\mu_0)(x) - x &= \beta \inf_{\psi\in\mathcal{P}} \int_{\mathbb{W}} w\psi(dw) \\
&= \beta \sup_{\lambda\ge 0} -\lambda\delta + \int_{[0,1]} \inf_{y\in[0,1]} \left(y + \lambda(w - y)^2\right) \mu_0(dw) \\
&= \beta \sup_{\lambda\ge 0} -\lambda\delta + p_0 \frac{4\lambda - 1}{4\lambda} \mathbb{1}\left\{\lambda \ge \frac{1}{2}\right\} + p_0\lambda \mathbb{1}\left\{0 \le \lambda < \frac{1}{2}\right\} \\
&= \beta \max\left\{ \sup_{\lambda\ge 1/2} p_0 - \lambda\delta - \frac{p_0}{4\lambda}, \sup_{0\le\lambda<1/2} p_0\lambda - \lambda\delta \right\} \\
&= \beta \max\left\{ p_0 - \sqrt{p_0\delta}, \frac{p_0 - \delta}{2} \right\}
\end{aligned}$$

It is not hard to see the max is always achieved by $p_0 - \sqrt{p_0\delta}$.

So, if we construct the local alternative $\mu_1 = p_1\delta_{\{1\}} + (1 - p_1)\delta_{\{0\}}$, then $\mathcal{K}(\mu_1)(x) - x = \beta(p_1 - \sqrt{p_1\delta})$. Therefore, choosing any $p_1 = p_0 + c$ with $\frac{1}{2} \ge c > 0$, we have

$$\begin{aligned}
|\mathcal{K}(\mu_0)(x) - \mathcal{K}(\mu_1)(x)| &= \beta \left| c + \sqrt{\delta}\left(\sqrt{p_0} - \sqrt{p_0 + c}\right) \right| \\
&\ge \beta \inf_{\xi\in[p_0, p_0+c]} \left| c - \sqrt{\delta}\frac{1}{2}\xi^{-1/2}c \right| \\
&\ge \beta \left(1 - \frac{\sqrt{\delta}}{2\sqrt{p_0}}\right) c \\
&\ge \frac{\beta c}{2}
\end{aligned}$$

Hence, we can choose $p_1 = p_0 + 4\beta^{-1}\eta$ when $\eta \le \frac{\beta}{8}$ to achieve separation $|\mathcal{K}(\mu_0)(x) - \mathcal{K}(\mu_1)(x)| \ge 2\eta$.

By properties of TV-distance, KL, and χ^2-divergence we have that

$$\begin{aligned}\|\mu_1^n - \mu_0^n\|_{\mathrm{TV}} &\le \frac{n}{2} D_{\mathrm{KL}}(\mu_1||\mu_0) \\ &\le \frac{n}{2}\chi^2(\mu_1||\mu_0) \\ &\le \frac{n}{2}\frac{(p_0-p_1)^2}{p_0(1-p_0)} \\ &= \frac{8n\eta^2}{\beta^2 p_0(1-p_0)}\end{aligned}$$

So, for all $n \ge 1$, we can choose

$$\eta = \frac{\beta\sqrt{p_0(1-p_0)}}{4\sqrt{n}} \le \frac{\beta}{8}.$$

With this η, we conclude that $\|\mu_1^n - \mu_0^n\|_{\mathrm{TV}} \le \frac{1}{2}$ and hence

$$\mathfrak{M}_n(\mathcal{U},\mathcal{K}) \ge \mathfrak{M}_n(\mathcal{U},\mathcal{K},x) \ge \frac{\eta}{4} = \frac{\beta\sqrt{p_0(1-p_0)}}{16} n^{-1/2}.$$

Since p_0 is arbitrary, we can choose the maximizer $p_0 = \frac{1}{2}$.

D.3 Proof of Theorem 6

We lower bound the uniform risk by

$$\mathfrak{M}_n(\mathcal{U},\mathcal{K}) \ge \mathfrak{M}_n(\mathcal{U},\mathcal{K},0).$$

To achieve this, we would like to apply Duchi Theorem 3.

Notice that for two-point distribution μ with support $\{0,1\}$,

$$\mathcal{K}(\mu)(0) = \inf_{D_{f_k}(\mu'||\mu)\le\delta} E_{\mu'}\beta Z = -\beta + \sup_{D_{f_k}(\mu'||\mu)\le\delta} E_{\mu'}\beta(1-Z).$$

Here, $\beta(1-Z)$ has a two-point distribution on $\{0,\beta\}$ under μ. Therefore, Theorem 3 of Duchi and Namkoong (2021) applies. Define

$$p_k(\delta) = (1+k(k-1)\delta)^{-\frac{1}{k-1}}, \quad \chi_k(\delta) = \frac{k(k-1)\delta}{2(1+k(k-1)\delta)}.$$

We obtain that with n s.t.

$$\sqrt{\frac{p_k(\delta)(1-p_k(\delta))}{8n}} \le \frac{1-p_k(\delta)}{2} \wedge p_k(\delta), \quad \frac{1}{4n} \le p_k(\delta) \wedge (1-(1-\chi_k(\delta))^{1-k'}p_k(\delta)),$$

then

$$\mathfrak{M}_n(\mathcal{U},\mathcal{K},0) \ge \beta \max\left\{ \frac{\sqrt{p_k(\delta)(1-p_k(\delta))}}{16\sqrt{2}k'p_k(\delta)} n^{-\frac{1}{2}}, \frac{\chi_k(\delta)^{\frac{1}{k}} c_k(\delta)}{8\cdot 4^{k'}} n^{-\frac{1}{k'}} \right\}.$$

This implies the statement of Theorem 6.

E Algorithm Design for the CAU Case

To parameterize randomized controller policies, we employ a generative model by considering $\pi_\eta : \mathbb{R}^d \times \mathbb{X} \to \mathbb{A}$ where an action is generated by $A \sim \pi_\eta(N,x)$ using an independence source of randomness $N \sim N(0,I)$ is a standard Gaussian vector independent of the state.

Under this definition, we overwrite the notation and consider

$$\widehat{\mathbf{T}}_{\eta,\theta}(\lambda,x) := E_N r(x,\pi_\eta(N,x)) + \alpha\left[\lambda - c_k(\delta)\left[\int_{\mathbb{W}} (u_\theta(f(x,\pi_\eta(N,x),w)) - \lambda)_+^{k'}\hat{\mu}(dw)\right]^{1/k'}\right]$$

By strong duality, the Bellman operator under policy π_η applied to u_θ is

$$E_N r(x, \pi_\eta(N,x)) + \alpha \sup_{\lambda \in \mathbb{R}} \left(\lambda - c_k(\delta) \left[\int_{\mathbb{W}} \left(E_N u_\theta(f(x, \pi_\eta(N,x), w)) - \lambda \right)^{k'} \hat{\mu}(dw) \right]^{1/k'} \right) = \widehat{\mathbf{T}}_{\eta,\theta}(\lambda^*, x)$$

where $\lambda^* = \lambda^*(\eta, \theta, x)$ is the optimal dual multiplier. We note that λ^* doesn't depend on the realizations of N, and can be computed via bisection search.

Bellman Error Minimization: Analogous to the CAA case, we minimize the L^2 Bellman error:

$$\min_\theta \int_{\mathbb{X}} [u_\theta(x) - \widehat{\mathbf{T}}_{\eta,\theta}(\lambda^*, x)]^2 \nu(dx)$$

The gradient is evaluated using the envelope theorem

$$\nabla_\theta \mathbf{T}_{\eta,\theta}(x) = \nabla_\theta \widehat{\mathbf{T}}_{\eta,\theta}(\lambda^*(\eta,\theta,x), x)$$

and the expectation over N can be approximated using the sample average over m i.i.d. samples $N_1, \ldots, N_m$.

Therefore, we update the θ using first order methods; e.g. mini-batch stochastic gradient descent (SGD):

$$\theta_{t+1} = \theta_t - \ell_t G_{n,m,t}$$

where ℓ_t is the learning rate and $G_{n,m,t}$ is a stochastic gradient.

The gradient estimate $G_{n,m,t}$ can be obtained as follows: we first sample $\mathbf{N} = \{N_1, \ldots, N_m\}$ and $X_i \sim \nu$ i.i.d. and compute $\lambda^*_{m,i} = \lambda^*(\eta, \theta, \mathbf{N}, X_i)$ that maximize

$$\lambda - c_k(\delta) \left[\int_{\mathbb{W}} \left(\frac{1}{m} \sum_{j=1}^m u_\theta(f(X_i, \pi_\eta(N_j, X_i), w)) - \lambda \right)^{k'} \hat{\mu}(dw) \right]^{1/k'}$$

using bisection search. Then, we set

$$G_{n,m,t} = \frac{1}{n} \sum_{i=1}^n 2(u_\theta(X_i) - \widehat{\mathbf{T}}_{\eta,\theta}(\lambda^*_{m,i}, X_i))(\nabla_\theta u_\theta(X_i) - \nabla_\theta \overline{\mathbf{T}}_{\eta,\theta}(\lambda^*_{m,i}, X_i)).$$

Policy Improvement: The policy improvement step parallels the CAA setting:

$$\eta_{t+1} = \eta_t + \ell'_t \frac{1}{n} \sum_{i=1}^n \nabla_\eta \widehat{\mathbf{T}}_{\eta,\theta}(\lambda^*_{m,i}, X_i),$$

for some possibly different learning rate ℓ'_t.