Approximate Bilevel Difference Convex Programming for Bayesian Risk Markov Decision Processes

Yifan Lin, Enlu Zhou
Industrial and Systems Engineering
Georgia Institute of Technology
755 Ferst Dr
Atlanta, GA 30332 USA
ylin429@gatech.edu, enlu.zhou@isye.gatech.edu

Abstract

We consider infinite-horizon Markov Decision Processes where parameters, such as transition probabilities, are unknown and estimated from data. The popular distributionally robust approach to addressing the parameter uncertainty can sometimes be overly conservative. In this paper, we utilize the recently proposed formulation, Bayesian risk Markov Decision Process (BR-MDP), to address parameter (or epistemic) uncertainty in MDPs. To solve the infinite-horizon BR-MDP with a class of convex risk measures, we propose a computationally efficient approach called approximate bilevel difference convex programming (ABDCP). The optimization is performed offline and produces the optimal policy that is represented as a finite state controller with desirable performance guarantees. We also demonstrate the empirical performance of the BR-MDP formulation and the proposed algorithm.

1 Introduction

In a Markov decision process (MDP), an agent must make decisions in a sequence while facing uncertainty. In this situation, some parameters of the MDP, such as the transition probabilities and costs, may be unknown and must be estimated from available data. The problem then becomes how to determine the best course of action, given the limited or possibly absent data, in order to minimize the expected total cost and optimize the decision-making process under these uncertain parameters.

An alternative approach to addressing the epistemic uncertainty in MDP is through the use of distributionally robust MDPs (DR-MDP, Xu and Mannor (2010)). It considers unknown parameters as random variables and assumes that their distributions belong to an ambiguity set determined by the available data. The optimal policy is then found by minimizing the expected total cost using the most adversarial distribution within this ambiguity set. However, these distributionally robust approaches may lead to overly conservative solutions that do not perform well in scenarios that are more likely to occur than the worst case. Additionally, the DR-MDP framework does not explicitly incorporate the dynamics of the problem, as the distribution of the unknown parameters does not depend on the data process, and is therefore not time-consistent, as noted by Shapiro (2021). In light of these limitations, Lin, Ren, and Zhou (2022) propose a Bayesian risk MDP (BR-MDP) framework to address epistemic uncertainty in MDPs. However, the approximation algorithm proposed by Lin, Ren, and Zhou (2022) only applies to finite-horizon MDPs and does not scale well with long horizon. It only provides an upper bound on the exact value function, without any theoretical guarantee on the gap.

We reformulate the considered BR-MDP as a bilevel difference convex program (DCP) such that we can employ the powerful optimization methods for DCP to solve infinite-horizon BR-MDP. Since the space of posterior distributions (beliefs) is uncountably infinite, we approximate the bilevel DCP by considering only a subset of posterior distributions. Although the DCP is approximate, we show that its solution is a lower bound on the exact optimal value function. Using the representation of a finite state controller of the resulting policy, we further show an upper bound on the exact optimal value function. We develop an iterative approach to reduce the gap between upper and lower bounds by incrementally generating new sets of posterior distributions, and show the convergence of the proposed algorithm.

To summarize, the contributions of this paper are twofold. First, we analyze the infinite-horizon MDP with epistemic uncertainty under BR-MDP via a Bayesian perspective and show the existence and uniqueness of stationary optimal policy. Second, we propose an approximate difference convex programming algorithm to solve the proposed formulation and show the convergence of the proposed algorithm. The rest of the paper is organized as follows. We conduct literature review and introduce the BR-MDP framework in Section 2. We show the existence and uniqueness of a stationary optimal policy to the infinite-horizon BR-MDP in Section 3.1. We provide a bilevel DCP solution to the infinite-horizon BR-MDP in Section 3.2. A computationally efficient approximate DCP algorithm is shown in Section 3.3. We verify the theoretical results and demonstrate the performance of our algorithms via numerical experiments in Section 4. Finally, we conclude the paper in Section 5.

2 Background

2.1 Related Literature

If data used to estimate the true but unknown underlying MDP are not sufficient, the estimated MDP may signifi-

cantly differ from the true MDP, leading to poor policy performance. This discrepancy (between the estimated MDP and the true MDP) can be seen tightly linked to the epistemic uncertainty about the model. There have been numerous approaches that address epistemic uncertainty in MDPs, with robust MDP and its variants (Nilim and Ghaoui (2004); Iyengar (2005); Delage and Mannor (2010); Wiesemann, Kuhn, and Rustem (2013); Petrik and Russel (2019); Zhou et al. (2021); Yang, Zhang, and Zhang (2022); Cousins et al. (2023); Derman et al. (2020)) being one of the most widely used methods. In robust MDPs, the optimal decisions are made based on their performance under the most unfavorable conditions within a known ambiguity set of possible parameter values.

Apart from the overly conservative robust MDP approach which only considers the worst-case scenario, the risk-averse approach has been proposed to address the epistemic uncertainty, but with more flexibility in choosing the risk functional. Risk-averse approach is originally proposed to address the aleatoric uncertainty due to the inherent stochasticity of the underlying MDP (Howard and Matheson (1972); Ruszczyński (2010); Petrik and Subramanian (2012); Osogami (2012)). It replaces the risk-neutral expectation by some general risk measures, such as conditional value-at-risk (CVaR, see Rockafellar and Uryasev (2000)). However, most of the existing approaches assume the agent has access to the true underlying MDP, and optimize some risk measures such as CVaR in that single MDP (Chow and Ghavamzadeh (2014); Tamar et al. (2015); Tamar, Glassner, and Mannor (2015); Sharma et al. (2019)). In this paper, we consider the offline planning problem in MDPs, where we only have access to prior belief distribution over MDPs that is constructed by the offline data. It should be noted that offline planning problem has also been considered by Duff (2002), where the author proposes a Bayes-adaptive MDP (BA-MDP) formulation with an augmented state composed of the underlying MDP state and the posterior distribution of the unknown parameters. Mostly close to the problem setting in this work are Rigter, Lacerda, and Hawes (2021); Lin, Ren, and Zhou (2022). Rigter, Lacerda, and Hawes (2021) optimize a CVaR risk functional over the total cost and simultaneously addresses both epistemic and aleatoric uncertainty, while Lin, Ren, and Zhou (2022) consider a nested risk functional to ensure the time consistency of the obtained policy.

While there are many works proposing different models and frameworks to address the epistemic uncertainty, developing computationally efficient solutions is also of great interest. In robust MDPs, with some mild conditions on the ambiguity set such as rectangularity, the proposed formulation can be solved by a second-order cone program when the horizon is finite, or policy iteration when the horizon is infinite (Mannor and Xu (2019)). In BA-MDP and its variants, Rigter, Lacerda, and Hawes (2021) propose an approximate algorithm based on Monte Carlo tree search and Bayesian optimization. Lin, Ren, and Zhou (2022) develop an α-function approximation algorithm using the convexity of the CVaR risk measure. However, the aforementioned works consider a finite-horizon MDP and do not generalize well to the infinite-horizon setting.

Compared to standard MDPs, our considered problem has two distinct features that make it difficult to apply value iteration, policy iteration, or linear programming (Puterman (2014)). First is the resulting continuous-state MDP due to augmented belief state. We note that this continuous-state MDP is similar to a belief-MDP, which is the equivalent way to represent a partially observable MDP (POMDP) by treating the posterior distribution of the hidden state as a belief state. Second is the risk measure taken with respect to the unknown parameters in the MDPs. In this work, we propose an optimization-based method to solve the infinite-horizon BR-MDPs. It has been empirically shown by Alagoz, Ayvaci, and Linderoth (2015) that linear programming can efficiently solve a significant number of MDPs in comparison to standard dynamic programming methods, such as value iteration and policy iteration. Furthermore, linear programming requires less memory and can handle MDPs with a larger number of states and still achieve optimality. It has been widely used in risk-sensitive MDP (that deals with intrinsic or aleatoric uncertainty that is due to the inherent stochasticity of the underlying MDP, see Zhang et al. (2021)). Works that are most related to our proposed optimization-based approach include Poupart et al. (2015) who propose an approximate linear programming algorithm for the risk-neutral constrained POMDPs, and Ahmadi et al. (2021) who propose a difference convex program (DCP) for the constrained risk-averse MDPs. Our approach for infinite-horizon BR-MDP significantly differs from the above approaches in two aspects. First, compared to the linear programming approach in Poupart et al. (2015), we use bilevel DCP, due to the additional risk measure used for mitigating the epistemic uncertainty. Our considered risk measure brings additional challenge to exactly evaluating the policy, whereas policy evaluation can be easily solved by a system of linear equations in Poupart et al. (2015). Second, compared to the DCP for the risk-averse MDP with aleatoric uncertainty in Ahmadi et al. (2021), the resulting continuous-state MDP in our problem has an infinite number of constraints and requires appropriate approximation to make the problem computationally feasible.

2.2 Preliminary: Bayesian Risk MDPs

Consider an infinite-horizon MDP that is defined as $(\mathcal{S}, \mathcal{A}, P, C, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is the transition probability with $P(s'|s,a)$ denoting the probability of transitioning to state s' from state s when action a is taken, C is the cost function with $C(s,a,s')$ denoting the cost when action a is taken and state transitions from s to s', $0 \leq \gamma < 1$ is the discount factor. We assume the state space and action space are finite and the cost is bounded. A Markovian deterministic policy π is a function mapping from $\mathcal{S}$ to $\mathcal{A}$. Given an initial state s, the goal is to find an optimal policy that minimizes the expected discounted total cost: $\min_{\pi} \mathbb{E}^{\pi,P,C}\left[\sum_{t=1}^{\infty} \gamma^{t-1} C\left(s_t, a_t, s_{t+1}\right) | s_1 = s\right]$, where $\mathbb{E}^{\pi,P,C}$ is the expectation with policy π when the transition probability is P and the cost is C. In practice, P and

C are often unknown and estimated from data.

BR-MDP is a recently proposed framework that deals with the epistemic uncertainty in MDPs (see Lin, Ren, and Zhou (2022)). It is assumed that the state transition is specified by the state equation $s' = g(s, a, \xi)$ with a known transition function g which involves state $s \in \mathcal{S} \subseteq \mathbb{R}^{k_s}$, action $a \in \mathcal{A} \subseteq \mathbb{R}^{k_a}$, and randomness $\xi \in \Xi \subseteq \mathbb{R}^{k_\xi}$, where k_s, k_a, k_ξ are the dimensions of the state, action, and randomness space, respectively. The state equation together with the distribution of ξ uniquely determines the transition probability of the MDP, i.e., $P(s' \in S'|s, a) = P(\{\xi \in \Xi : g(s, a, \xi) \in S'\}|s, a)$, where S' is a measurable set in $\mathcal{S}$. We refer the readers to Chapter 3.5 in Puterman (2014) for the equivalence between stochastic optimal control and MDP formulation. We use the representation of state equations instead of transition probabilities in MDPs, for the purpose of decoupling the randomness and the policy, leading to a cleaner formulation in the nested form. The cost is assumed to be a function of state s, action a, and randomness ξ, i.e., $C(s, a, \xi)$.

The distribution of ξ, denoted by $f(\cdot; \theta^c)$, is assumed to belong to a parametric family $\{f(\cdot; \theta)|\theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ is a convex parameter space, d is the dimension of the parameter space Θ, and $\theta^c \in \Theta$ is the true but unknown parameter value. Many real-world problems exhibit the characteristic of relying on a parametric assumption. For example, it is commonly assumed that the demand of customers follows a Poisson distribution with an unknown arrival rate in inventory control. We begin by assuming a prior distribution, denoted by μ, over the parameter space Θ. This prior accounts for the uncertainty of the parameter estimate that comes from an initial set of data, and it can also take expert opinions into consideration. Then, given an observed realization of the data process, we update the posterior distribution μ according to the Bayes' rule. Let the policy be a sequence of mappings from state s and posterior μ to the action space, i.e., $\pi = \{\pi : \mathcal{S} \times \mathcal{M} \to \mathcal{A}\}$, where $\mathcal{M}$ is the space of posterior distributions. This representation implies the policy is stationary. Now we present the BR-MDP formulation below.

$$\min_{\pi} \ \rho_{\mu_1}\mathbb{E}_{\theta_1}\Big[C_1(s_1, a_1, \xi_1) + \cdots + \gamma^{t-1}\rho_{\mu_t}\mathbb{E}_{\theta_t}\big[C_t(s_t, a_t, \xi_t) + \cdots\big]|s_1 = s, \mu_1 = \mu\Big] \tag{1}$$

$$\text{s.t.} \quad s_{t+1} = g(s_t, a_t, \xi_t), \tag{2}$$

$$a_t = \pi(s_t, \mu_t),$$

$$\mu_{t+1}(\theta) = \frac{\mu_t(\theta) f\left(\xi_t; \theta\right)}{\int_\Theta \mu_t(\theta) f\left(\xi_t; \theta\right) d\theta}, \tag{3}$$

where ρ is a risk measure (we defer the definition and form of the risk measure ρ to Section 2.3), θ_t is a random vector following distribution μ_t, $\mathbb{E}_{\theta_t}$ denotes the expectation with respect to $\xi_t \sim f(\cdot; \theta_t)$ conditional on θ_t, and ρ_{μ_t} denotes a risk functional with respect to $\theta_t \sim \mu_t$ applied in nested form to the expected total cost with respect to the Bayesian posterior distributions of the unknown parameters. Equation (2) is the transition of the state s_t, and without loss of generality we assume the initial state s_1 takes a deterministic value s. Equation (3) is the updating of the posterior μ_t. For a given dataset with size N, the prior distribution converges to a Dirac delta function concentrated on the true parameter θ^c with probability 1, and the optimal value function of BR-MDP converges to the optimal value function of the true MDP.

2.3 Preliminary: Risk Measure

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{Z}$ be a linear space of $\mathcal{F}$-measurable functions $Z : \Omega \to \mathbb{R}$. A risk measure is a function $\rho : \mathcal{Z} \to \mathbb{R}$ which assigns a random variable Z to a real number representing its risk. It is said that risk measure ρ is convex if it possesses the properties of convexity, monotonicity, and translation invariance (see Föllmer and Schied (2002)). In this paper we consider a class of convex risk measures which can be represented in the following parametric form: $\rho_\mu(Z) := \inf_{\phi\in\Phi} \mathbb{E}_\mu[\Psi(Z, \phi)]$, where $\Phi \subset \mathbb{R}^m$ and $\Psi : \mathbb{R} \times \Phi \to \mathbb{R}$ is a real-valued convex function, and $\Psi(\cdot, \phi)$ is finite-valued and continuous on a compact set of ϕ. There is a large class of risk measures which can be represented in the parametric form. For example, conditional value-at-risk (CVaR), defined as $\mathrm{CVaR}_\alpha(X) = \min_{\phi\in\mathbb{R}} \left\{\phi + \frac{1}{1-\alpha}\mathbb{E}\left[(X - \phi)^+\right]\right\}$, where $(\cdot)^+$ stands for $\max(0, \cdot)$, is widely used (see Rigter, Lacerda, and Hawes (2021); Chow et al. (2015)). Another example is risk measures constructed from ϕ-divergence ambiguity sets (see Example 3 in Guigues, Shapiro, and Cheng (2024)). We refer the readers to Shapiro, Dentcheva, and Ruszczynski (2021) for a comprehensive discussion.

3 Algorithm and Analysis

3.1 Bellman Equation and Optimality

We can write the value function under policy π of BR-MDP in the following recursive forms.

$$V^\pi(s, \mu) = \rho_\mu\mathbb{E}_\theta\big[C(s, a, \xi) + \gamma V^\pi(s', \mu')\big]$$
$$\text{s.t.} \quad s' = g(s, a, \xi), a = \pi(s, \mu);$$
$$\mu'(\theta) = \frac{\mu(\theta) f\left(\xi; \theta\right)}{\int_\Theta \mu(\theta) f\left(\xi; \theta\right) d\theta}.$$

We refer the readers to Lin, Ren, and Zhou (2022) for a discussion on the preference of dynamic risk measure over static risk measure in consideration of time consistency and derivation of the Bellman equation. For simplicity we only consider deterministic policies, but all the analysis below can be extended to stochastic policies. For the stochastic policies, the expectation in (1) is taken with respect to the randomness ξ and the action a. As a consequence of Theorem 5.5.3b in Puterman (2014), it is sufficient to consider the Markovian policy. The optimal value function is then denoted as $V^*(s, \mu) = \min_{\pi\in\Pi^{MD}} V^\pi(s, \mu)$, where Π^{MD} is the set of Markovian deterministic policies. It should be noted that the Bayes optimality is with respect to the prior belief μ. In the following, we derive the intermediate results to show V^* is the unique optimal value function to the infinite-horizon BR-MDP.

Definition 3.1 (Bellman Operator)**.** Let $B(\mathcal{S}, \mathcal{M})$ be the space of real-valued bounded measurable functions on $(\mathcal{S} \times$

$\mathcal{M}$). For any bounded value function $V \in B(\mathcal{S}, \mathcal{M})$, define an operator $\mathcal{T} : B(s,\mu) \to B(s,\mu)$ as:

$$(\mathcal{T}V)(s,\mu) = \min_{a\in\mathcal{A}} \rho_\mu \left[\mathbb{E}_\theta \left[C(s,a,\xi) + \gamma V(s',\mu')\right]\right].$$

Also let $\mathcal{T}^\pi : B(s,\mu) \to B(s,\mu)$, where

$$(\mathcal{T}^\pi V)(s,\mu) = \rho_\mu \left[\mathbb{E}_\theta \left[C(s,\pi(s,\mu),\xi) + \gamma V^\pi(s',\mu')\right]\right].$$

The next two lemmas show the above Bellman operators are monotonic and contraction mappings.

Lemma 3.2 (Monotonicity)**.** *The operators $\mathcal{T}^\pi$ and $\mathcal{T}$ are monotonic, in the sense that $V \le V'$ implies $\mathcal{T}^\pi V \le \mathcal{T}^\pi V'$ and $\mathcal{T}V \le \mathcal{T}V'$.*

Lemma 3.3 (Contraction Mapping)**.** *The operators $\mathcal{T}^\pi$ and $\mathcal{T}$ are γ contraction for $||\cdot||_\infty$ norm. That is, for any two bounded value functions $V, V' \in B(\mathcal{S}, \mathcal{M})$, we have*

$$||\mathcal{T}^\pi V - \mathcal{T}^\pi V'||_\infty \le \gamma ||V - V'||_\infty.$$

The following proposition shows that sub-solutions V_{sub} and super-solutions V_{sup} of the optimality equations $V = \mathcal{T}V$ provide lower and upper bounds on V^*. As a result, when a solution is obtained, both bounds are satisfied, meaning that the solution must be equivalent to V^*. Additionally, this outcome serves as an important algorithmic tool for optimization-based methods.

Proposition 3.4. *For any $V_{\text{sub}}, V_{\text{sup}} \in B(\mathcal{S}, \mathcal{M})$, (i) if $V_{\text{sup}} \ge \mathcal{T}V_{\text{sup}}$, then $V_{\text{sup}} \ge V^*$; (ii) if $V_{\text{sub}} \le \mathcal{T}V_{\text{sub}}$, then $V_{\text{sub}} \le V^*$.*

According to Proposition 3.4, we have $V^* = \mathcal{T}V^*$. By Banach fixed-point theorem, V^* is the unique optimal value function to the infinite horizon BR-MDP. We also have that the value V of a stationary policy π is the unique bounded solution of the equation $V = \mathcal{T}^\pi V$. Similar analysis shows the existence and uniqueness of the optimal stationary policy π^* that satisfies $V^* = \mathcal{T}^{\pi^*} V^*$.

Applying the operator $\mathcal{T}$ on any initial value function V, we have the value iteration algorithm for the infinite-horizon BR-MDP problem. The following corollary of convergence rate is similar to the standard with the contraction property.

Corollary 3.5. *For any initial bounded value function V, the convergence rate is shown to be $||(\mathcal{T}^k V)(s,\mu) - V^*(s,\mu)||_\infty \le \gamma^k ||V(s,\mu) - V^*(s,\mu)||_\infty$.*

3.2 Bilevel Difference Convex Programming

The main challenge of executing the value iteration algorithm (and similarly policy iteration algorithm) lies in the continuous augmented state. In this work, we propose an optimization-based method to solve the infinite-horizon BR-MDPs. According to Proposition 3.4, the infinite-horizon BR-MDP can be solved as follows:

$$\begin{aligned}
&\max_V \sum_{s\in\mathcal{S},\mu\in\mathcal{M}} \alpha(s,\mu) V(s,\mu) \\
&\text{s.t. } V(s,\mu) \le \rho_\mu \mathbb{E}_\theta \Big[C(s,a,\xi) + \gamma V(s',\mu')\Big], \\
&\quad \forall a \in \mathcal{A}, s \in \mathcal{S}, \mu \in \mathcal{M},
\end{aligned}$$

where we choose $\alpha(s,\mu)$ to be positive scalars which satisfy $\sum_{s\in\mathcal{S},\mu\in\mathcal{M}} \alpha(s,\mu) = 1$. For the considered class of convex risk measures, we can rewrite the above formulation as a bilevel difference convex program:

$$\begin{aligned}
&\min_V \quad - \sum_{s\in\mathcal{S},\mu\in\mathcal{M}} \alpha(s,\mu) V(s,\mu) \qquad (4) \\
&\text{s.t.} V(s,\mu) - \min_\phi \mathbb{E}_\mu \Big[\Psi\big(\mathbb{E}_\theta[C(s,a,\xi) + \gamma V(s',\mu')], \phi\big)\Big] \\
&\quad \le 0, \forall a \in \mathcal{A}, s \in \mathcal{S}, \mu \in \mathcal{M}.
\end{aligned}$$

Since $\Psi(Z,\phi)$ is convex in (Z,ϕ) and expectation is a linear operator, the minimum of $\mathbb{E}[\Psi(Z,\phi)]$ over a convex set Θ remains convex in Z. Thus, (4) is a bilevel difference convex program (see Horst and Thoai (1999) for the definition of DCP). It should be noted that Ahmadi et al. (2021) show that the minimum over ϕ can be absorbed into the overall minimum problem, and ϕ is treated as a single variable. However, it is clear that the minimum is achieved at different ϕ for different augmented state (s,μ), thus turning (4) into a bilevel optimization problem. When the lower-level problem is convex and satisfies certain regularity conditions, we can use the Karush-Kuhn-Tucker (KKT) conditions to reformulate the lower-level optimization problem, which allows us to transform the original bilevel optimization problem into a single-level (constrained) optimization problem.

After being reduced to a single-level DCP problem, (4) can be solved by the convex-concave procedure (see Lipp and Boyd (2016) for such procedure), wherein the concave terms are replaced by a convex upper bound. We employ the method of disciplined convex-concave programming (DCCP, Shen et al. (2016)), which converts a DCP problem into a disciplined convex program and subsequently into an equivalent cone program. However, one problem remains to be solved: the number of constraints in (4) is infinite, due to the continuous belief state. To tackle this problem, we take a similar approach as Poupart et al. (2015). The main idea is to start with a finite posterior set (belief space) $\hat{\mathcal{M}}$, and then problem (4) can be solved efficiently by DCCP, where the posterior distribution (belief point) not in the set $\hat{\mathcal{M}}$ is replaced by convex combination of the points in $\hat{\mathcal{M}}$. We then iteratively add to the posterior set new posterior distributions that are reachable from the current set and re-solve (4). It should be noted that the proposed approach could be extended to value iteration and policy iteration, but the analysis would be more complicated, since there would be a trade-off between the optimization (e.g. value iteration) and the belief point generation, and it could be quite tricky to decide the optimal number of steps for value iteration and optimal intervals for belief point generation. We formally introduce the approximate bilevel DCP algorithm in the next section.

3.3 Approximate Bilevel Difference Convex Programming

Let $\hat{\mathcal{M}}$ be the current posterior set. Let $\mu^{sas'}$ be the one-step posterior distribution with observed randomness ξ indicated by state transition $s' = g(s,a,\xi)$ and current posterior μ. Initially the posterior set is constructed from corner

(degenerate) points. In case the parameter space Θ is finite, the corner points are $(1, 0, \cdots, 0)$, $(0, 1, 0, \cdots)$, $\cdots$, and $(0, \cdots, 0, 1)$. In case the parameter space is continuous, it is impossible to express one-step posterior distribution (i.e., $\mu^{sas'}$) as a convex combination of those degenerate points. Therefore, we assume the parameter space is finite, which is practical in many real-world problems. It can also be viewed as a discrete approximation of a continuous parameter set, and the discretization can be chosen of any precision.

Algorithm 1: Approximate Bilevel DCP

input: posterior set $\hat{\mathcal{M}}$
output: policy $\hat{\pi}^*$, approximate value function $\hat{V}^*$
1. solve the following approximate bilevel DCP:

$$\min_{V} \quad -\sum_{s\in\mathcal{S},\mu\in\hat{\mathcal{M}}} \alpha(s,\mu)V(s,\mu) \tag{5}$$
$$\text{s.t. } V(s,\mu) \le \min_{\phi} \sum_{\theta\in\Theta} \mu(\theta)\Big[\Psi\big(\gamma \sum_{\mu'\in\hat{\mathcal{M}},s'\in\mathcal{S}} P(s'|s,a,\theta)$$
$$w(\mu',\mu^{sas'})V(s',\mu') + C(s,a,\theta),\phi\big)\Big], \forall a\in\mathcal{A}, s\in\mathcal{S}, \mu\in\hat{\mathcal{M}}$$

where $w(\mu', \mu^{sas'})$ is obtained by solving (6).
2. obtain the approximate solution $\hat{V}^*$ to (5); obtain the approximate policy

$$\hat{\pi}^*(s,\mu) = \arg\min_{\phi,a\in\mathcal{A}} \sum_{\theta\in\Theta} \mu(\theta)\Big[\Psi\big(\gamma \sum_{\mu'\in\hat{\mathcal{M}},s'\in\mathcal{S}} P(s'|s,a,\theta)$$
$$w(\mu',\mu^{sas'})\hat{V}^*(s',\mu') + C(s,a,\theta),\phi\big)\Big], \forall s\in\mathcal{S}, \mu\in\hat{\mathcal{M}}.$$

To interpolate all $\mu^{sas'}$ that can be reached from some $\mu_i \in \hat{\mathcal{M}}$ in one step, we use some convex combination of points μ_i in $\hat{\mathcal{M}}$. Let $w(\mu_i, \mu^{sas'})$ be the weight w_i associated with μ_i when interpolating $\mu^{sas'}$. We can use this interpolation weight to define an approximate transition probability for posterior as:

$$\tilde{P}(\mu'|s,a,\mu,\theta) = \sum_{s'\in\mathcal{S}} P(s'|s,a,\theta)w(\mu',\mu^{sas'}).$$

A sanity check that $\tilde{P}(\mu'|s,a,\mu,\theta)$ is indeed a transition probability: $\sum_{\mu'\in\hat{\mathcal{M}}} \tilde{P}(\mu'|s,a,\mu,\theta) = 1$ and $\tilde{P}(\mu'|s,a,\mu,\theta) \ge 0$. We choose the convex combination that minimizes the weighted Euclidean norm of the difference between μ and each μ_i by solving the following linear program:

$$\min_{w} \sum_i w_i ||\mu_i - \mu^{sas'}||_2^2 \tag{6}$$
$$\text{s.t. } \sum_i w_i\mu_i(\theta) = \mu^{sas'}(\theta), \forall\theta\in\Theta$$
$$\sum_i w_i = 1, w_i \ge 0, \forall i.$$

With the approximation in the constraint in (4), we obtain the following approximate bilevel DCP Algorithm 1 for a given posterior set. For ease of notation, we denote by $C(s,a,\theta) = \mathbb{E}_\theta[C(s,a,\xi)]$ the average cost at state s when action a is taken, under the parameter value θ.

Theorem 3.6. *The approximate value function $\hat{V}^*$ found by running Algorithm 1 is a lower bound on the exact optimal value function V^*.*

We also develop an upper bound on the exact optimal value function, using the obtained policy from Algorithm 1. The obtained policy is a finite state controller (see Hansen (2013) for the definition of finite state controller). Let $\mathcal{N}$ be the set of nodes in the controller such that we associate a node $n_{s,\mu}$ to each (s,μ) pair. The action chosen in node $n_{s,\mu}$ is determined by the policy $\hat{\pi}^*(a|s,\mu)$. For a given parameter θ, the transition probability to the next node is $P(n_{s',\mu'}|n_{s,\mu},a) = w(\mu',\mu^{sas'})P(s'|s,a)$. The value function of the finite state controller can be computed by

$$\hat{V}^{\hat{\pi}^*}(n_{s,\mu}) = \min_{\phi} \sum_{\theta\in\Theta} \mu(\theta)\Big[\Psi\big(c(s,a,\theta) + \gamma \sum_{n_{s',\mu'}\in\mathcal{N}}$$
$$w(\mu',\mu^{sas'})P(s'|s,a,\theta)\hat{V}^{\hat{\pi}^*}(n_{s',\mu'}),\phi\big)\Big].$$

Similar to Ahmadi et al. (2021), the value function can be solved efficiently by DCP. It is also known from Hansen (2013) that the value function obtained by the finite state controller $\hat{V}^{\hat{\pi}^*}$ serves as an upper bound for the optimal value function.

Note that the inequality $\hat{V}^* \le V^* \le \hat{V}^{\hat{\pi}^*}$ provides information about how well the optimal value function V^* is approximated. As the posterior set $\hat{\mathcal{M}}$ contains more beliefs to accurately evaluate the policies, the gap between the approximate value function and the optimal value function gets smaller.

Algorithm 2: New Posterior Set Generation

input: policy $\hat{\pi}^*$, posterior set $\hat{\mathcal{M}}$, maximum number of newly added posterior distributions n
output: newly added posterior set $\hat{\mathcal{M}}'$
initialization: $\hat{\mathcal{M}}' \leftarrow \emptyset$.
for each $(s,\mu) \in (\mathcal{S}\times\hat{\mathcal{M}})$ and $s'\in\mathcal{S}$ **do**
  $\mu'(\theta) \propto \mu(\theta)f(\xi|\theta)$, where $s' = g(s,\hat{\pi}^*(a|s,\mu),\xi)$;
  $\text{dist}_{\mu'} \leftarrow$ distance of μ' to $\hat{\mathcal{M}}\bigcup\hat{\mathcal{M}}'$.
  if $\text{dist}_{\mu'} > 0$ (i.e., μ' not in $\hat{\mathcal{M}}\bigcup\hat{\mathcal{M}}'$) **then**
    $\hat{\mathcal{M}}' \leftarrow \hat{\mathcal{M}}'\bigcup\{\mu'\}$.
  end if
  if $|\hat{\mathcal{M}}'| > n$ (to reduce the size of $\hat{\mathcal{M}}'$) **then**
    for each $\mu'\in\hat{\mathcal{M}}'$ **do**
      $\text{dist}_{\mu'} \leftarrow$ distance of μ' to $\hat{\mathcal{M}}\bigcup\hat{\mathcal{M}}'\backslash\{\mu'\}$;
      $\hat{\mathcal{M}}' \leftarrow \hat{\mathcal{M}}'\backslash\{\arg\min_{\mu'\in\hat{\mathcal{M}}'}\text{dist}_{\mu'}\}$.
    end for
  end if
end for

Next we incrementally add new posterior distributions to the posterior set $\hat{\mathcal{M}}$. Different methods can be employed to

produce new posterior distributions that are added to the set $\hat{\mathcal{M}}$ at each iteration. We take a similar approach as Poupart et al. (2015), which is based on envelope techniques. It considers the posterior distributions that can be reached in one step from any posterior distribution in $\hat{\mathcal{M}}$ by executing the policy $\hat{\pi}^*$. As the number of posterior distributions to be added might be excessive, we can prioritize them by including the n reachable posterior distributions with the largest Euclidean distance to the posterior distributions in $\hat{\mathcal{M}}$. Note that the point-based value iteration approach in Pineau et al. (2003) shares the similar idea, that is, to include new posterior distribution that improves the worst-case density as rapidly as possible, where density is defined as the maximum distance from any posterior distribution to $\hat{\mathcal{M}}$. We summarize the new posterior set generation in Algorithm 2.

Combining Algorithm 1 and Algorithm 2, we now present the full algorithm below (ABDCP), which iteratively adds to the new posterior set and solves a bilevel difference convex program at each iteration. Theorem 3.7 shows that Algorithm 3 converges to a near-optimal policy.

Algorithm 3: ABDCP for infinite-horizon BR-MDPs

input: threshold ϵ, number of newly added posterior distributions n, initial state s_1, dataset $\mathcal{D}$
output: policy $\hat{\pi}^*$
initialization: compute prior distribution μ_1 using dataset $\mathcal{D}$; $\hat{M} \leftarrow \{\text{degenerate beliefs}\} \bigcup \{\mu_1\}$.
repeat
  obtain $(\hat{\pi}^*, \hat{V}^*)$ by running Algorithm 1;
  evaluate policy $\hat{\pi}^*$ by solving a DCP and obtain $\hat{V}^{\hat{\pi}^*}$;
  $\hat{\mathcal{M}} \leftarrow \hat{\mathcal{M}} \bigcup \hat{\mathcal{M}}'$ generated by Algorithm 2.
until $||\hat{V}^{\hat{\pi}^*} - \hat{V}^*||_\infty \leq \epsilon$

Theorem 3.7. *Algorithm 3 converges to a near-optimal policy $\hat{\pi}^*$, i.e., $||\hat{V}^{\hat{\pi}^*} - V^*||_\infty \leq \epsilon$, where ϵ is the desired threshold.*

As the number of iterations in Algorithm 3 increases, the gap between the optimal value function and the lower bound becomes arbitrarily small, which shows the lower bound in Theorem 3.6 is non-trivial.

4 Numerical Experiments

We illustrate the performance of the infinite-horizon BR-MDP formulation with different choices of risk measures and the proposed approximate bilevel DCP algorithm with an offline path planning problem.

We adapt two methods to our offline planning problems and compare their performances. The first method (CALP) comes from Poupart et al. (2015) with a risk-neutral POMDP formulation. The second method (DR-MDP) comes from Xu and Mannor (2010) with a distributionally robust MDP formulation. Note that the BPO approach from Lee et al. (2019) solves a risk-neutral BA-MDP formulation, where two separate encoders for the physical state and belief state are designed to deal with the continuous latent parameter space. It could have been a good benchmark if its encoder design were made available. Apart from the two benchmarks, we also compare with the nominal approach (MLE), where a maximal likelihood estimator for the parameter is computed from the given dataset and then a policy is obtained by solving the MDP with the plugged-in parameter value. In our proposed algorithm (ABDCP) for the infinite-horizon BR-MDP formulation, we consider two particular risk measures, namely expectation and CVaR with different risk levels α. It should be noted that, when the considered risk measure is expectation, our algorithm can be modified and reduced to CALP. Similar observation is verified in Poupart et al. (2006), where the BA-MDP formulation is transformed into a POMDP formulation.

For each of the considered algorithms, we obtain the corresponding optimal policy with the same dataset. It should be noted that the calculations are carried out offline. The obtained policy is then applied for risk-averse path planning and evaluated on the true system, i.e., MDP with the true parameter. This is referred to as one replication, and we repeat the experiments for 200 replications on different independent datasets. Results for the path planning problem can be found in Table 1 and Table 2, with different data size $N = 10$ and $N = 1000$. The columns report the running time, expected performance (cost), and the CVaR performance (cost) of our proposed algorithm and benchmarks over the 200 replications. ABDCP-EXP stands for our proposed algorithm ABDCP with expectation as the risk measure. ABDCP-CVaR stands for our proposed algorithm ABDCP with CVaR as the risk measure. We also show the histogram of the actual performance over 200 replications for our proposed algorithm and the nominal benchmark on the path planning problem in Figure 1. We summarize the main observations for the path planning problem below.

BR-MDP hedges against epistemic uncertainty: in each replication, data points are randomly sampled from the true distribution. While facing the epistemic uncertainty, BR-MDP formulation optimizes over a dynamic risk measure that provides robustness. Table 1 shows that our proposed ABDCP algorithm is the most robust in the sense of balancing the mean and variability of the actual cost. The CVaR cost of our proposed algorithm is also lower than the other benchmarks, showing that it avoids large costs. In contrast, the nominal approach performs badly when the data size is small, e.g. $N = 10$, indicating that it is not robust against the epistemic uncertainty and suffers from the scarcity of data. On the other hand, DR-MDP is overly conservative, even though it has the smallest variability. This conservativeness comes from two aspects. First, it always chooses to optimize over the worst-case scenario, which rarely happens in the true system. Second, the static worst-case risk measure prevents it from adapting to the data realizations, which is one of the motivations for the dynamic risk measure considered in the BR-MDP formulation. In contrast, BR-MDP formulation learns from the future data realization and updates its posterior distribution on θ.

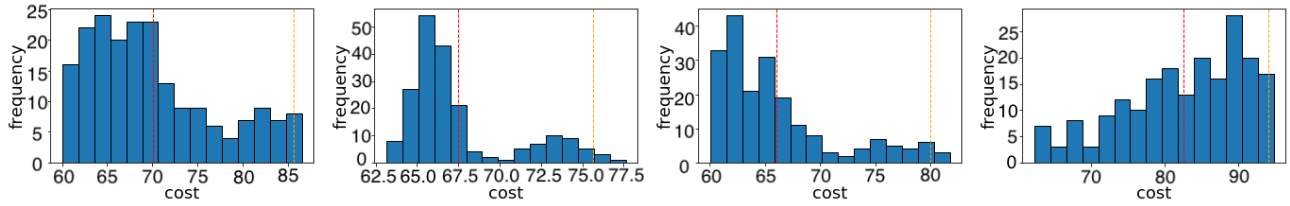
Larger data size reduces epistemic uncertainty: when there are more data, the posterior distribution used in BR-MDP formulation and the MLE estimator used in the nomi-

Approach	time (sec)	expected cost	CVaR ($\alpha = 0.95$) cost	CVaR ($\alpha = 0.8$) cost
ABDCP-EXP (CALP)	969.13(0.18)	70.06(0.51)	85.72	82.06
ABDCP-CVaR ($\alpha = 0.95$)	2639.38(0.22)	67.51(0.24)	**75.67**	**73.72**
ABDCP-CVaR ($\alpha = 0.8$)	2545.74(0.24)	**66.02(0.38)**	79.97	75.50
DR-MDP	62.34(0.11)	79.43(0.15)	81.64	80.60
Nominal	61.44(0.08)	82.59(0.59)	94.10	92.46

Table 1: Results for path planning problem. Running time for each replication, expected cost, and CVaR cost at different risk levels α are reported for different algorithms. Standard errors are reported in parentheses. Number of data points is set to $N = 10$.

Approach	time (sec)	expected cost	CVaR ($\alpha = 0.95$) cost	CVaR ($\alpha = 0.8$) cost
ABDCP-EXP (CALP)	967.25(0.17)	64.15(0.05)	66.34	65.97
ABDCP-CVaR ($\alpha = 0.95$)	2642.26(0.21)	65.18(0.03)	66.14	65.76
ABDCP-CVaR ($\alpha = 0.8$)	2643.48(0.25)	65.17(0.04)	66.26	65.84
DR-MDP	63.15(0.09)	65.22(0.03)	66.43	66.01
Nominal	62.47(0.08)	64.31(0.12)	67.55	65.59

Table 2: Results for path planning problem. Running time for each replication, expected cost, and CVaR cost at different risk levels α are reported for different algorithms. Standard errors are reported in parentheses. Number of data points is set to $N = 1000$.



(a) ABDCP-EXP (b) ABDCP-CVaR($\alpha = 0.95$) (c) ABDCP-CVaR($\alpha = 0.8$) (d) Nominal

Figure 1: Histogram of the actual performance over 200 replications for different algorithms. Number of data points $N = 10$. The left vertical line represents the mean cost. The right vertical line represents the CVaR ($\alpha = 0.95$) cost.

nal approach converge to the true parameter, which reduces to solving an MDP with known transition probability and cost function. Therefore, the optimal policies and the actual costs tend to be the same.

Convergence of ABDCP: the running time for a single replication on the path planning problem using our proposed ABDCP algorithm is affordable, and the proposed algorithm solves the infinite-horizon BR-MDP in finite time. In contrast, the infinite-horizon BR-MDP is intractable with standard value iteration or policy iteration.

Effect of risk measures: although both risk measures (expectation and CVaR) result in time-consistent optimal policy for each considered formulation, they provide different levels of robustness. Even though the expectation case is faster to compute, it provides the least robustness, especially when the data size is small. For the CVaR risk measure, different risk level α also affects the robustness. As α increases, the agent is more risk-averse, and the CVaR cost is smaller since it avoids more severe costs, as is shown in Figure 1(b) and Figure 1(c). But this comes with a price: its expected cost is higher. This is intuitive: even though the agent avoids severe costs, it also forfeits a chance to traverse a path that is likely to have less traffic, even though the likelihood is small. This is shown as a right-shift of the actual performance distribution from Figure 1(c) to Figure 1(b).

5 Conclusion

In this paper, we consider the offline planning problem in MDPs with epistemic uncertainty, where we only have access to a prior belief distribution over MDPs that is constructed by the offline data. We consider the infinite-horizon BR-MDP that produces a time-consistent formulation and provides the robustness against epistemic uncertainty. We develop an efficient optimization-based approximation algorithm that converges to the optimal policy. Our experimental results demonstrate the efficiency of the proposed approximate algorithm, and show the robustness of the infinite-horizon BR-MDP formulation. One of the future directions is to conduct the iteration complexity analysis on the proposed algorithm. Another interesting direction is to utilize function approximation to improve the scalability of the proposed approach to more complex domains. Separate encoders for the physical state and belief state have been proposed in Lee et al. (2019) and adaptation from their risk-neutral BA-MDP formulation to our risk averse BR-MDP formulation could be interesting.

Acknowledgments

The authors gratefully acknowledge the support by the Air Force Office of Scientific Research under Grant FA9550-22-1-0244, the National Science Foundation under Grant NSF-ECCS-2419562 and the NSF AI Institute for Advances in Optimization under Grant NSF-2112533.

References

Ahmadi, M.; Rosolia, U.; Ingham, M. D.; Murray, R. M.; and Ames, A. D. 2021. Constrained risk-averse Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11718–11725.

Alagoz, O.; Ayvaci, M. U.; and Linderoth, J. T. 2015. Optimally solving Markov decision processes with total expected discounted reward function: linear programming revisited. *Computers & Industrial Engineering*, 87: 311–316.

Chow, Y.; and Ghavamzadeh, M. 2014. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, 3509–3517.

Chow, Y.; Tamar, A.; Mannor, S.; and Pavone, M. 2015. Risk-sensitive and robust decision-making: a CVaR optimization approach. In *Advances in Neural Information Processing Systems*, 1522–1530.

Cousins, C.; Lobo, E.; Petrik, M.; and Zick, Y. 2023. Percentile criterion optimization in offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 9322–9352.

Delage, E.; and Mannor, S. 2010. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1): 203–213.

Derman, E.; Mankowitz, D.; Mann, T.; and Mannor, S. 2020. A Bayesian approach to robust reinforcement learning. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, 648–658.

Duff, M. O. 2002. *Optimal learning: computational procedures for Bayes-adaptive Markov decision processes*. Ph.D. diss., University of Massachusetts Amherst.

Föllmer, H.; and Schied, A. 2002. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4): 429–447.

Guigues, V.; Shapiro, A.; and Cheng, Y. 2024. Risk-averse stochastic optimal control: An efficiently computable statistical upper bound. *Operations Research Letters*, 393–400.

Hansen, E. A. 2013. Solving POMDPs by searching in policy space. *arXiv preprint arXiv:1301.7380*.

Horst, R.; and Thoai, N. V. 1999. DC programming: overview. *Journal of Optimization Theory and Applications*, 103(1): 1–43.

Howard, R. A.; and Matheson, J. E. 1972. Risk-sensitive Markov decision processes. *Management Science*, 18(7): 356–369.

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

Lee, G.; Hou, B.; Mandalika, A.; Lee, J.; Choudhury, S.; and Srinivasa, S. S. 2019. Bayesian policy optimization for model uncertainty. In *Proceedings of the International Conference on Learning Representations*.

Lin, Y.; Ren, Y.; and Zhou, E. 2022. Bayesian risk Markov decision processes. In *Advances in Neural Information Processing Systems*, 17430–17442.

Lipp, T.; and Boyd, S. 2016. Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2): 263–287.

Mannor, S.; and Xu, H. 2019. Data-driven methods for Markov decision problems with parameter uncertainty. In *Operations Research & Management Science in the Age of Analytics*, 101–129. INFORMS.

Nilim, A.; and Ghaoui, L. 2004. Robustness in Markov decision problems with uncertain transition matrices. In *Advances in Neural Information Processing Systems*.

Osogami, T. 2012. Robustness and risk-sensitivity in Markov decision processes. In *Advances in Neural Information Processing Systems*.

Petrik, M.; and Russel, R. H. 2019. Beyond confidence regions: tight Bayesian ambiguity sets for robust MDPs. In *Advances in Neural Information Processing Systems*.

Petrik, M.; and Subramanian, D. 2012. An approximate solution method for large risk-averse Markov decision processes. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 805–814.

Pineau, J.; Gordon, G.; Thrun, S.; et al. 2003. Point-based value iteration: an anytime algorithm for POMDPs. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1025–1032.

Poupart, P.; Malhotra, A.; Pei, P.; Kim, K.-E.; Goh, B.; and Bowling, M. 2015. Approximate linear programming for constrained partially observable Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Poupart, P.; Vlassis, N.; Hoey, J.; and Regan, K. 2006. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, 697–704.

Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rigter, M.; Lacerda, B.; and Hawes, N. 2021. Risk-averse Bayes-adaptive reinforcement learning. In *Advances in Neural Information Processing Systems*, 1142–1154.

Rockafellar, R. T.; and Uryasev, S. 2000. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2: 21–41.

Ruszczyński, A. 2010. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125(2): 235–261.

Shapiro, A. 2021. Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming. *European Journal of Operational Research*, 288(1): 1–13.

Shapiro, A.; Dentcheva, D.; and Ruszczynski, A. 2021. *Lectures on stochastic programming: modeling and theory*. SIAM.

Sharma, A.; Harrison, J.; Tsao, M.; and Pavone, M. 2019. Robust and adaptive planning under model uncertainty. In *Proceedings of the 29th International Conference on Automated Planning and Scheduling*, 410–418.

Shen, X.; Diamond, S.; Gu, Y.; and Boyd, S. 2016. Disciplined convex-concave programming. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 1009–1014. IEEE.

Tamar, A.; Chow, Y.; Ghavamzadeh, M.; and Mannor, S. 2015. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*.

Tamar, A.; Glassner, Y.; and Mannor, S. 2015. Optimizing the CVaR via sampling. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2993–2999.

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 *Advances in Neural Information Processing Systems*.

Yang, W.; Zhang, L.; and Zhang, Z. 2022. Toward theoretical understandings of robust Markov decision processes: sample complexity and asymptotics. *The Annals of Statistics*, 50(6): 3223–3248.

Zhang, J.; Bedi, A. S.; Wang, M.; and Koppel, A. 2021. Cautious reinforcement learning via distributional risk in the dual domain. *IEEE Journal on Selected Areas in Information Theory*, 2(2): 611–626.

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 *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 3331–3339.