Distributionally Robust Counterpart in Markov Decision Processes

Pengqian Yu and Huan Xu

***Abstract*—This technical note studies Markov decision processes under parameter uncertainty. We adapt the distributionally robust optimization framework, assume that the uncertain parameters are random variables following an unknown distribution, and seek the strategy which maximizes the expected performance under the most adversarial distribution. In particular, we generalize a previous study [1] which concentrates on distribution sets with very special structure to a considerably more generic class of distribution sets, and show that the optimal strategy can be obtained efficiently under mild technical conditions. This significantly extends the applicability of distributionally robust MDPs by incorporating probabilistic information of uncertainty in a more flexible way.**

***Index Terms*—Distributional robustness, Markov decision processes, parameter uncertainty.**

I. Introduction

Markov Decision Processes (MDPs) are widely used tools to model stochastic sequential decision making problems (e.g., [2]–[4]). A strategy that achieves maximal expected accumulated reward is considered optimal. However, in practice, the transition probabilities and reward parameters are typically estimated from finite and possibly noisy data, which often deviate from their true values. Such deviation, called "parameter uncertainty," can cause the performance of the optimal policies to degrade significantly (see experiments in [5]).

Inspired by the "robust optimization" framework in mathematical programming (e.g., [6]–[9]), many efforts have been made to alleviate the effect of parameter uncertainty in MDPs (e.g., [1], [10]–[15]). Most previous study (e.g., [10], [11], [14]–[16]) focuses on the "robust MDP" which treats the uncertain parameter as a fixed yet unknown element of a given "uncertainty set," and aims to find the strategy that achieves best performance under the worst parameter. This set-inclusive formulation of uncertainty can be conservative as it cannot incorporate probabilistic information of the uncertainty that is often available in practice (e.g., [12], [17]). To overcome this, [1] proposed the *distributionally robust MDP* approach, which can incorporate certain kind of probabilistic information of the uncertainty. More specifically, this approach treats the uncertain parameters as a random variable *following an unknown distribution*, while the distribution is known to belong to a set of distributions, called the "ambiguity set," and the goal is to seek a strategy that archives the maximum expected performance under the most adversarial distribution of the uncertain parameters. Indeed, this approach is the multi-stage counter-part of the distributionally robust optimization (e.g., [1], [18]) which considers the following: Given a utility function $u(x,\xi)$ where $x \in \mathcal{X}$ is the optimizing variable and ξ is the unknown parameter, distributionally robust optimization solves $\max_{x\in\mathcal{X}}[\inf_{\mu\in\mathcal{C}} \mathbb{E}_{\xi\sim\mu} u(x,\xi)]$, where $\mathcal{C}$ is an *a priori* known set of distributions.

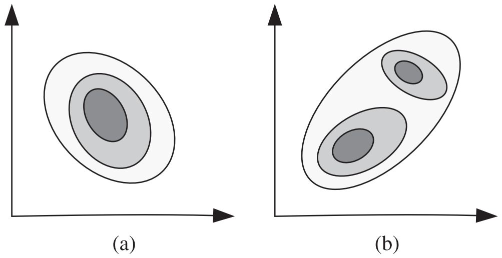

(a) (b)

Fig. 1. Illustration of the confidence sets.

We highlight our contributions by comparing with [1]. In [1] the state-wise ambiguity set is restricted to the following form: $\tilde{\mathcal{C}}_s = \{\mu_s | \mu_s(O_s^i) \geq \underline{\alpha}_s^i \ \forall i = 1,\ldots,n_s\}$, where $\underline{\alpha}_s^i \leq \underline{\alpha}_s^j$ and O_s^i is a proper set of uncertain parameters with a "nested-set" structure, i.e., satisfying $O_s^i \subseteq O_s^j$, for all $i < j$ [see Fig. 1(a)]. This setup can effectively model distributions with a single mode (such as a Gaussian distribution), but less so when modeling multi-mode distributions such as a mixture Gaussian distribution. Moreover, other probabilistic information such as mean, variance etc. cannot be incorporated. Thus, in this technical note, we extend the distributionally robust MDP approach to handle ambiguity sets with more general structures. In particular, we consider a class of ambiguity sets, first proposed in [18] as a unifying framework for modeling and solving distributionally robust single-stage optimization problems, and embed them into the distributionally robust MDPs setup. These ambiguity sets are considerably more general: they are characterized by a class of O_s^i which can either be nested or disjoint [as shown in Fig. 1(b)], and moreover, additional linear constraints are allowed to define the ambiguity set, which can be used to incorporate probabilistic information such as mean, covariance or other variation measures. We show that, under this more general class of ambiguity sets, the resulting distributionally robust MDPs remain tractable under mild technical conditions, and often outperform previous methods thanks to the fact that it can model uncertainty in a more flexible way.

II. Preliminaries

Throughout the technical note, we use capital letters to denote matrices, and bold face letters to denote column vectors. We use $\mathbf{e}_i(m)$ to denote the ith elementary vector of length m, and use $\mathbb{R}_+^n$ to denote the nonnegative orthant of $\mathbb{R}^n$. If $\mathcal{C}$ is the set of joint probability distributions of three random vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, then $\prod_{(\mathbf{a},\mathbf{b})} \mathcal{C}$ denotes the set of marginal distributions of $(\mathbf{a},\mathbf{b})$. We use $\oplus$ to represent mixture distribution: given two probability distributions $\mathcal{F}_1, \mathcal{F}_2$ and a Bernoulli random variable x which takes value 1 w.p. p, $x\mathcal{F}_1 \oplus (1-x)\mathcal{F}_2$ is a random variable such that it follows distribution $\mathcal{F}_1$ w.p. p, and follows $\mathcal{F}_2$ w.p. $1-p$. We use $\mathcal{N}(m,\sigma^2)$ to represent a Gaussian distribution with mean m and variance σ^2.

A (finite) Markov Decision Process (MDP) is defined as a 6-tuple $\langle T, \gamma, S, A, \mathbf{p}, \mathbf{r}\rangle$. Here, T is the (possibly infinite) decision horizon;

Manuscript received May 14, 2015; revised September 18, 2015; accepted October 16, 2015. Date of publication October 27, 2015; date of current version August 26, 2016. This work was supported in part by the Ministry of Education of Singapore through AcRF Tier Two Grant R266000098112. Recommended by Associate Editor H. S. Chang.

P. Yu is with the Department of Mechanical Engineering, National University of Singapore, Singapore 117575 (e-mail: yupengqian@nus.edu.sg).

H. Xu is with the Department of Industrial and Systems Engineering, National University of Singapore, Singapore 117575 (e-mail: isexuh@nus.edu.sg).

Color versions of one or more of the figures in this technical note are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2015.2495174

$\gamma \in (0, 1]$ is the discount factor; S is the state set and A_s is the action set of state $s \in S$, both assumed to be finite. The parameter $\mathbf{p}$ and $\mathbf{r}$ are the transition probability and the expected reward, respectively. That is, for $s \in S$ and $a \in A_s$, $r(s, a)$ is the expected reward and $p(s'|s, a)$ is the probability that the next state is s'. Following [2], we denote the set of all history-dependent randomized strategies by Π^{HR}. We use subscript s to denote the value associated with the state s: e.g., $\mathbf{r}_s$ denotes the vector form of the rewards associated with the state s, and π_s is the (randomized) action chosen at state s for strategy π. The elements in the vector $\mathbf{p}_s$ are listed in the following way: the transition probabilities of the same action are arranged in the same block, and inside each block they are listed according to the order of the next state. We use $\underline{s}$ to denote the (random) state following s, and $\Delta(s)$ to denote the probability simplex on A_s. We use $\bigotimes$ to represent Cartesian product, e.g., $\mathbf{p} = \bigotimes_{s \in S} \mathbf{p}_s$. For a given strategy $\pi \in \Pi^{HR}$, we denote the expected (discounted) total-reward under parameters pair $(\mathbf{p}, \mathbf{r})$ as $u(\pi, \mathbf{p}, \mathbf{r}) \triangleq \mathbb{E}_{\pi}^{(\mathbf{p},\mathbf{r})}\{\sum_{i=1}^{T} \gamma^{i-1} r(s_i, a_i)\}$.

A Distributionally Ambiguous MDP (DAMDP) is defined as a tuple $\langle T, \gamma, S, A, \tilde{\mathcal{C}}_S \rangle$, where the transition probability $\mathbf{p}$ and the expected reward $\mathbf{r}$ are unknown. Instead, they are assumed to obey a joint distribution μ_0 (also unknown) that belongs to a known ambiguity set $\mathcal{C}_S \triangleq \prod_{(\mathbf{p},\mathbf{r})} \tilde{\mathcal{C}}_S$.

While the DAMDP framework can be very general, most $\tilde{\mathcal{C}}_S$ result in formulations that are computationally intractable (e.g., [1], [19]). Hence, we make the following requirement of $\tilde{\mathcal{C}}_S$ such that the parameters among different states are independent.

Assumption 1: The ambiguity set $\tilde{\mathcal{C}}_S$ has the following property:

$$\tilde{\mathcal{C}}_S \triangleq \left\{ \mu \middle| \mu = \bigotimes_{s \in S} \mu_s, \mu_s \in \tilde{\mathcal{C}}_s, \ \forall\, s \in S \right\}$$

where "state-wise ambiguity set" $\tilde{\mathcal{C}}_s$ is a set of distributions of parameters of state s.

By the definition of $\tilde{\mathcal{C}}_S$, the state-wise property applies to $\mathcal{C}_S$ as well. This property is the same as the concept of "s-rectangularity" in [16], and is essential for reducing DAMDP to robust MDP in Lemma 1. In addition, [20] showed that the robust MDP with coupled uncertainty sets is computationally challenging, which implies solving DAMDP with nonrectangular ambiguity sets is even harder.

We now discuss the admissible state-wise ambiguity set. Our formulation of the state-wise ambiguity set follows the unifying framework of [18]. In specific, given $s \in S$, the state-wise ambiguity set is representable with the following standard form:

$$\tilde{\mathcal{C}}_s \triangleq \left\{ \mu_s \begin{pmatrix} \mathbf{p}_s \\ \mathbf{r}_s \\ \tilde{\mathbf{u}}_s \end{pmatrix} \middle| \begin{array}{c} \mathbb{E}_{(\mathbf{p}_s,\mathbf{r}_s,\tilde{\mathbf{u}}_s)\sim\mu_s}[F_s\mathbf{p}_s + G_s\mathbf{r}_s \\ +H_s\tilde{\mathbf{u}}_s] = \mathbf{c}_s, \\ \mu_s\left(O_s^i\right) \in [\underline{\alpha}_s^i, \overline{\alpha}_s^i], \quad \forall\, i \in I_s \end{array} \right\}. \tag{1}$$

Here, $F_s \in \mathbb{R}^{k\times(|A_s|\times|s|)}$, $G_s \in \mathbb{R}^{k\times|A_s|}$, $H_s \in \mathbb{R}^{k\times Q}$, $\mathbf{c}_s \in \mathbb{R}^k$; $I_s = \{1, 2, \ldots, n_s\}$ is an index set and $O_s^i \subseteq \mathbb{R}^{|A_s|\times|s|} \times \mathbb{R}^{|A_s|} \times \mathbb{R}^Q$ is a set of possible values of the parameters $(\mathbf{p}_s, \mathbf{r}_s, \tilde{\mathbf{u}}_s)$, termed "confidence set"; $\underline{\alpha}_s^i, \overline{\alpha}_s^i \in [0, 1]$, $\underline{\alpha}_s^i \leq \overline{\alpha}_s^i$ for all $i \in I_s$, are the lower and upper bounds of the probability that parameters belong to a confidence set. Thus, each confidence set O_s^i provides an estimation of the uncertain parameters pair $(\mathbf{p}_s, \mathbf{r}_s, \tilde{\mathbf{u}}_s)$ subject to a different confidence level. Ambiguity sets $\tilde{\mathcal{C}}_s$ contain prescribed conic representable confidence sets and mean values residing on an affine manifold, which is rich enough to encompass and extend several ambiguity sets considered in recent literature (e.g., [1], [19], [21]). The set of joint distribution of $(\mathbf{p}_s, \mathbf{r}_s)$ is hence $\mathcal{C}_s \triangleq \prod_{(\mathbf{p}_s,\mathbf{r}_s)} \tilde{\mathcal{C}}_s$. Notice that a classical technique called "lifting" is used here: We introduce an auxiliary random vector $\tilde{\mathbf{u}}$, so that some non-linear relationship can be modeled linearly. For example, a constraint on the variance can be modeled using this standard form (see [22, Example 2]), which is otherwise impossible without the auxiliary variable. This lifting technique thus allows us to model a rich variety of structural information about the marginal distribution of $(\mathbf{p}, \mathbf{r})$ in a unified manner. Note when the ambiguity set only contains the support of random variables, i.e., $\tilde{\mathcal{C}}_s = \{\mu_s(\mathbf{p}_s, \mathbf{r}_s, \tilde{\mathbf{u}}_s) | \mu_s(O_s^1) = 1, \ \forall\, i \in I_s, s \in S\}$, DAMDP reduces to classical robust MDP formulation, where the a-priori information of unknown parameters is that they belong to an uncertainty set.

Assumptions 2 to 4 are standard requirements for the confidence sets, proposed in [18]. The first one asserts the relationship between different confidence sets.

Assumption 2 (Nesting Condition): For any $s \in S$, all $i, i' \in I_s$ and $i \neq i'$, we have either $O_s^i \Subset O_s^{i'}$, $O_s^{i'} \Subset O_s^i$ or $O_s^i \cap O_s^{i'} = \emptyset$.

Here $O_s^i \Subset O_s^{i'}$ means that a set O_s^i is strictly included in a set $O_s^{i'}$, i.e., O_s^i is contained in the interior of $O_s^{i'}$. The nesting condition is illustrated in Fig. 1(b). Next, for any $s \in S$ we require that $\tilde{\mathcal{C}}_s$ satisfies the following regularity condition.

Assumption 3 (Regularity Conditions for $\tilde{\mathcal{C}}_s$):

1) The confidence set $O_s^{n_s}$ is bounded and has probability one, that is, $\underline{\alpha}_s^{n_s} = \overline{\alpha}_s^{n_s} = 1$.
2) There is a probability distribution $\mu_s(\mathbf{p}_s, \mathbf{r}_s, \tilde{\mathbf{u}}_s) \in \tilde{\mathcal{C}}_s$ such that $\mu_s(O_s^i) \in (\underline{\alpha}_s^i, \overline{\alpha}_s^i)$ whenever $\underline{\alpha}_s^i < \overline{\alpha}_s^i$, $i \in I_s$.

The condition 1 of Assumption 3 ensures the confidence set with largest index, $O_s^{n_s}$, contains the support of the joint unknown parameters pair $(\mathbf{p}_s, \mathbf{r}_s, \tilde{\mathbf{u}}_s)$. The second condition stipulates that there is a probability distribution $\mu_s(\mathbf{p}_s, \mathbf{r}_s, \tilde{\mathbf{u}}_s) \in \tilde{\mathcal{C}}_s$ that satisfies the probability bounds in (1) as strict inequalities whenever the corresponding probability interval $[\underline{\alpha}_s^i, \overline{\alpha}_s^i]$ is non-degenerate. For each individual O_s^i, we make the following assumption to ensure tractability.

Assumption 4: For $s \in S$, $i \in I_s$, each nonempty and convex confidence set O_s^i is defined as

$$O_s^i = \left\{ \begin{pmatrix} \mathbf{p}_s \\ \mathbf{r}_s \\ \tilde{\mathbf{u}}_s \end{pmatrix} \in \begin{array}{c} \mathbb{R}^{|A_s|\times|s|} \\ \times\mathbb{R}^{|A_s|} \\ \times\mathbb{R}^Q \end{array} \middle| \begin{array}{c} B_s^i\mathbf{p}_s + D_s^i\mathbf{r}_s \\ +E_s^i\tilde{\mathbf{u}}_s \preceq_{K_s^i} \mathbf{b}_s^i \end{array} \right\}$$

where $B_s^i \in \mathbb{R}^{L_i\times(|A_s|\times|s|)}$, $D_s^i \in \mathbb{R}^{L_i\times|A_s|}$, $E_s^i \in \mathbb{R}^{L_i\times Q}$, $\mathbf{b}_s^i \in \mathbb{R}^{L_i}$, K_s^i are proper cones (i.e., a closed, convex and pointed cone with nonempty interior).

III. Finite Horizon Distributionally Robust MDPs

This section focuses on DAMDP with a finite number of decision stages. We show that a strategy defined through backward induction, which we call S-robust strategy, is distributionally robust. We further show such a strategy is solvable in polynomial time under mild technical conditions. This generalizes results in [1] to a significantly more general class of ambiguity sets.

Similar to [10], we assume that when a state is visited multiple times, each time it can take a different parameter realization (nonstationary model). This assumption is justified mainly because the stationary model is generally intractable and a lower-bound of it is given by the non-stationary model. Therefore, multiple visits to a state can be treated as visiting different states. By introducing dummy states as in [1, Assumption 2.2], for finite horizon DAMDP we make the following assumption without loss of generality. This will simplify our exposition.

Assumption 5: 1) Each state belongs to only one stage. 2) The terminal reward equals zero. 3) The first stage only contains one state s^{ini}.

Using the condition 1 of Assumption 5, we partition S according to the stage each state belongs to. That is, we let S_t be the set of states belong to tth stage.

For $\pi \in \Pi^{HR}$ and $\mu \in \mathcal{C}_S$, we denote the expected performance of a DAMDP as

$$w(\pi, \mu, (s^{\mathrm{ini}})) \triangleq \mathbb{E}_{(\mathbf{p},\mathbf{r})\sim\mu}\{u(\pi, \mathbf{p}, \mathbf{r})\} = \int u(\pi, \mathbf{p}, \mathbf{r}) d\mu(\mathbf{p}, \mathbf{r}).$$

Definition 1: A strategy $\pi^* \in \Pi^{HR}$ is *distributionally robust* with respect to $\mathcal{C}_S$ if it satisfies that for all $\pi \in \Pi^{HR}$, $\inf_{\mu \in \mathcal{C}_S} w(\pi, \mu, (s^{\mathrm{ini}})) \le \inf_{\mu' \in \mathcal{C}_S} w(\pi^*, \mu', (s^{ini}))$.

In words, each strategy is evaluated by its expected performance under the (respective) most adversarial distribution of the uncertain parameters, and a distributionally robust strategy is the optimal strategy according to this metric. The main focus of this section is deriving approaches to solve the distributionally robust strategy. To this end, we need the following definition.

Definition 2: Given a DAMDP $\langle T, \gamma, S, A, \tilde{\mathcal{C}}_S \rangle$, we define the *S- robust strategy* as follows

1) For $s \in S_T$, the *S- robust value* $\tilde{v}_T(s) \triangleq 0$.
2) For $s \in S_t$, where $t < T$, the *S- robust value* $\tilde{v}_t(s)$ and *S- robust action* $\tilde{\pi}_s$ are defined as

$$\tilde{v}_t(s) \triangleq \max_{\pi_s \in \Delta(s)} \left\{ \min_{\mu_s \in \mathcal{C}_s} \mathbb{E}_{(\mathbf{p}_s, \mathbf{r}_s)\sim\mu_s} \times \left\{ \mathbb{E}_{\pi_s}^{(\mathbf{p}_s, \mathbf{r}_s)} \left[r(s,a) + \gamma \tilde{v}_{t+1}(\underline{s}) \right] \right\} \right\}$$
$$\tilde{\pi}_s \in \arg\max_{\pi_s \in \Delta(s)} \left\{ \min_{\mu_s \in \mathcal{C}_s} \mathbb{E}_{(\mathbf{p}_s, \mathbf{r}_s)\sim\mu_s} \times \left\{ \mathbb{E}_{\pi_s}^{(\mathbf{p}_s, \mathbf{r}_s)} \left[r(s,a) + \gamma \tilde{v}_{t+1}(\underline{s}) \right] \right\} \right\}. \quad (2)$$

3) A strategy $\tilde{\pi}^*$ is a *S- robust strategy* if $\forall\, s \in S$, and every history h that ends at s, we have $\tilde{\pi}_s^*$, conditioned on history h, is a S-robust action.

The definition requires that the strategy must be robust w.r.t. each sub-problem, and hence the name "S-robust." The following theorem shows any S-robust strategy π^* is distributionally robust, and is the main result of this technical note.

Theorem 1: Let $T < \infty$. Under Assumptions 1, 2, 4, and 5, if π^* is a S-robust strategy, then

1) π^* is a distributionally robust strategy with respect to $\mathcal{C}_S$.
2) There exists $\mu^* \in \mathcal{C}_s$ such that (π^*, μ^*) is a saddle point. That is

$$\sup_{\pi \in \Pi^{HR}} w\left(\pi, \mu^*, (s^{\mathrm{ini}})\right) = w\left(\pi^*, \mu^*, (s^{\mathrm{ini}})\right) = \inf_{\mu \in \mathcal{C}_S} w\left(\pi^*, \mu, (s^{\mathrm{ini}})\right).$$

Proof: We first state a Lemma from [1, Lemma 3.2] without proof.

Lemma 1: Under Assumption 1, fix $\pi \in \Pi^{HR}$ and $\mu \in \mathcal{C}_S$, denote $\overline{\mathbf{p}} = \mathbb{E}_\mu(\mathbf{p})$ and $\overline{\mathbf{r}} = \mathbb{E}_\mu(\mathbf{r})$. We have $w(\pi, \mu, (s^{\mathrm{ini}})) = u(\pi, \overline{\mathbf{p}}, \overline{\mathbf{r}})$.

Lemma 1 means for any strategy, the expected performance under an admissible distribution μ only depends on the expected value of parameters under μ. Thus, the distributionally robust MDPs reduce to robust MDPs. Next we characterize the set of expected value of the parameters.

Lemma 2: For $s \in S$ and $\pi_s \in \Delta(s)$, we define the set $\mathcal{Z}_s = \{\mathbb{E}_{\mu_s}(\mathbf{p}_s, \mathbf{r}_s) | \mu_s \in \mathcal{C}_s\}$. Then set $\mathcal{Z}_s$ is convex and compact.

Proof: First, we show that, for $s \in S$ and $\pi_s \in \Delta(s)$, the set defined as $\tilde{\mathcal{Z}}_s = \{\mathbb{E}_{\mu_s}(\mathbf{p}_s, \mathbf{r}_s, \tilde{\mathbf{u}}_s) | \mu_s \in \tilde{\mathcal{C}}_s\}$ is convex and compact. The convexity can be easily shown, which is omitted due to space constraints (see [22] for details). To show the compactness, notice that $\tilde{\mathcal{C}}_s$ is weakly closed (i.e., closed w.r.t. to the weak topology) since the feasible set of each of constraint is weakly closed which implies their intersection is also weakly closed. Thus, $\tilde{\mathcal{Z}}_s$ is closed since it is the image of $\tilde{\mathcal{C}}_s$ under expectation (which is a continuous function). This implies $\tilde{\mathcal{Z}}_s$ is compact since $O_s^{n_s}$ is bounded and hence $\tilde{\mathcal{Z}}_s$ is bounded. Finally, since $\mathcal{Z}_s$ is the projection onto the first two coordinates of set $\tilde{\mathcal{Z}}_s$, its convexity and compactness thus follow. ■

Lemma 2 implies that, for $s \in S$ and $\pi_s \in \Delta(s)$, there exists $(\mathbf{p}_s^*, \mathbf{r}_s^*) \in \mathcal{Z}_s$ that satisfies $\inf_{(\mathbf{p}_s, \mathbf{r}_s) \in \mathcal{Z}_s} u(\pi_s, \mathbf{p}_s, \mathbf{r}_s) = u(\pi_s, \mathbf{p}_s^*, \mathbf{r}_s^*)$. Since saddle point of the minimax objective exists for robust MDPs (e.g., [10], [11]), we can complete the proof of part 2) following a similar procedure as the last portion of proof for [1, Theorem 3.1]. We omit the details due to space constraint (see [22] for details). Part 1) then follows part 2) immediately. ■

We now investigate the computational aspect of finding the S-robust action.

Theorem 2: Under Assumption 2, 3, 4, and 5, for $s \in S_t$ where $t < T$, the S-robust action is the optimal solution of the following optimization problem (termed *S- robust problem* hereafter):

$$\begin{aligned}
\underset{w,\pi_s,\beta,\kappa,\lambda,\nu_i}{\text{minimize}} \quad & w \\
\text{subject to} \quad & \mathbf{c}_s^\top \beta + \sum_{i \in I_s} \left[\overline{\alpha}_s^i \kappa_i - \underline{\alpha}_s^i \lambda_i \right] \le w \\
& \nu_i^\top \mathbf{b}_s^i - \sum_{i' \in A(i)} \left[\kappa_{i'} - \lambda_{i'} \right] \le 0, \quad i \in I_s \\
& {B_s^i}^\top \nu_i + \tilde{V}_s \pi_s + F_s^\top \beta = 0, \quad i \in I_s \\
& {D_s^i}^\top \nu_i + \pi_s + G_s^\top \beta = 0, \quad i \in I_s \\
& {E_s^i}^\top \nu_i + H_s^\top \beta = 0, \quad i \in I_s \\
& \pi_s \in \Delta(s),\ \beta \in \mathbb{R}^k,\ \kappa, \lambda \in \mathbb{R}_+^{n_s},\ \nu_i \in {K_s^i}^*. \quad (3)
\end{aligned}$$

Here, ${K_s^i}^*$ represents the cone dual to K_s^i; set $A(i) \triangleq \{i\} \cup \{i' \in I_s : O_s^i \Subset O_s^{i'}\}$; $\tilde{\mathbf{v}}_{t+1}$ is the vector form of $\tilde{v}_{t+1}(s')$ for all $s' \in S_{t+1}$; and $\tilde{V}_s \triangleq [\mathbf{e}_1(|A_s|)\tilde{\mathbf{v}}_{t+1}^\top, \ldots, \mathbf{e}_{|A_s|}(|A_s|)\tilde{\mathbf{v}}_{t+1}^\top]^\top$.

Proof: The proof essentially follows from [18] and duality of convex optimization [23], and can be found in the longer version [22] of this technical note. ■

Thus, since for $s \in S_t$, $\Delta(s)$ is compact, we can solve the S-robust action in polynomial time if all K_s^i are "easy" cones such as linear, conic quadratic or semidefinite cones. Moreover, using Theorem 1, by backward induction, we can obtain the S-robust strategy efficiently.

By virtue of the lifting technique [18, Theorem 5], we show below several widely used ambiguity sets are indeed special cases of $\tilde{\mathcal{C}}_s$ defined in (1). We further derive their corresponding S-robust problems. See [22] for additional examples (variance and expected Huber loss function).

Example 1 (Mean Absolute Deviation): Assume that $\mathbb{E}_{\mathbf{r}_s \sim \mu_s(\mathbf{r}_s)}[|\mathbf{r}_s - \mathbf{m}|] \le \mathbf{f}$ for $\mathbf{m}, \mathbf{f} \in \mathbb{R}^{|A_s|}$. [18] shows that $\tilde{\mathcal{C}}_s$, which involves the auxiliary random vector $\tilde{\mathbf{u}}_s \in \mathbb{R}^{|A_s|}$, can be expressed as $\tilde{\mathcal{C}}_s = \{\mu_s(\mathbf{r}_s, \tilde{u}_s) | \mathbb{E}_{\tilde{\mathbf{u}}_s \sim \mu_s}[\tilde{\mathbf{u}}_s] = \mathbf{f}, \mu_s(\tilde{\mathbf{u}}_s \ge \mathbf{r}_s - \mathbf{m}, \tilde{\mathbf{u}}_s \ge \mathbf{m} - \mathbf{r}_s) = 1\}$. Note that $\mu_s(\mathbf{r}_s) \in \prod_{\mathbf{r}_s} \tilde{\mathcal{C}}_s$. In this case Problem (3) can be rewritten as

$$\begin{aligned}
\underset{w,\pi_s,\kappa,\nu}{\text{minimize}} \quad & w \\
\text{subject to} \quad & \kappa - \mathbf{f}^\top \nu \le w \\
& \kappa + \mathbf{p}_s^\top \tilde{V}_s \pi_s + \mathbf{m}^\top \pi_s \ge 0 \\
& \pi_s \in \Delta(s), \quad \nu \ge 0.
\end{aligned}$$

Example 2 (Mean): Assume that we only know a noisy empirical estimator of the exact mean of $\mathbf{p}_s$. That is, given $G \in \mathbb{R}^{M\times(|A_s|\times|s|)}$, $\mathbf{f} \in \mathbb{R}^M$ and $\mathbf{p}_s \sim \mu_s(\mathbf{p}_s)$, $G\mathbb{E}_{\mathbf{p}_s\sim\mu_s(\mathbf{p}_s)}[\mathbf{p}_s] \preceq_K \mathbf{f}$, where K is a proper cone. [18] shows that $\tilde{\mathcal{C}}_s$, which involves the auxiliary random vector $\tilde{\mathbf{u}}_s \in \mathbb{R}^M$, can be expressed as $\tilde{\mathcal{C}}_s = \{\mu_s(\mathbf{p}_s, \tilde{\mathbf{u}}_s) | \mathbb{E}_{\tilde{\mathbf{u}}_s\sim\mu_s}[\tilde{\mathbf{u}}_s] = \mathbf{f}, \mu_s(G\mathbf{p}_s \preceq_K \tilde{\mathbf{u}}_s) = 1\}$. Note that $\mu_s(\mathbf{p}_s) \in \prod_{\mathbf{p}_s} \tilde{\mathcal{C}}_s$. Problem (3) now takes the form

$$\begin{aligned}
\underset{w,\pi_s,\kappa,\nu}{\text{minimize}} \quad & w \\
\text{subject to} \quad & \kappa + \mathbf{f}^\top \nu \le w \\
& \kappa + \mathbf{r}_s^\top \pi_s \ge 0 \\
& \tilde{V}_s \pi_s + G^\top \nu = 0 \\
& \pi_s \in \Delta(s), \quad \nu \in K^*.
\end{aligned}$$

This example can also be treated via "classical" robust optimization by virtue of Lemma 1.

The finite horizon DAMDP can be easily extended to discounted-reward infinite horizon setup. We can generalize the notion of S-robust strategy, which turns to be distributionally robust in both stationary and non-stationary models. This extension is similar to [1] and can be found in [22].

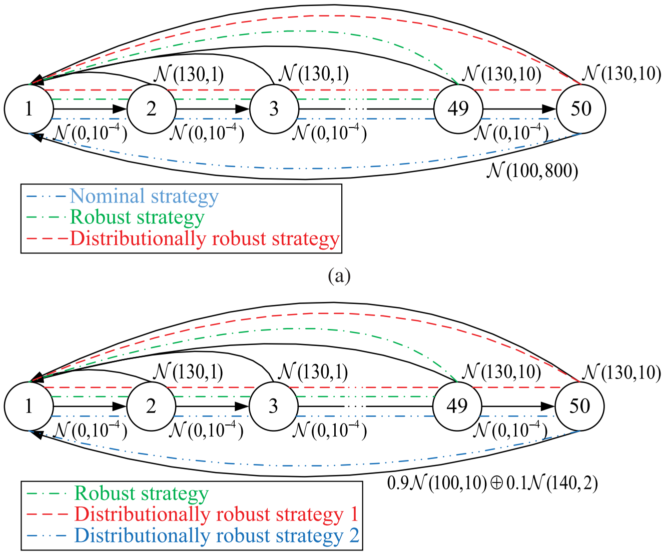
IV. Simulation

In this section, we study two synthetic numerical examples: a machine replacement problem and a path planning problem. In the machine replacement problem, the reward parameters are uncertain; whereas in the path planning problem, the transition probabilities are uncertain. All results were generated on desktop with Intel Core i5-3570 CPU of 3.40 GHz clock speed and 8 GB RAM. The S-robust problems are solved in Matlab using the CVX package [24].

A. Reward Uncertainty in the Machine Replacement Problem

We consider a machine replacement problem similar to the one in [12]. Consider the repair cost incurred by a factory that holds a large number of machines, given that each of these machines is modeled with a same underlying MDP for which rewards are subject to uncertainty.

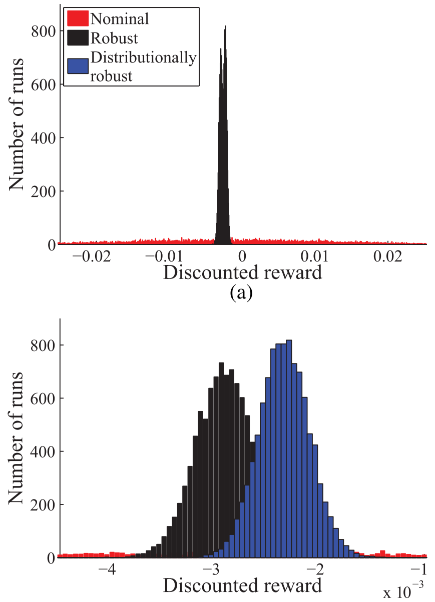
1) Machine Replacement as a MDP With Gaussian Rewards: We first consider a machine replacement problem with 50 states, 2 actions ("repair" and "not repair") for each state, deterministic transitions, a discount factor of 0.8, and uncertain rewards following Gaussian distributions independently [see Fig. 2(a)]: For the first 48 states, the "repair" action has a cost $\mathcal{N}(130, 1)$. The 49th and 50th states of the machine's life are designed to be risky: not repairing at state 50 incurs a highly uncertain cost $\mathcal{N}(100, 800)$, while repairing at both states is a more secure but still uncertain option with a cost $\mathcal{N}(130, 10)$. The detailed implementation is as follows: We use the mean value of uncertain rewards to compute the nominal strategy. For both robust and distributionally robust strategy, we construct confidence sets using $\hat{m} \pm 3\hat{\sigma}$ for the first 49 states, and $\hat{m} \pm 4\hat{\sigma}$ for state 50 where $\hat{m}$ and $\hat{\sigma}^2$ are mean and variance estimated from samples (see [22] for details), as it is more risky and thus hard to estimate. In addition, we construct an extra confidence set (centered at the mean) with 60%–70% confidence level (i.e., $\underline{\alpha}_{50}^1 = 0.6$, $\overline{\alpha}_{50}^1 = 0.7$) for distributionally robust strategy. The optimal paths followed by three strategies are shown in Fig. 2(a).



(a)

(b)

Fig. 2. Two instances of a machine replacement problem. Fig. 2(a) shows Gaussian uncertainty in the rewards, while Fig. 2(b) shows mixed Gaussian uncertainty in the rewards.



(a)

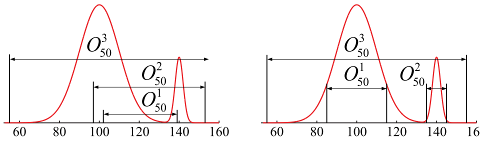
(b)

Fig. 3. Performance comparisons between nominal, robust, and distributionally robust strategies on 10,000 runs of the machine replacement problem with Gaussian rewards (The bottom figure focuses on the interval [−0.0045, −0.001]).

The performance of the strategies obtained by using the nominal, the robust and the distributionally robust approaches is presented in Fig. 3. The corresponding average total discounted rewards and computational times are shown in Table I. The nominal strategy results in the highest average total discounted rewards. This is well expected as we are using the exact mean value of the reward as the nominal parameter. However, the nominal strategy is highly risky: it cannot prevent bad performance (e.g., −0.025) from happening, which is undesirable. While the nominal strategy, blind to any form of risk, finds no advantage in ever repairing, the robust strategy ends up following a highly conservative policy (repairing the machine at state 49 to avoid state 50). In contrast, the distributionally robust optimal strategy makes use of more distributional information and handles the risk efficiently by waiting until state 50 and then repair the machine. Therefore, this strategy beats the nominal and robust strategies in that it strikes a good tradeoff between high mean reward and low variance over 10,000 different trials. These results coincide with what one would typically expect from the three solution concepts.

TABLE I
AVERAGE TOTAL DISCOUNTED REWARDS AND COMPUTATIONAL TIMES OF NOMINAL, ROBUST, AND DISTRIBUTIONALLY ROBUST STRATEGIES IN MACHINE REPLACEMENT PROBLEM WITH GAUSSIAN REWARDS

Strategies	Nominal	Robust	Distributionally robust
Average total discounted rewards	-1.8×10^{-3}	-2.9×10^{-3}	-2.3×10^{-3}
Computational times (seconds)	0.643	815	820



(a) (b)

Fig. 4. Illustration of the confidence sets for two distributionally robust strategies.

2) Machine Replacement as a MDP With Mixed Gaussian Rewards: The second experiment has a similar setup as the previous one, except that not repairing at the 50th state has a reward which follows a *mixed Gaussian distribution* [see Fig. 2(b)]. This experiment illustrates the effect of the two different nested-set structures shown in Fig. 1. In specific, we apply the two different distributionally robust approaches (proposed in [1] and this technical note respectively), and show that our method outperforms. The detailed implementation is as follows: For the robust and two distributionally robust strategies, we construct uncertainty set corresponding to 99% probability support of the rewards for the first 49 states, and 99.9% for the 50th state that is more risky, using estimated mean and variance (see [22] for details). For the first distributionally robust strategy proposed in [1], we construct two additional *nested* confidence sets O^1_{50} and O^2_{50} [see Fig. 4(a)], which w.p. 40%–50% and 60%–70% respectively the uncertain rewards belong to. In contrast, for the second distributionally robust strategy proposed in this technical note, we construct two *disjoint* confidence sets O^1_{50} and O^2_{50} [see Fig. 4(b)] with 70%–80% and 0%–10% confidence level, respectively. Specifically, we select these two intervals around the peaks of the two Gaussian elements [i.e., $\mathcal{N}(100, 10)$ and $\mathcal{N}(140, 2)$] to better model this mixed distribution. The optimal paths followed for the three strategies are shown in Fig. 2(b).

The performance of the three strategies obtained is presented in Fig. 5. The corresponding average total discounted rewards and computational times are shown in Table II. As expected, the robust strategy ends up following a highly conservative policy repairing the machine at state 49 to avoid state 50. The first distributionally robust strategy, not modeling the mixture Gaussian distribution well, finds it advantageous to repair at the 50th state. In contrast, capable of capturing the distribution information in a more flexible way, the second distributionally robust strategy better models the uncertainty and finds not repairing the machine at state 50 is optimal. The performance comparison clearly shows the second distributionally robust strategy is more desirable, which highlights the distributionally robust approach with general structure of confidence sets can be beneficial in practice.

We remark that, in practice, one can obtain the modality structure of uncertain parameters in a data-driven way by applying clustering algorithms to an initial primitive data set. For example, one may check the histogram of historical observations. If the data concentrates on several distinct and disjoint bins, our multi-model DAMDP approach can be applied. Moreover, we note that networked control systems (NCSs) have recently emerged as a topic of significant interest in the control community. A typical application of NCSs is in modern industrial systems, in which the components are often connected over network media. Our multi-model DAMDP approach might be extended for network-based performance tracking control of complex industrial processes, where recent work [25] and [26] proposed a novel two-layer structure to solve the setpoints compensation problem for industrial processes under network-based environment.

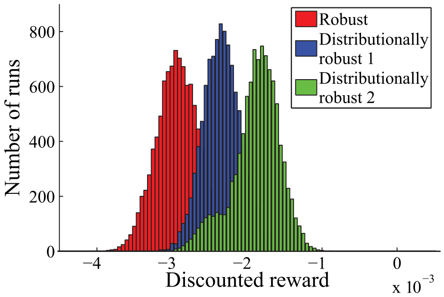


Fig. 5. Performance comparisons between robust and two distributionally robust strategies on 10,000 runs of the machine replacement problem with mixed Gaussian rewards.

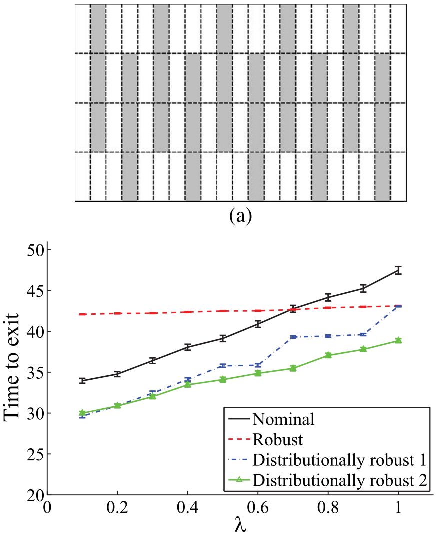
TABLE II
AVERAGE TOTAL DISCOUNTED REWARDS AND COMPUTATIONAL TIMES OF ROBUST AND TWO DISTRIBUTIONALLY ROBUST STRATEGIES IN MACHINE REPLACEMENT PROBLEM WITH MIXED GAUSSIAN REWARDS

Strategies	Robust	Distributionally robust 1	Distributionally robust 2
Average total discounted rewards	-2.9×10^{-3}	-2.3×10^{-3}	-1.9×10^{-3}
Computational times (seconds)	849	862	820

B. *Transition Uncertainty in the Path Planning Problem*

We now consider a path planning problem, similar to the one presented in [1]: an agent wants to exit a 4 × 21 maze [shown in Fig. 6(a)] using the least possible time. Starting from the upper-left corner, the agent can move up, down, left and right, but can only exit the grid at the lower-right corner. Here, a white box stands for a normal place where the agent needs one time unit to pass through. A shaded box represents a "shaky" place: if an agent reaches a "shaky" place, then he may risk jumping to the starting point ("reboot"). The true transition probability of the jump follows a distribution $(1-\lambda)\mathcal{N}(0.1, 10^{-4}) \oplus \lambda\mathcal{N}(0.2, 10^{-4})$ where $\lambda \in (0, 1]$. The four approaches are implemented as follows: The nominal approach neglects this random jump. The robust approach takes a worst-case analysis, i.e., it assumes that with 30%, the whole probability support of transition, the agent will jump to the spot with the highest cost-to-go. The first distributionally robust approach takes into account an additional information by using two *nested* confidence sets: the jump probability parameter belonging to 9%–11% is of a confidence $1-\lambda$. The second distributionally robust approach, which is proposed in this technical note, incorporates more information. In specific, we construct an extra confidence interval *disjoint* with the above 9%–11% interval. It states that the chance of jumping with probability 20% is λ.

The performance of strategies of the nominal, the robust and the two distributionally robust approaches is shown in Fig. 6(b), where the error bars show the standard error of the expected time to exit. The CPU times of computing optimal policies for four strategies are 0.461, 549, 642, and 654 seconds, respectively. The second distributionally robust approach achieves the best performance over virtually the whole spectrum of λ. This is well expected, since additional probabilistic



(a)

(b)

Fig. 6. Fig. 6(a) illustrates the maze for the path plawnning problem. Fig. 6(b) shows the performance comparisons between nominal, robust and two distributionally robust strategies over 3,000 runs of the path planning problem.

information is available to and incorporated by the second distributionally robust approach which considers ambiguity sets with more general structures.

V. Conclusion

In this technical note, we considered Markov decision problems with uncertainty. Specifically, we generalized the distributionally robust approach proposed in [1] to incorporate more general ambiguity sets proposed in [18] to model *a-priori* probabilistic information of the uncertain parameters. We proposed a way to compute the distributionally robust strategy through a Bellman type backward induction. We showed that the strategy, which achieves maximum expected utility under the worst admissible distributions of uncertain parameters, can be solved in polynomial time under some mild technical conditions. We believe that many important problems that are usually addressed using standard MDP models could be revisited and better resolved using the proposed models when parameter uncertainty exists, as this formulation naturally enables the decision maker to account for more general parameter uncertainty.

Acknowledgment

The authors wish to thank the associate editor and anonymous reviewers for their constructive comments which resulted in a significant improvement in this technical note.

References

[1] H. Xu and S. Mannor, "Distributionally robust Markov decision processes," *Math. Oper. Res.*, vol. 37, no. 2, pp. 288– 300, 2012.

[2] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: Wiley, 2014.

[3] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming (optimization and neural computation series, 3)," *Athena Scientific*, vol. 7, pp. 15– 23, 1996.

[4] A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[5] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis, "Bias and variance approximation in value function estimates," *Manag. Sci.*, vol. 53, no. 2, pp. 308– 322, 2007.

[6] A. L. Soyster, "Convex programming with set-inclusive constraints and applications to inexact linear programming," *Oper. Res.*, vol. 21, no. 5, pp. 1154– 1157, 1973.

[7] A. Ben-Tal and A. Nemirovski, "Robust solutions of uncertain linear programs," *Oper. Res. Lett.*, vol. 25, no. 1, pp. 1– 13, 1999.

[8] D. P. Bertsimas and M. Sim, "The price of robustness," *Oper. Res.*, vol. 52, no. 1, pp. 35– 53, 2004.

[9] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton, NJ: Princeton Univ. Press, 2009.

[10] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Oper. Res.*, vol. 53, no. 5, pp. 780–798, 2005.

[11] G. N. Iyengar, "Robust dynamic programming," *Math. Oper. Res.*, vol. 30, no. 2, pp. 257– 280, 2005.

[12] E. Delage and S. Mannor, "Percentile optimization for Markov decision processes with parameter uncertainty," *Oper. Res.*, vol. 58, no. 1, pp. 203–213, 2010.

[13] C. C. White and H. K. Eldeib, "Markov decision processes with imprecise transition probabilities," *Oper. Res.*, vol. 42, no. 4, pp. 739– 749, 1994.

[14] A. Bagnell, A. Y. Ng, and J. Schneider, "Solving uncertain Markov decision problems," Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-01-25, 2001.

[15] L. G. Epstein and M. Schneider, "Learning under ambiguity," *Rev. Econ. Studies*, vol. 74, no. 4, pp. 1275– 1303, 2007.

[16] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust Markov decision processes," *Math. Oper. Res.*, vol. 38, no. 1, pp. 153– 183, 2013.

[17] H. Xu, S. Mannor, B. Schölkopf, J. Platt, and T. Hofmann, "The robustness-performance tradeoff in Markov decision processes," in *Proc. NIPS*, 2006, pp. 1537– 1544.

[18] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Oper. Res.*, vol. 62, no. 6, pp. 1358– 1376, 2014.

[19] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Oper. Res.*, vol. 58, no. 3, pp. 595– 612, 2010.

[20] S. Mannor, O. Mebel, and H. Xu, "Lightning does not strike twice: Robust MDPs with coupled uncertainty," in *Proc. 29th Int. Conf. Machine Learning (ICML'12)*, Edinburgh, U.K., 2012, pp. 385–392. [Online]. Available: http://arxiv.org/abs/1206.4643.

[21] S. Zymler, D. Kuhn, and B. Rustem, "Distributionally robust joint chance constraints with second-order moment information," *Math. Programm.*, vol. 137, no. 1–2, pp. 167– 198, 2013.

[22] P. Yu and H. Xu, "Distributionally robust counterpart in Markov decision processes," CoRR, vol. abs/1501.07418, 2015. [Online]. Available: http://arxiv.org/abs/1501.07418.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[24] M. Grant and S. Boyd, CVX: Matlab Software for Disciplined Convex Programming, Version 2.1, Mar. 2014. [Online]. Available: http://cvxr.com/cvx.

[25] T. Wang, H. Gao, and J. Qiu, "A combined adaptive neural network and nonlinear model predictive control for multirate networked industrial process control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 2, pp. 416– 425, Feb. 2016.

[26] F. Liu, H. Gao, J. Qiu, S. Yin, J. Fan, and T. Chai, "Networked multirate output feedback control for setpoints compensation and its application to rougher flotation process," *IEEE Trans. Ind. Electron.*, vol. 61, no. 1, pp. 460– 468, 2014.