MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxxx 20xx, pp. xxx–xxx
ISSN 0364-765X | EISSN 1526-5471 |xx|0000|0xxx
DOI 10.1287/moor.xxxx.xxxx

Distributionally Robust Markov Decision Processes

Huan Xu
Department of Mechanical Engineering, National University of Singapore, Singapore
email: mpexuh@nus.edu.sg

Shie Mannor
Department of Electrical Engineering, Technion, Israel
email: shie@ee.technion.ac.il

We consider Markov decision processes where the values of the parameters are uncertain. This uncertainty is described by a sequence of nested sets (that is, each set contains the previous one), each of which corresponds to a probabilistic guarantee for a different confidence level. Consequently, a set of admissible probability distributions of the unknown parameters is specified. This formulation models the case where the decision maker is aware of and wants to exploit some (yet imprecise) a-priori information of the distribution of parameters, and arises naturally in practice where methods for estimating the confidence region of parameters abound. We propose a decision criterion based on distributional robustness: the optimal strategy maximizes the expected total reward under the most adversarial admissible probability distributions. We show that finding the optimal distributionally robust strategy can be reduced to the standard robust MDP where parameters are known to belong to a *single* uncertainty set, hence it can be computed in polynomial time under mild technical conditions.

Key words: Markov decision process, parameter uncertainty, distributional robustness.

MSC2000 Subject Classification: Primary: 90C40; Secondary: 90B50

OR/MS subject classification: Primary: Dynamic programming/optimal control, Markov, finite state; Secondary: Decision analysis, Applications

History: Received: March 16, 2010; Revised: April 14, 2011 and January 24, 2012.

1. Introduction. Sequential decision making in stochastic dynamic environments is often modeled using Markov decision processes (e.g., Puterman [27], Bertsekas and Tsitsiklis [7]). A strategy that achieves maximal expected accumulated reward is considered optimal. However, in many applications, the practical performance of such a strategy can significantly differ from the model's prediction due to *parameter uncertainty* – the deviation of the model parameters from the true ones (cf experiments in Mannor et al. [24]). Most attempts to reduce such performance variation consider the robust MDP formulation (e.g., Nilim and El Ghaoui [25], Bagnell et al. [4], White and El Deib[33], Iyengar [20], Epstein and Schneider[16]). In this context, it is assumed that the parameters can be any member of a known set (termed the *uncertainty set*), and solutions are ranked based on their performance under the (respective) worst parameter realizations. One main advantage of the robust MDP formulation is its computational efficiency: the optimal solution to a robust MDP is obtained in polynomial time when parameters are state-wise independent and the uncertainty set is compact and convex.

Because the robust approach considers the set-inclusive formulation of uncertainty, it is difficult to incorporate probabilistic information of the uncertainty, such as "r is no larger than 5 with at least 90% of probability." Such a-priori information is often available in applications, from either domain knowledge or sampling of parameter values. Neglecting this a-priori information, as robust MDP approach does, can lead to overly conservative solutions (cf Delage and Mannor [11]).

To effectively incorporates a-priori probabilistic information of the unknown parameters, distributionally robust formulation has been extensively studied and broadly applied in *single stage* optimization problems (e.g., Scarf [29], Dupacová [14], Kall [21], Shapiro [30], Popescu [26], Delage and Ye [12], Calafiore and El Ghaoui [10], and Goh and Sim [18]). In this framework, the uncertain parameters are regarded as stochastic, with a distribution μ that is not precisely observed, yet assumed to belong to an

a-priori known set $\mathcal{C}$. The objective of this problem is then formulated based on the worst-case analysis over distributions in $\mathcal{C}$. That is, 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 $\max_{x\in\mathcal{X}} \left[\inf_{\mu\in\mathcal{C}} \mathbb{E}_{\xi\sim\mu} u(x,\xi)\right]$.

From a decision theory perspective, the distributionally robust approach coincides with the celebrated MaxMin Expected Utility framework (Gilboa and Schemeidler [17] and Kelsey [23]), which states that if a preference relationship among actions satisfies certain axioms, then the optimal action maximizes the minimal expected utility with respect to a class of distributions. This approach addresses the famous *neglect of probability cognitive bias* (e.g., Baron [5]), i.e., the tendency to completely disregard probability when making a decision under uncertainty. Two extreme cases of such biases are the *normalcy bias*, which roughly speaking, can be stated as "since a disaster has never occurred then it never will occur," and the *zero-risk bias*, which stands for the tendency of individuals to prefer small benefits that are certain to large ones that are uncertain, regardless of the size of the certain benefit and the expected magnitude of the uncertain one. It is easy to see that the nominal approach and the robust approach suffer from normalcy bias and zero-risk bias, respectively. Additionally, the distributionally robust approach is also parallel to the minimax estimation used in statistical decision theory (Blackwell and Girshick [8]), which has been extensively explored through its connection with two-player zero-sum game (e.g., Karlin [22] and Dvoretzky *et al* [15]).

In this paper, we adapt the distributionally robust approach to multi-stage decision making and propose a new formulation for MDPs under parameter uncertainty. This distributionally robust formulation effectively incorporates a-priori probabilistic information of the uncertain parameters, and has a computational complexity comparable to robust MDP approach. Specifically, we consider a set of distributions that satisfy: (1) the parameters are state-wise independent; and (2) parameters of each state s is constrained by n different levels of (set) estimation, that is, for some $\mathcal{C}_s^1 \subseteq \mathcal{C}_s^2 \subseteq \cdots \mathcal{C}_s^n$, the probability that the parameters of state s belongs to $\mathcal{C}_s^i$ is at least λ_i (termed as the *nested-set condition* in the sequel). Strategies are then ranked based on their expected performance under the (respective) most adversarial distribution. Observe that the robust MDP formulation is a special case of the proposed distributionally robust formulation with $n = 1$.

The nested-set formulation is motivated by estimating the distributions of parameters via sampling. Such estimation is often imprecise especially when only a small number of samples is available. Instead, estimating uncertainty sets with high confidence can be made more accurate, which provides a lower-bound on the performance under the true distribution. In addition, one can easily sharpen the approximation by incorporating more layers of confidence sets (i.e, to increase n). Thus, the nested-set formulation provides a data-driven framework that can model a-priori information in a flexible way. Another setup that can be modeled using the nested-set condition is the "multi-scenario" setup, where in different scenarios the parameters are subject to different levels of uncertainty. For example, consider a "stable-shaky" model: imagine the system is the economy and the parameter value is the price of a stock, while the system usually (say with probability 90%) belongs to a "stable" scenario, corresponding to small fluctuation of the parameters from their nominal value, from time to time it can falls into a "shaky" scenario, which causes the parameters to be less predictable and subject to larger uncertainty. This can be modeled as a nested-set condition with two uncertainty sets, such that the parameter distribution with at least 90% belongs to the small uncertainty set, and guaranteed to belong to the large uncertainty set.

We remark that this paper considers the "large uncertainty case," that is, the parameter uncertainty is not diminishing. Readers interested in the diminishing uncertainty case may refer to literature of perturbed Markov decision processes (e.g., Delebecque and Quadrat [13], Abbad and Filar [1], and Abbad, Filar and Bielecki [2]; see Avrachenkov, Filar and Haviv [3] for a detailed survey) that deals with this problem in a comprehensive way.

This paper is organized as follows. In Section 2 we provide some background and assumptions on uncertain MDPs. We then formulate and solve the distributionally robust MDP for both finite-horizon and discounted reward infinite-horizon case, in Section 3 and Section 4, respectively. Some concluding remarks are offered in Section 5.

2. Preliminaries. Throughout the paper, we use capital letters to denote matrices, and bold face letters to denote column vectors. Row vectors are represented as the transpose of column vectors. We use $\mathbf{1}$ to denote the vectors of appropriate length with all elements 1, and use $\mathbf{e}_i(m)$ to denote the i^{th}

elementary vector of length m.

A (finite) Markov Decision Process (MDP) is defined as a 6-tuple $< T, \gamma, S, A_s, \mathbf{p}, \mathbf{r} >$ where: T is the possibly infinite decision horizon; $\gamma \in (0, 1]$ is the discount factor; S is the finite state set; A_s is the finite action set of state s; $\mathbf{p}$ is the transition probability; and $\mathbf{r}$ is the expected reward. 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 to reach state s'. Following Puterman [27], we denote the set of all history-dependent randomized strategies by Π^{HR}, and the set of all Markovian randomized strategies by Π^{MR}. We use subscript s to denote the value associated with state s, e.g., $\mathbf{r}_s$ denotes the vector form of rewards associated with state s, and π_s is the (randomized) action chosen at state s for strategy π. The elements in 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., $\mathbf{p} = \bigotimes_{s \in S} \mathbf{p}_s$.

An n-layer Uncertain MDP (*UMDP*) is defined as a tuple $< T, \gamma, S, A_s, (\mathcal{P}^1, \cdots, \mathcal{P}^n) >$ such that $\mathcal{P}^1 \subseteq \mathcal{P}^2 \subseteq \cdots \subseteq \mathcal{P}^n$. Each uncertainty set $\mathcal{P}^i$ provides an estimation of the unknown parameters (both rewards and transitions) subject to a different confidence level, which we will make precise in the next section. Note that when $n = 1$ the UDMP reduces to the standard robust MDP formulation where the only a-priori information of the unknown parameters is that they belong to the uncertainty set.

We make the following assumption about the uncertainty set, which basically means that the parameters of different states are independent (we use the term "independent" but there is no probabilistic interpretation here). Such assumption is made, to the best of our knowledge, by all papers investigating UMDPs to date (e.g. Nilim and El Ghaoui [25] and Iyengar [20]).

Assumption 2.1 *State-wise Cartesian uncertainty sets:*

(i) For $i = 1, \cdots, n$, $\mathcal{P}^i = \bigotimes_{s \in S} \mathcal{P}^i_s$.

(ii) For $i = 1, \cdots, n$, $\mathcal{P}^i_s$ *is nonempty, convex and compact.*

Note that Assumption 2.1 implies that statewise, the uncertainty sets have an incremental inclusive structure as well, that is, for any $s \in S$ and $i < j$, $\mathcal{P}^i_s \subseteq \mathcal{P}^j_s$.

Similarly to Nilim and El Ghaoui [25], we assume that when a state is visited multiple times, each time it can take a different parameter realization (*non-stationary model*). This assumption is justified mainly because the stationary model is generally intractable and a lower-bound on 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, for finite horizon case we can make the following assumption without loss of generality. This will simplify our derivations.

Assumption 2.2 *(i) Each state belongs to only one stage.*

(ii) The terminal reward equals zero.

(iii) The first stage only contains one state s^{ini}.

Using Assumption 2.2 (i), we partition S according to the stage each state belongs to. That is, we let S_t be the set of states belong to t^{th} stage. For a strategy π, we denote the expected (discounted) total-reward under parameters $\mathbf{p}$, $\mathbf{r}$ by $u(\pi, \mathbf{p}, \mathbf{r})$, that is,

$$u(\pi, \mathbf{p}, \mathbf{r}) \triangleq \mathbb{E}^{\mathbf{p}}_{\pi}\{\sum_{i=1}^{T} \gamma^{i-1} r(s_i, a_i)\}.$$

3. Distributionally robust MDPs: finite-horizon case. This section focuses on uncertain MDPs with a finite number of decision stages. We propose a decision criterion we term *distributionally robustness*, which incorporates a-prior information of how parameters are distributed. We show that a strategy defined through backward induction, which we called S-robust strategy, is a distributionally robust strategy. We further show that such strategy is solvable in polynomial time under mild technical conditions.

Recall that distributionally robust formulation (in general) solves the optimal solution evaluated with the expected utility for the most adversarial distribution. Thus, to adapt this framework into MDPs, we need to specify the set of admissible distributions of uncertain parameters $(\mathbf{p}, \mathbf{r})$. Let $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = 1$. We use the following set of distributions $\mathcal{C}_S(\lambda_1, \cdots, \lambda_n)$ for our model.

$$\begin{aligned} \mathcal{C}_S(\lambda_1, \cdots, \lambda_n) &\triangleq \{\mu | \mu = \bigotimes_{s\in S} \mu_s;\ \mu_s \in \mathcal{C}_s(\lambda_1, \cdots, \lambda_n), \forall s \in S\}, \\ \text{where: } \mathcal{C}_s(\lambda_1, \cdots, \lambda_n) &\triangleq \{\mu_s | \mu_s(\mathcal{P}_s^n) = 1;\ \mu_s(\mathcal{P}_s^i) \geq \lambda_i,\ i = 1, \cdots, n-1\}. \end{aligned} \tag{1}$$

We briefly explain this set of distributions. For a state s, the condition $\mu_s(\mathcal{P}_s^n) = 1$ means that the unknown parameters $(\mathbf{p}_s, \mathbf{r}_s)$ are restricted to the outermost uncertainty set; and the condition $\mu_s(\mathcal{P}_s^i) \geq \lambda_i$ means that with probability at least λ_i, $(\mathbf{p}_s, \mathbf{r}_s) \in \mathcal{P}_s^i$. Thus, $\mathcal{P}_s^1, \cdots, \mathcal{P}_s^n$ provides probabilistic guarantees of $(\mathbf{p}_s, \mathbf{r}_s)$ for n different uncertainty sets (or equivalently confidence levels). Note that $\bigotimes_{s\in S} \mu_s$ stands for the product measure generated by μ_s, which indicates that the parameters among different states are independent.

Since we fix $\lambda_1, \cdots, \lambda_n$ throughout the paper, to avoid heavy notations, in the sequel we simply write $\mathcal{C}_s$ and $\mathcal{C}_S$ without explicitly referring to λ. To simplify notations, we also let $\lambda_0 \equiv 0$.

Definition 3.1 *A strategy $\pi^* \in \pi^{HR}$ is* distributionally robust *with respect to $\mathcal{C}_S$ if it satisfies that for all $\pi \in \Pi^{HR}$,*

$$\inf_{\mu\in\mathcal{C}_S} \int u(\pi, \mathbf{p}, \mathbf{r})\, d\mu(\mathbf{p}, \mathbf{r}) \leq \inf_{\mu'\in\mathcal{C}_S} \int u(\pi^*, \mathbf{p}, \mathbf{r})\, d\mu'(\mathbf{p}, \mathbf{r}).$$

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 (if there exists one) is the optimal strategy according to this measure. We show that the following S-robust strategy defined through a backward induction is the distributionally robust strategy, by reducing the distributionally robust MDP into a standard robust MDP. Note that the definition essentially requires that the strategy must be robust with respect to each sub-problem, and hence the name "S-robust."

Definition 3.2 *Given UMDP $< T, \gamma, S, A_s, (\mathcal{P}^1, \cdots, \mathcal{P}^n) >$ with $T < \infty$, denote $\hat{\mathcal{P}}_s = \{\sum_{i=1}^n (\lambda_i - \lambda_{i-1})(\mathbf{r}_s(i), \mathbf{p}_s(i)) | (\mathbf{p}_s(i), \mathbf{r}_s(i)) \in \mathcal{P}_s^i\}$, we define the following:*

(*i*) *For $s \in S_T$, the* S-robust value $\tilde{v}_T(s) \triangleq 0$.

(*ii*) *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*

$$\tilde{v}_t(s) \triangleq \max_{\pi_\mathbf{s}\in\Delta(s)} \Big\{ \min_{(\mathbf{p}_s, \mathbf{r}_s)\in\hat{P}_s} \mathbb{E}_{\pi_s}^{\mathbf{p}_s}[r(s, a) + \gamma \tilde{v}_{t+1}(\underline{s})] \Big\}.$$

$$\tilde{\pi}_s \in \arg\max_{\pi_\mathbf{s}\in\Delta(s)} \Big\{ \min_{(\mathbf{p}_s, \mathbf{r}_s)\in\hat{P}_s} \mathbb{E}_{\pi_s}^{\mathbf{p}_s}[r(s, a) + \gamma \tilde{v}_{t+1}(\underline{s})] \Big\}.$$

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

Note that for each s, $\mathbb{E}_{\pi_s}^{\mathbf{p}_s}[r(s, a) + \gamma \tilde{v}_{t+1}(\underline{s})]$ is bilinear to π_s and $(\mathbf{p}_s, \mathbf{r}_s)$. Furthermore, $\hat{\mathcal{P}}_s$ is convex and compact. Thus, by a standard game theoretic argument, max and min in Definition 3.2 are attainable, which further implies the existence of S-robust action (and hence S-robust strategy). Indeed, readers familiar with literature in robust MDP (cf Nilim and El Ghaoui [25], White and El Deib[33], Iyengar [20]) may find that S-robust strategy is the solution to the robust MDP where the uncertainty set is $\bigotimes_s \hat{\mathcal{P}}_s$. The following theorem shows that any S-robust strategy π^* is distributionally robust.

Theorem 3.1 *Let $T < \infty$. Under Assumptions 2.1 and 2.2, if π^* is an S-robust strategy, then*

(*i*) *π^* is a distributionally robust strategy with respect to $\mathcal{C}_s$;*

(*ii*) *there exists $\mu^* \in \mathcal{C}_s$ such that (π^*, μ^*) is a saddle point. That is,*

$$\sup_{\pi\in\Pi^{HR}} \int u(\pi, \mathbf{p}, \mathbf{r})\, d\mu^*(\mathbf{p}, \mathbf{r}) = \int u(\pi^*, \mathbf{p}, \mathbf{r})\, d\mu^*(\mathbf{p}, \mathbf{r}) = \inf_{\mu\in\mathcal{C}_S} \int u(\pi^*, \mathbf{p}, \mathbf{r})\, d\mu(\mathbf{p}, \mathbf{r}).$$

PROOF. The outline of the proof is as follows: We first show that for a given strategy, the expected performance under admissible μ depends only on the expected value of the parameters. Then we show that the set of expected parameters is indeed $\bigotimes_{s\in S}\hat{\mathcal{P}}_s$. Thus the distributionally robust MDP reduces to the robust MDP with $\bigotimes_{s\in S}\hat{\mathcal{P}}_s$ being the uncertainty set. Finally, by applying results from robust MDP we prove the theorem.

Let h_t denote a history up to stage t and $s(h_t)$ denote the last state of history h_t. We use $\pi_{h_t}(a)$ to represent the probability of choosing an action a at state $s(h_t)$, following a strategy π and under a history h_t. A $t+1$ stage history, with h_t followed by action a and state s' is written as (h_t,a,s').

With an abuse of notation, we denote the expected reward-to-go under a history as:

$$u(\pi,\mathbf{p},\mathbf{r},h_t)\triangleq\mathbb{E}_\pi^{\mathbf{p}}\{\sum_{i=t}^{T}\gamma^{i-t}r(s_i,a_i)|(s_1,a_1\cdots,s_t)=h_t\}.^1$$

For $\pi\in\Pi^{HR}$ and $\mu\in\mathcal{C}_S(\lambda)$, define

$$w(\pi,\mu,h_t)\triangleq\mathbb{E}_{(\mathbf{p},\mathbf{r})\sim\mu}u_s(\pi,\mathbf{p},\mathbf{r},h(t))=\int u(\pi,\mathbf{p},\mathbf{r},h(t))d\mu(\mathbf{p},\mathbf{r}).$$

Thus, $w(\pi,\mu,(s^{\text{ini}}))=\int u(\pi,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r})$ is the minimax objective. We establish the following recursion formula for $w(\cdot)$.

LEMMA 3.1 *Fix $\pi\in\Pi^{HR}$, $\mu\in\mathcal{C}_S$ and a history h_t where $t<T$, denote $\overline{\mathbf{r}}=\mathbb{E}_\mu(\mathbf{r})$, $\overline{\mathbf{p}}=\mathbb{E}_\mu(\mathbf{p})$, then we have:*

$$\begin{aligned}w(\pi,\mu,h_t)&=\int\sum_{a\in A_{s(h_t)}}\pi_{h_t}(a)\Big(r\big(s(h_t),a\big)+\sum_{s'\in S}\gamma p\big(s'|s(h_t),a\big)w\big(\pi,\mu,(h_t,a,s')\big)\Big)d\mu_{s(h_t)}(\mathbf{p}_{s(h_t)},\mathbf{r}_{s(h_t)})\\&=\sum_{a\in A_{s(h_t)}}\pi_{h_t}(a)\Big(\overline{r}\big(s(h_t),a\big)+\sum_{s'\in S}\gamma\overline{p}\big(s'|s(h_t),a\big)w\big(\pi,\mu,(h_t,a,s')\big)\Big).\end{aligned}\tag{2}$$

PROOF. By definition $\mu(\mathbf{p},\mathbf{r})\in\mathcal{C}_S$ implies $\mu(\mathbf{p},\mathbf{r})=\bigotimes_{s\in S}\mu_s(\mathbf{p}_s,\mathbf{r}_s)$ while $\mu_s(\mathbf{p}_s,\mathbf{r}_s)\in\mathcal{C}_s$. When it is clear on what variable the distribution is, we will simply write μ and μ_s. Denote $\mu(t)=\bigotimes_{s\in\bigcup_{i=t}^{T}S_i}\mu_s$, that is, the probability distribution for the parameters from stage t on. We thus have:

$$\begin{aligned}w(\pi,\mu,h_t)&=\mathbb{E}_{(\mathbf{p},\mathbf{r})\sim\mu}u(\pi,\mathbf{p},\mathbf{r},h_t)=\int u(\pi,\mathbf{p},\mathbf{r},h_t)d\mu\\&=\int u(\pi,\mathbf{p},\mathbf{r},h_t)d\mu(t)=\int\int u(\pi,\mathbf{p},\mathbf{r},h_t)d\mu(t+1)d\mu_{s(h_t)},\end{aligned}\tag{3}$$

due to the fact that $u\big(\pi,\mathbf{p},\mathbf{r},h(t)\big)$ only depends on the parameters from the t^{th} stage on. Notice that for a fixed parameter realization and a fixed strategy, the Bellman Equation holds. That is,

$$u(\pi,\mathbf{p},\mathbf{r},h_t)=\sum_{a\in A_{s(h_t)}}\pi_{h_t}(a)\Big(r\big(s(h_t),a\big)+\sum_{s'\in S_{t+1}}\gamma p\big(s'|s(h_t),a\big)u\big(\pi,\mathbf{p},\mathbf{r},(h_t,a,s')\big)\Big).$$

The right-hand-side of Equation (3) therefore equals

$$\begin{aligned}&\int\int\Big\{\sum_{a\in A_{s(h_t)}}\pi_{h_t}(a)\Big(r\big(s(h_t),a\big)+\sum_{s'\in S_{t+1}}\gamma p\big(s'|s(h_t),a\big)u\big(\pi,\mathbf{p},\mathbf{r},(h_t,a,s')\big)\Big)\Big\}d\mu(t+1)d\mu_{s(h_t)}\\\overset{(a)}{=}&\int\sum_{a\in A_{s(h_t)}}\pi_{h_t}(a)\Big(r\big(s(h_t),a\big)+\sum_{s'\in S_{t+1}}\gamma p\big(s'|s(h_t),a\big)\int u\big(\pi,\mathbf{p},\mathbf{r},(h_t,a,s')\big)d\mu(t+1)\Big)d\mu_{s(h_t)}\\=&\int\sum_{a\in A_{s(h_t)}}\pi_{h_t}(a)\Big(r\big(s(h_t),a\big)+\sum_{s'\in S}\gamma p\big(s'|s(h_t),a\big)w\big(\pi,\mu,(h_t,a,s')\big)\Big)d\mu_{s(h_t)},\end{aligned}$$

[1] When h_t is not possible under $\mathbf{p}$ and π, $u(\pi,\mathbf{p},\mathbf{r},h_t)$ is the expected accumulated reward of a fictitious MDP, which is the sub-problem starting from s_t of the original MDP, and the strategy and the parameters are π, $\mathbf{p}$ and $\mathbf{r}$ conditioned on h_t.

which proves the first equality of (2). Here (a) holds because $\mu(t+1)$ by definition is the probability distribution of parameters from stage $t+1$ on, and $s(h_t)$ belongs to stage t.

Since $\mu(\mathbf{p},\mathbf{r}) = \bigotimes_{s\in S}\mu_s(\mathbf{p}_s,\mathbf{r}_s)$, we have $\overline{r}(s(h_t),a) = \int r(s(h_t),a)d\mu_{s(h_t)}$ and $\overline{p}\big(s'|s(h_t),a\big) = \int p\big(s'|s(h_t),a\big)d\mu_{s(h_t)}$. The second equality of (2) follows. □

LEMMA 3.2 *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:*

$$w\big(\pi,\mu,(s^{\text{ini}})\big) = u(\pi,\overline{\mathbf{p}},\overline{\mathbf{r}}).$$

PROOF. Note that Bellman Equation gives

$$u(\pi,\overline{\mathbf{p}},\overline{\mathbf{r}},h_t) = \sum_{a\in A_{s(h_t)}}\pi_{h_t}(a)\Big(\overline{r}\big(s(h_t),a\big)+\sum_{s'\in S}\gamma\overline{p}\big(s'|s(h_t),a\big)u\big(\pi,\overline{\mathbf{p}},\overline{\mathbf{r}},(h_t,a,s')\big)\Big).$$

Furthermore for any h_T

$$w\big(\pi,\mu,h_T\big) = 0 = u\big(\pi,\overline{\mathbf{p}},\overline{\mathbf{r}},h_T\big).$$

Thus, Lemma 3.1 and backward induction on t prove the lemma. □

Lemma 3.2 essentially means that for any strategy, the expected performance under an admissible distribution μ only depends on the expected value of the parameters under μ. Thus, the distributionally robust MDP reduces to the robust MDP. Next we characterize the set of expected value of the parameters.

LEMMA 3.3 *Fix $s\in S$, we have*

$$\{\mathbb{E}_{\mu_s}(\mathbf{p}_s,\mathbf{r}_s)|\mu_s\in\mathcal{C}_s\} = \hat{\mathcal{P}}_s.$$

PROOF. To simplify notation, let $c_i=\lambda_i-\lambda_{i-1}$, and use $\mathbf{x}$ to represent the parameter vector $(\mathbf{p}_s,\mathbf{r}_s)$. Thus, we need to show that $\{\mathbb{E}_{\mu_s}\mathbf{x}|\mu_s\in\mathcal{C}_s\} = \{\sum_{i=1}^n(\lambda_i-\lambda_{i-1})\mathbf{x}_i|\mathbf{x}_i\in\mathcal{P}_s^i\}$.

Note that given any $(\mathbf{x}_1,\cdots,\mathbf{x}_n)$ such that $\mathbf{x}_i\in\mathcal{P}_s^i$, we let $\mu_s(\mathbf{x}_i) = \sum_{j:\mathbf{x}_i=\mathbf{x}_j}c_j$. Observe that μ_s belongs to $\mathcal{C}_s$, and satisfies that $\mathbb{E}_{\mathbf{x}\sim\mu_s}(\mathbf{x}) = \sum_{i=1}^n c_i\mathbf{x}_i$. Therefore, $\{\mathbb{E}_{\mathbf{x}\sim\mu_s}(\mathbf{x})|\mu_s\in\mathcal{C}_s\}\supseteq\hat{\mathcal{P}}_s$.

We next show that $\{\mathbb{E}_{\mathbf{x}\sim\mu_s}(\mathbf{x})|\mu_s\in\mathcal{C}_s\}\subseteq\hat{\mathcal{P}}_s$; or equivalently, for any $\mu_s\in\mathcal{C}_s$, $\mathbb{E}_{\mathbf{x}\sim\mu_s}\mathbf{x}\in\hat{\mathcal{P}}_s$. To this end, we show that for any $\mathbf{y}$,

$$\mathbb{E}_{\mathbf{x}\sim\mu_s}\mathbf{y}^\top\mathbf{x}\le\max_{\mathbf{x}'\in\hat{\mathcal{P}}_s}\mathbf{y}^\top\mathbf{x}'.$$

Let $\alpha_i\triangleq\mu_s(\mathcal{P}_s^i\backslash\mathcal{P}_s^{i-1})$, by total expectation we have

$$\mathbb{E}_{\mathbf{x}\sim\mu_s}\mathbf{y}^\top\mathbf{x}\le\sum_{i=1}^n\alpha_i\max_{\mathbf{x}_i\in(\mathcal{P}_s^i\backslash\mathcal{P}_s^{i-1})}\mathbf{y}^\top\mathbf{x}_i\le\sum_{i=1}^n\alpha_i\max_{\mathbf{x}_i\in\mathcal{P}_s^i}\mathbf{y}^\top\mathbf{x}_i. \tag{4}$$

Furthermore, note that $\sum_{i=1}^k\alpha_i=\mu_s(\mathcal{P}_s^k)\ge\lambda_k=\sum_{i=1}^k c_i$, $\sum_{i=1}^n\alpha_i=1=\sum_{i=1}^n c_i$, and $\max_{\mathbf{x}_i\in\mathcal{P}_s^i}\mathbf{y}^\top\mathbf{x}_i$ is non-decreasing in i. Thus we have

$$\sum_{i=1}^n\alpha_i\max_{\mathbf{x}_i\in\mathcal{P}_s^i}\mathbf{y}^\top\mathbf{x}_i\le\sum_{i=1}^n c_i\max_{\mathbf{x}_i'\in\mathcal{P}_s^i}\mathbf{y}^\top\mathbf{x}_i' = \max_{\mathbf{x}'\in\hat{\mathcal{P}}_s}\mathbf{y}^\top\mathbf{x}'.$$

Combining this with Inequality (4), we have that for any $\mathbf{y}$,

$$\mathbb{E}_{\mathbf{x}\sim\mu_s}\mathbf{y}^\top\mathbf{x}\le\max_{\mathbf{x}'\in\hat{\mathcal{P}}_s}\mathbf{y}^\top\mathbf{x}'.$$

This implies $\mathbb{E}_{\mathbf{x}\sim\mu_s}\mathbf{x}\in\hat{\mathcal{P}}_s$, because otherwise by the strict separating hyperplane theorem (recall that $\hat{\mathcal{P}}_s$ is convex and compact), there exists some $\mathbf{y}$ such that the inequality does not hold. □

Note that Lemma 3.3 implies that $\{\mathbb{E}_\mu(\mathbf{p},\mathbf{r})|\mu\in\mathcal{C}_S\} = \bigotimes_{s\in S}\hat{\mathcal{P}}_s$, by the statewise decomposibility of $\mathcal{C}_S$.

We complete the proof of the Theorem 3.1 using the equivalence of distributionally robust MDsP and robust MDPs where the uncertainty set is $\bigotimes_{s\in S}\hat{\mathcal{P}}_s$. Recall that for each $s\in S$, $\hat{\mathcal{P}}_s$ is convex and compact. It is well known that for robust MDPs, a saddle point of the minimax objective exists (Nilim and El Ghaoui [25], Iyengar [20]). More precisely, there exists $\pi^*\in\Pi^{HR},(\mathbf{p}^*,\mathbf{r}^*)\in\bigotimes_{s\in S}\hat{\mathcal{P}}_s$ such that

$$\sup_{\pi\in\Pi^{HR}}u(\pi,\mathbf{p}^*,\mathbf{r}^*) = u(\pi^*,\mathbf{p}^*,\mathbf{r}^*) = \inf_{(\mathbf{r},\mathbf{p})\in\bigotimes_{s\in S}\hat{\mathcal{P}}_s}u(\pi^*,\mathbf{p},\mathbf{r}).$$

Moreover, π^* and $(\mathbf{p}^*, \mathbf{r}^*)$ can be constructed state-wise: $\pi^* = \bigotimes_{s\in S} \pi_s^*$ and $(\mathbf{p}^*, \mathbf{r}^*) = \bigotimes_{s\in S}(\mathbf{p}_s^*, \mathbf{r}_s^*)$, and for each $s \in S_t$, $\pi_s^*, (\mathbf{p}_s^*, \mathbf{r}_s^*)$ solves the following zero-sum game

$$\max_{\pi_s} \min_{(\mathbf{p}_s,\mathbf{r}_s)\in\hat{P}_s} \mathbb{E}_\pi^{\mathbf{p}_s}\big(r(s,a) + \gamma\tilde{v}_{t+1}(\underline{s})\big).$$

Thus π_s^* is any S-robust action, and hence π^* can be any S-robust strategy. From Lemma 3.3, there exists $\mu_s^* \in \mathcal{C}_s$ that satisfies $\mathbb{E}_{\mu_s^*}(\mathbf{p}_s, \mathbf{r}_s) = (\mathbf{p}_s^*, \mathbf{r}_s^*)$. Let $\mu^* = \bigotimes_{s\in S} \mu_s^*$. By Lemma 3.2 we have

$$\begin{aligned}
&\sup_{\pi\in\Pi^{HR}} w\big(\pi, \mu^*, (s^{\text{ini}})\big) = \sup_{\pi\in\Pi^{HR}} u(\pi, \mathbf{p}^*, \mathbf{r}^*);\\
&w\big(\pi^*, \mu^*, (s^{\text{ini}})\big) = u\big(\pi^*, \mathbf{p}^*, \mathbf{r}^*\big);\\
&\inf_{\mu\in\mathcal{C}_S} w\big(\pi^*, \mu, (s^{\text{ini}})\big) = \inf_{(\mathbf{p},\mathbf{r})\in\bigotimes_s \hat{\mathcal{P}}_s} u(\pi^*, \mathbf{p}, \mathbf{r}).
\end{aligned}$$

This leads to

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

Thus, part (ii) of Theorem 3.1 holds. Note that part (ii) immediately implies part (i) of Theorem 3.1. □

We briefly discuss the relationship between finding a distributionally robust strategy and solving a zero-sum two player stochastic game (e.g., Sion [32], Karlin [22], Shapley [31]). Indeed, finding a distributionally robust strategy can be modeled as a zero-sum stochastic game with perfect information, where the first player (the decision maker) chooses π, and the second player (Nature) chooses μ – the distribution over uncountably many uncertain parameters $\mathbf{p}$ and $\mathbf{r}$. In standard robust MDPs, where one player chooses π and the other chooses $\mathbf{p}$ and $\mathbf{r}$, a stochastic game theoretic argument has been used to establish mini-max results that are parallel to Theorem 3.1 (e.g., Section 5 of Iyengar [20]), thanks to the fact that π, $\mathbf{p}$ and $\mathbf{r}$ can all be embedded into a finite dimensional Euclidean space. Such an argument may not directly generalize to the distributionally robust case because μ cannot be trivially embedded into a finite dimensional space. This may raise issues concerning the compactness of the action set of the second player and measurability of the objective function. While it may be possible to apply more advanced game-theoretic argument to the distributionally robust case, a refined analysis on the topology of the set of μ is necessary. The constructive proof we have provided seems to be more accessible.

We now investigate the computational aspect of the S-robust action. By backward induction we thus find the S-robust strategy.

THEOREM 3.2 *Denote $\lambda_0 = 0$. For $s \in S_t$ where $t < T$, the S-robust action is given by*

$$\mathbf{q}^* = \arg\max_{\mathbf{q}\in\Delta(s)} \Big\{ \sum_{i=1}^{n} (\lambda_i - \lambda_{i-1}) \min_{(\mathbf{p}_s^i,\mathbf{r}_s^i)\in\mathcal{P}_s^i} \big[(\mathbf{r}_s^i)^\top \mathbf{q} + (\mathbf{p}_s^i)^\top \tilde{V}_s \mathbf{q}\big] \Big\}, \tag{5}$$

where $m = |A_s|$, $\tilde{\mathbf{v}}_{t+1}$ is the vector form of $\tilde{v}_{t+1}(s')$ for all $s' \in S_{t+1}$, and

$$\tilde{V}_s \triangleq \begin{bmatrix} \tilde{\mathbf{v}}_{t+1}\mathbf{e}_1^\top(m) \\ \vdots \\ \tilde{\mathbf{v}}_{t+1}\mathbf{e}_m^\top(m) \end{bmatrix}.$$

PROOF. Notice that for any $\mathbf{q} \in \Delta(s)$, the following holds:

$$\begin{aligned}
&\min_{(\mathbf{p}_s,\mathbf{r}_s)\in\hat{\mathcal{P}}_s} \mathbb{E}_{\mathbf{q}}^{\mathbf{p}_s}[r(s,a) + \tilde{v}_{t+1}(\underline{s})]\\
&= \min_{(\mathbf{p}_s,\mathbf{r}_s)=\sum_{i=1}^n(\lambda_i-\lambda_{i-1})(\mathbf{p}_s^i,\mathbf{r}_s^i);\,\forall i:\,(\mathbf{p}_s^i,\mathbf{r}_s^i)\in\mathcal{P}_s^i} \Big[\sum_{a\in A_s} q(a) r(s,a) + \sum_{a\in A_s}\sum_{s'\in S_{t+1}} q(a)p(s'|s,a)\tilde{v}_{t+1}(s') \Big]\\
&= \sum_{i=1}^{n} (\lambda_i - \lambda_{i-1}) \min_{(\mathbf{p}_s^i,\mathbf{r}_s^i)\in\mathcal{P}_s^i} \Big[\sum_{a\in A_s} q(a) r^i(s,a) + \sum_{a\in A_s}\sum_{s'\in S_{t+1}} q(a)p^i(s'|s,a)\tilde{v}_{t+1}(s') \Big]\\
&= \sum_{i=1}^{n} (\lambda_i - \lambda_{i-1}) \min_{(\mathbf{p}_s^i,\mathbf{r}_s^i)\in\mathcal{P}_s^i} \big[(\mathbf{r}_s^i)^\top \mathbf{q} + (\mathbf{p}_s^i)^\top \tilde{V}_s \mathbf{q}\big].
\end{aligned} \tag{6}$$

Taking maximization over $\mathbf{q} \in \Delta(s)$ on both side of Equation (6), we establish the theorem. □

Theorem 3.2 implies that the computation of the S-robust action at a state s critically depends on the structure of the sets $\mathcal{P}_s^i$. In fact, it can be shown that for "good" uncertainty sets, computing the S-robust action is tractable. To make this claim precise, we need the following definition.

DEFINITION 3.3 *A* polynomial separation oracle *of a convex set $\mathcal{H} \subseteq \mathbb{R}^n$ is a subroutine such that given $\mathbf{x} \in \mathbb{R}^n$, in polynomial time it reports whether $\mathbf{x} \in \mathcal{H}$, and if the answer is negative, it finds a hyperplane that separates $\mathbf{x}$ and $\mathcal{H}$.*

COROLLARY 3.1 *Under Assumption 2.1, The S-robust action for state s can be found in polynomial-time, if $\mathcal{P}_s^i$ has a polynomial separation oracle for each $i = 1, \cdots, n$.*

PROOF. We establish the following lemma first.

LEMMA 3.4 *Fix $c \in \mathbb{R}$, and let $\mathcal{F} \triangleq \{\mathbf{x} | g(\mathbf{x}) \leq c\}$ for a convex function $g(\cdot)$. If $\mathbf{x}_0 \notin \mathcal{F}$, and let $\mathbf{d}$ be a subgradient of $g(\cdot)$ evaluated at $\mathbf{x}_0$, then the following hyperplane*

$$\mathbf{d}^\top \mathbf{x} = \mathbf{d}^\top \mathbf{x}_0 + \frac{c - f(\mathbf{x}_0)}{2},$$

separates $\mathcal{F}$ and $\{\mathbf{x}_0\}$.

PROOF. By the definition of subgradient we have $f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{d}^\top(\mathbf{x} - \mathbf{x}_0)$, equivalently

$$\mathbf{d}^\top \mathbf{x} - f(\mathbf{x}) \leq \mathbf{d}^\top \mathbf{x}_0 - f(\mathbf{x}_0).$$

For any $\mathbf{x} \in \mathcal{F}$ we then have

$$\mathbf{d}^\top \mathbf{x} - c \leq \mathbf{d}^\top \mathbf{x} - f(\mathbf{x}) \leq \mathbf{d}^\top \mathbf{x}_0 - f(\mathbf{x}_0) < \mathbf{d}^\top \mathbf{x}_0 - \frac{c + f(\mathbf{x}_0)}{2}.$$

Here the first inequality holds due to $f(x) \leq c$, and the last holds due to $f(\mathbf{x}_0) > c$. This implies that if $\mathbf{x} \in \mathcal{F}$, then $\mathbf{d}^\top \mathbf{x} < \mathbf{d}^\top \mathbf{x}_0 + \frac{c - f(\mathbf{x}_0)}{2}$. On the other hand, $f(\mathbf{x}_0) > c$ leads to $\mathbf{d}^\top \mathbf{x}_0 > \mathbf{d}^\top \mathbf{x}_0 + \frac{c - f(\mathbf{x}_0)}{2}$. Thus, the hyperplane separates $\mathcal{F}$ and $\mathbf{x}_0$. □

We now turn to prove the corollary. It suffices to show that in polynomial time, the following optimization problem can be solved.

$$\begin{aligned} \text{Minimize:}_{\mathbf{q}} \ & \sum_{i=1}^{n} (\lambda_i - \lambda_{i-1}) \max_{(\mathbf{p}_s^i, \mathbf{r}_s^i) \in \mathcal{P}_s^i} \big[-(\mathbf{r}_s^i)^\top \mathbf{q} - (\mathbf{p}_s^i)^\top \tilde{V}_s \mathbf{q} \big] \\ \text{such that: } & \mathbf{q} \in \Delta(s). \end{aligned} \tag{7}$$

Notice that the objective function to be minimized is the maximum of a class of linear functions of $\mathbf{q}$, and hence convex. Thus, it suffices to show that for any $c \in \mathbb{R}$, in polynomial time we can either report a member of the following set or determine it is empty:

$$\{\mathbf{q} | \mathbf{q} \in \Delta(s), \ \sum_{i=1}^{n} (\lambda_i - \lambda_{i-1}) \max_{(\mathbf{p}_s^i, \mathbf{r}_s^i) \in \mathcal{P}_s^i} \big[-(\mathbf{r}_s^i)^\top \mathbf{q} - (\mathbf{p}_s^i)^\top \tilde{V}_s \mathbf{q} \big] \leq c\}.$$

A sufficient condition for this to hold is that we can construct a separation oracle for this set in polynomial time (Grötschel et al. [19]), which by Lemma 3.4 leads to checking the value and the sub-gradient of the function. Therefore, if the value and the sub-gradient of the objective function of (7) can be evaluated in polynomial time, the optimization problem (7) is solvable in polynomial time.

Due to the Envelope Theorem (e.g., Rockafellar [28]), it is known that for a function $f(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{C}} g(\mathbf{x}, \mathbf{y})$ where $g(\cdot, \cdot)$ is convex with respect to the first argument, the following holds (here we abuse the notation and use ∂ and $\partial_{\mathbf{x}}$ to represent the set of all sub-gradients),

$$\partial f(\mathbf{x}_0) \supseteq \partial_{\mathbf{x}} g(\mathbf{x}_0, \mathbf{y}^*), \quad \text{where: } \mathbf{y}^* = \arg\max_{\mathbf{y} \in \mathcal{C}} g(\mathbf{x}_0, \mathbf{y}).$$

Notice that for fixed $(\mathbf{p}_s^1, \mathbf{r}_s^1, \cdots, \mathbf{p}_s^n, \mathbf{r}_s^n)$, the objective function is linear. Hence, evaluation of the gradient is easy. Thus, we only need to show that in polynomial time, for any given $\mathbf{q}_0$, we can solve

$$\arg \max_{(\mathbf{p}_s^i, \mathbf{p}_s^i) \in \mathcal{P}_s^i} \big[-(\mathbf{r}_s^i)^\top \mathbf{q}_0 - (\mathbf{p}_s^i)^\top \tilde{V}_s \mathbf{q}_0 \big]; \quad i = 1, \cdots, n.$$

Note that these problems are maximizing a linear objective over a set. A sufficient condition for polynomial-time solvability of such problem is that the set is convex and compact, which holds by Assumption 2.1, and has a polynomial separation oracle (Grötschel et al. [19]). □

Having a polynomial separation oracle is a rather mild technical condition. Indeed, any convex set defined by finitely many convex constraints $g_i(\mathbf{x}) \leq 0$ has a polynomial (with respect to the number of constraints) separation oracle if both the value and the subgradient of $g_i(\cdot)$ can be evaluated in polynomial time (e.g., Ben-Tal and Nemirovski [6], Grötschel et al. [19]).

In practice, especially when the problem size is large, the theoretical guarantee of polynomial-time solvability may not ensure that the problem can be solved in reasonably short time. However, thanks to the duality theorem of convex programming (e.g., Boyd and Vandenberghe [9]), for many "nice" uncertainty sets, finding the S-robust action can be reduced to an "easy" optimization problem. For example, when the uncertainty sets are polytopes, which includes the arguably most natural uncertainty set – each parameter belongs to an interval, finding the S-robust action can be formulated as a linear program. Similarly, when uncertainty sets are ellipsoids, or even intersection of ellipsoids and polytopes, finding the S-robust action is a second order cone program. While these results are standard, for completeness we list them below.

Example 3.1 (Intersection of Polytopes and Ellipsoids) *The following minimization problem*

$$\begin{aligned} \textit{Minimize:}_{\mathbf{x},\mathbf{u}} \quad & \mathbf{h}^\top \mathbf{x} \\ \textit{Subject to:} \quad & \mathbf{x} = A\mathbf{u} + b \\ & \|\mathbf{u}\|_2 \leq 1 \\ & \mathbf{C}\mathbf{x} \geq \mathbf{d}; \end{aligned}$$

is equivalent to

$$\begin{aligned} \textit{Maximize:}_{\mathbf{y},\mathbf{z}} \quad & -\|A^\top \mathbf{y}\|_2 - \mathbf{b}^\top \mathbf{y} + \mathbf{d}^\top \mathbf{z} \\ \textit{Subject to:} \quad & -\mathbf{y} + C^\top \mathbf{z} = \mathbf{h} \\ & \mathbf{z} \geq 0. \end{aligned}$$

Example 3.2 (Polytopes) *If $\mathcal{P}_s^i$, are all polyhedral sets, defined as $\mathcal{P}_s^i = \{(\mathbf{p}_s^i, \mathbf{r}_s^i) | C^i \mathbf{p}_s^i + D^i \mathbf{r}_s^i \geq \mathbf{k}^i\}$, then the S-robust action equals the optimal $\mathbf{q}$ of the following Linear Program on $(\mathbf{q}, \mathbf{z}^1, \cdots, \mathbf{z}^n)$. In addition, the S-robust value equals its optimal value.*

$$\begin{aligned} \textit{Maximize:} \quad & \sum_{i=1}^{n} (\lambda_i - \lambda_{i-1})(\mathbf{k}^i)^\top \mathbf{z}^i \qquad (8) \\ \textit{Subject to:} \quad & (C^i)^\top \mathbf{z}^i = \mathbf{q}; \quad i = 1, \cdots, n \\ & (D^i)^\top \mathbf{z}^i = \tilde{V}_s \mathbf{q}; \quad i = 1, \cdots, n \\ & \mathbf{z}^i \geq 0; \quad i = 1, \cdots, n \\ & \mathbf{1}^\top \mathbf{q} = 1; \\ & \mathbf{q} \geq 0. \end{aligned}$$

Before concluding this section, let us remark that it is possible that no deterministic action is S-robust (and hence the distributionally robust strategy is not deterministic), because we allow parameters corresponding to different actions of a state to be coupled. In fact even for special case of robust MDP (i.e., distributionally robust with $n = 1$) this also holds. To see this, consider a simple example, with one state and two actions a and b. Suppose the reward parameters satisfy $r_a \in [0, 2]$, $r_b \in [0, 2]$ as well as $r_a + r_b = 2$. Then, it is clear that any deterministic strategy has a worst-case reward 0, whereas a randomized strategy that takes each action with probability 50% achieves a worst-case reward 1. Thus, the unique robust strategy is random.

4. Distributionally robust MDP: discounted reward infinite horizon case. In this section, we generalize the notion of S-robust strategy proposed in Section 3 to discounted-reward infinite-horizon UMDPs, and show that it is distributionally robust. Unlike the finite horizon case, we cannot model the

system as (1) having finitely many states, and (2) each visited at most once. In contrast, we have to relax either one of these two assumptions, which leads to two different natural formulations. The first formulation, termed *non-stationary model*, is to treat the system as having infinitely many states, each visited at most once. Therefore, we consider an equivalent MDP with an augmented state space, where each augmented state is defined by a pair (s,t) where $s \in S$ and t, meaning state s in the t^{th} horizon. Observe that each augmented state will be visited at most once. Similarly to the finite horizon case, the set of distributions to be considered is the Cartesian product of the admissible distribution of each (augmented) state. That is,

$$\bar{\mathcal{C}}_S^{\infty} \triangleq \{\mu | \mu = \bigotimes_{s\in S, t=1,2,\cdots} \mu_{s,t};\ \mu_{s,t} \in C_s, \forall s \in S,\ \forall t = 1,2,\cdots\}.$$

The second formulation, termed *stationary model*, treats the system as having a finite number of states, while multiple visits to one state is allowed. That is, if a state s is visited for multiple times, then each time the distribution (of uncertain parameters) μ_s is the same. Mathematically, we can adapt the augmented state space as in the non-stationary model, and requires that $\mu_{s,t}$ does not depend on t. Thus, the set of admissible distributions is

$$\bar{\mathcal{C}}_S \triangleq \{\mu | \mu = \bigotimes_{s\in S, t=1,2,\cdots} \mu_{s,t};\ \mu_{s,t} = \mu_s; \mu_s \in C_s, \forall s \in S,\ \forall t = 1,2,\cdots\}.$$

It turns out that for both model the distributionally robust strategies are the same, and given by the S-robust strategy defined as follows.

Definition 4.1 *Given UMDP* $< T, \gamma, S, A_s, (\mathcal{P}^1, \cdots, \mathcal{P}^n) >$ *with* $T = \infty$ *and* $\gamma < 1$, *denote* $\hat{\mathcal{P}}_s = \{\sum_{i=1}^n (\lambda_i - \lambda_{i-1})(\mathbf{r}_s(i), \mathbf{p}_s(i)) | (\mathbf{p}_s(i), \mathbf{r}_s(i)) \in \mathcal{P}_s^i\}$, *we define the following:*

(*i*) *The* S-robust value $\tilde{v}_\infty(s)$ *is the unique solution to the following set of equations:*

$$\tilde{v}_\infty(s) = \max_{\pi_{\mathbf{s}} \in \Delta(s)} \Big\{ \min_{(\mathbf{p}_s, \mathbf{r}_s) \in \hat{\mathcal{P}}_s} \mathbb{E}_{\pi_s}^{\mathbf{p}_s}[r(s,a) + \gamma \tilde{v}_\infty(\underline{s})] \Big\}.$$

(*ii*) *The* S-robust action $\tilde{\pi}_s$ *is given by*

$$\tilde{\pi}_s \in \arg\max_{\pi_{\mathbf{s}} \in \Delta(s)} \Big\{ \min_{(\mathbf{p}_s, \mathbf{r}_s) \in \hat{\mathcal{P}}_s} \mathbb{E}_{\pi_s}^{\mathbf{p}_s}[r(s,a) + \gamma \tilde{v}_\infty(\underline{s})] \Big\}.$$

(*iii*) *A strategy* $\tilde{\pi}^*$ *is a* S-robust strategy *if* $\forall s \in S$, $\tilde{\pi}_s^*$ *is an S-robust action.*

Note that both min and max are attainable following a similar argument as the remark after Definition 3.2. To see that the S-robust strategy is well defined, it suffices to show that the following operator $\mathcal{L} : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ is a γ contraction for $\|\cdot\|_\infty$ norm.

$$\{\mathcal{L}\mathbf{v}\}(s) \triangleq \max_{\mathbf{q} \in \Delta(s)} \min_{(\mathbf{p}, \mathbf{r}) \in \hat{\mathcal{P}}_s} \{\mathcal{L}_{\mathbf{p},\mathbf{r}}^{\mathbf{q}} \mathbf{v}\}(s);$$

$$\text{where:} \quad \{\mathcal{L}_{\mathbf{p},\mathbf{r}}^{\mathbf{q}} \mathbf{v}\}(s) \triangleq \sum_{a \in A_s} q(a) r(s,a) + \gamma \sum_{a \in A_s} \sum_{s' \in S} q(a) p(s'|s,a) v(s'). \tag{9}$$

Lemma 4.1 *Under Assumption 2.1,* $\mathcal{L}$ *is a* γ *contraction for* $\|\cdot\|_\infty$ *norm.*

Proof. Observe that $\mathcal{L}_{\mathbf{p},\mathbf{r}}^{\mathbf{q}}$ is a γ contraction for any given $(\mathbf{q}, \mathbf{p}, \mathbf{r})$. For arbitrary $\mathbf{v}_1$ and $\mathbf{v}_2$, let $\mathbf{q}_{1,2}$, $\mathbf{p}_{1,2}$, $\mathbf{r}_{1,2}$ be the respective maximizing and minimizing variables. We have

$$\begin{aligned} &\{\mathcal{L}\mathbf{v}_1\}(s) - \{\mathcal{L}\mathbf{v}_2\}(s) = \mathcal{L}_{\mathbf{p}_1(s),\mathbf{r}_1(s)}^{\mathbf{q}_1(s)} \mathbf{v}_1(s) - \mathcal{L}_{\mathbf{p}_2(s),\mathbf{r}_2(s)}^{\mathbf{q}_2(s)} \mathbf{v}_2(s) \\ &\le \mathcal{L}_{\mathbf{p}_2(s),\mathbf{r}_2(s)}^{\mathbf{q}_1(s)} \mathbf{v}_1(s) - \mathcal{L}_{\mathbf{p}_2(s),\mathbf{r}_2(s)}^{\mathbf{q}_1(s)} \mathbf{v}_2(s) \le \gamma \|\mathbf{v}_1 - \mathbf{v}_2\|_\infty; \end{aligned}$$

Similarly, $\{\mathcal{L}\mathbf{v}_2\}(s) - \{\mathcal{L}\mathbf{v}_1\}(s) \le \gamma \|\mathbf{v}_2 - \mathbf{v}_1\|_\infty = \gamma \|\mathbf{v}_1 - \mathbf{v}_2\|_\infty$. Hence we establish the lemma. □

Note that given any $\mathbf{v}$, applying $\mathcal{L}$ is equivalent to solving a minimax problem, which by Theorem 3.2 can be efficiently computed. Lemma 4.1 implies that by applying $\mathcal{L}$ on any initial $\mathbf{v}^0 \in \mathbb{R}^{|S|}$ repeatedly, the resulting value vector will converge to the S-robust value $\tilde{\mathbf{v}}$ exponentially fast. Indeed, as the following lemma shows, we can compute the S-robust action for each s (and hence S-robust strategy) using this procedure.

LEMMA 4.2 *Given $s \in S$. Let $\mathbf{v}^n \triangleq \mathcal{L}^n(\mathbf{v}^0)$, and*

$$\pi_s^n \in \arg\max_{\pi_\mathbf{s} \in \Delta(s)} \Big\{ \min_{(\mathbf{p}_s, \mathbf{r}_s) \in \hat{\mathcal{P}}_s} \mathbb{E}_{\pi_s}^{\mathbf{p}_s}[r(s,a) + \gamma v^n(\underline{s})] \Big\}.$$

Then the sequence $\{\pi_s^n\}_{n=1}^\infty$ has convergent subsequences, and any of its limiting points is an S-robust action of state s.

PROOF. Note that Δ_s is compact, hence $\{\pi_s^n\}_{n=1}^\infty$ has convergent subsequences. To show that any limiting point is an S-robust action, without loss of generality we assume that $\pi_s^n \to \pi_s^*$, and denote

$$f(\pi_s, \mathbf{v}, \mathbf{p}_s, \mathbf{r}_s) \triangleq \mathbb{E}_{\pi_s}^{\mathbf{p}_s}[r(s,a) + \gamma v(\underline{s})].$$

Observe that

$$f(\pi_s, \mathbf{v}, \mathbf{p}_s, \mathbf{r}_s) \le f(\pi_s, \mathbf{v}', \mathbf{p}_s, \mathbf{r}_s) + \gamma \|\mathbf{v}' - \mathbf{v}\|.$$

By definition, for any $\hat{\pi}_s \in \Delta_s$, we have

$$\min_{(\mathbf{p}_s, \mathbf{r}_s) \in \hat{\mathcal{P}}_s} f(\pi_s^n, \mathbf{v}^n, \mathbf{p}_s, \mathbf{r}_s) \ge \min_{(\mathbf{p}_s', \mathbf{r}_s') \in \hat{\mathcal{P}}_s} f(\hat{\pi}_s, \mathbf{v}^n, \mathbf{p}_s', \mathbf{r}_s'),$$

which implies that

$$\min_{(\mathbf{p}_s, \mathbf{r}_s) \in \hat{\mathcal{P}}_s} f(\pi_s^n, \tilde{\mathbf{v}}, \mathbf{p}_s, \mathbf{r}_s) \ge \min_{(\mathbf{p}_s', \mathbf{r}_s') \in \hat{\mathcal{P}}_s} f(\hat{\pi}_s, \tilde{\mathbf{v}}, \mathbf{p}_s', \mathbf{r}_s') - 2\gamma \|\mathbf{v}^n - \tilde{\mathbf{v}}\|_\infty. \tag{10}$$

Let $(\mathbf{p}_s^*, \mathbf{r}_s^*) = \arg\min_{(\mathbf{p}_s, \mathbf{r}_s) \in \hat{\mathcal{P}}_s} f(\pi_s^*, \tilde{\mathbf{v}}, \mathbf{p}_s, \mathbf{r}_s)$, where the minimum is attainable since f is linear in $(\mathbf{p}, \mathbf{r})$, and $\hat{\mathcal{P}}_s$ is compact. We have

$$\lim_n \min_{(\mathbf{p}_s, \mathbf{r}_s) \in \hat{\mathcal{P}}_s} f(\pi_s^n, \tilde{\mathbf{v}}, \mathbf{p}_s, \mathbf{r}_s) \le \lim_n f(\pi_s^n, \tilde{\mathbf{v}}, \mathbf{p}_s^*, \mathbf{r}_s^*) = f(\pi^*, \tilde{\mathbf{v}}, \mathbf{p}_s^*, \mathbf{r}_s^*) = \min_{(\mathbf{p}_s', \mathbf{r}_s') \in \hat{\mathcal{P}}_s} f(\pi_s^*, \tilde{\mathbf{v}}, \mathbf{p}_s', \mathbf{r}_s'). \tag{11}$$

Here the first equality holds since f is continuous on π_s and $\pi_s^n \to \pi^*$, and the second equality holds due to definition of $\mathbf{p}_s^*, \mathbf{r}_s^*$. Combining Equation (10) and (11), and noting that $\mathbf{v}^n \to \tilde{\mathbf{v}}$ due to Lemma 4.1, we have

$$\min_{(\mathbf{p}_s, \mathbf{r}_s) \in \hat{\mathcal{P}}_s} f(\pi_s^*, \tilde{\mathbf{v}}, \mathbf{p}_s, \mathbf{r}_s) \ge \min_{(\mathbf{p}_s', \mathbf{r}_s') \in \hat{\mathcal{P}}_s} f(\hat{\pi}_s, \tilde{\mathbf{v}}, \mathbf{p}_s', \mathbf{r}_s'),$$

which establishes the lemma. □

In the rest of this section, we show that any S-robust strategy is distributionally robust. We consider the non-stationary model first.

THEOREM 4.1 *Under Assumption 2.1, given $T = \infty$ and $\gamma < 1$, any S-robust strategy is distributionally robust with respect to $\bar{\mathcal{C}}_S^\infty$.*

PROOF. We introduce the following $\hat{T}$-truncated problem with the total reward

$$u_{\hat{T}}(\pi, \mathbf{p}, \mathbf{r}) \triangleq \mathbb{E}_\pi^{\mathbf{p}}\{\sum_{i=1}^{\hat{T}} \gamma^{i-1} r(s_i, a_i) + \gamma^{\hat{T}} \tilde{v}_\infty(s_{\hat{T}})\}.$$

That is, the problem stops at stage $\hat{T}$ with a termination reward $\tilde{v}_\infty(\cdot)$. Note that $|S|$ is finite and all $\mathcal{P}_s^n$ are bounded. Hence there exists a universal constant c (independent of $\hat{T}$) such that for any $(\pi, \mathbf{p}, \mathbf{r})$ where $(\mathbf{p}, \mathbf{r}) \in \mathcal{P}^n$, the following holds:

$$\big|u_{\hat{T}}(\pi, \mathbf{p}, \mathbf{r}) - u(\pi, \mathbf{p}, \mathbf{r})\big| \le \gamma^{\hat{T}} c.$$

This implies for any $\mu \in \bar{\mathcal{C}}_S^\infty(\lambda)$,

$$\Big| \int u_{\hat{T}}(\pi, \mathbf{p}, \mathbf{r})\, d\mu(\mathbf{p}, \mathbf{r}) - \int u(\pi, \mathbf{p}, \mathbf{r})\, d\mu(\mathbf{p}, \mathbf{r}) \Big| \le \gamma^{\hat{T}} c, \tag{12}$$

which further leads to

$$\Big| \inf_{\mu \in \bar{\mathcal{C}}_S^\infty} \int u_{\hat{T}}(\pi, \mathbf{p}, \mathbf{r})\, d\mu(\mathbf{p}, \mathbf{r}) - \inf_{\mu' \in \bar{\mathcal{C}}_S^\infty} \int u(\pi, \mathbf{p}, \mathbf{r})\, d\mu'(\mathbf{p}, \mathbf{r}) \Big| \le \gamma^{\hat{T}} c. \tag{13}$$

By Theorem 3.1, it is easy to see that the S-robust strategy π^* is a distributionally robust strategy of the (finite horizon) $\hat{T}$ truncated problem, regardless of $\hat{T}$. That is,

$$\inf_{\mu\in\bar{\mathcal{C}}_S^\infty}\int u_{\hat{T}}(\pi^*,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r})\geq\inf_{\mu'\in\bar{\mathcal{C}}_S^\infty}\int u_{\hat{T}}(\pi',\mathbf{p},\mathbf{r})\,d\mu'(\mathbf{p},\mathbf{r}),\quad\forall\pi'\in\Pi^{HR}.$$

Combining it with Inequality (13), we have

$$\inf_{\mu\in\bar{\mathcal{C}}_S^\infty}\int u(\pi^*,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r})\geq\inf_{\mu'\in\bar{\mathcal{C}}_S^\infty}\int u(\pi',\mathbf{p},\mathbf{r})\,d\mu'(\mathbf{p},\mathbf{r})-2\gamma^{\hat{T}}c,\quad\forall\pi'\in\Pi^{HR}.$$

Notice that this holds for arbitrary $\hat{T}$, hence we have

$$\inf_{\mu\in\bar{\mathcal{C}}_S^\infty}\int u(\pi^*,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r})\geq\inf_{\mu'\in\bar{\mathcal{C}}_S^\infty}\int u(\pi',\mathbf{p},\mathbf{r})\,d\mu'(\mathbf{p},\mathbf{r}),\quad\forall\pi'\in\Pi^{HR}.$$

Thus, the S-robust strategy π^* is a distributionally robust strategy with respect to $\bar{\mathcal{C}}_S^\infty$ of the infinite horizon UMDP. □

Similar results hold for the stationary model, as the next theorem shows.

THEOREM 4.2 *Under Assumption 2.1, given $T=\infty$ and $\gamma<1$, any S-robust strategy is distributionally robust with respect to $\bar{\mathcal{C}}_S$.*

PROOF. Similar to the proof of Theorem 4.1, we consider the $\hat{T}$ truncated problem. From the proof of Theorem 3.1, for each $s\in S$ and $t\leq\hat{T}$, let $\pi_{s,t}^*,(\mathbf{p}_{s,t}^*,\mathbf{r}_{s,t}^*)$ solves the following zero-sum game

$$\max_{\pi_s}\min_{(\mathbf{p}_s,\mathbf{r}_s)\in\hat{P}_s}\mathbb{E}_\pi^{\mathbf{p}_s}\big(r(s,a)+\gamma\tilde{v}_\infty(\underline{s})\big),$$

and let $\mu_{s,t}^*\in\mathcal{C}_s$ that satisfies $\mathbb{E}_{\mu_{s,t}^*}(\mathbf{p}_{s,t},\mathbf{r}_{s,t})=(\mathbf{p}_{s,t}^*,\mathbf{r}_{s,t}^*)$, then $\pi^*=\bigotimes_{s\in S,t\leq\hat{T}}\pi_{s,t}^*$, $\mu^*=\bigotimes_{s\in S,t\leq\hat{T}}\mu_{s,t}^*$ satisfy that

$$\sup_{\pi\in\Pi^{HR}}\int u_{\hat{T}}(\pi,\mathbf{p},\mathbf{r})\,d\mu^*(\mathbf{p},\mathbf{r})=\int u_{\hat{T}}(\pi^*,\mathbf{p},\mathbf{r})\,d\mu^*(\mathbf{p},\mathbf{r})=\inf_{\mu\in\bar{\mathcal{C}}_S^\infty}\int u_{\hat{T}}(\pi^*,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r}),$$

which leads to

$$\sup_{\pi\in\Pi^{HR}}\int u(\pi,\mathbf{p},\mathbf{r})\,d\mu^*(\mathbf{p},\mathbf{r})\leq\inf_{\mu\in\bar{\mathcal{C}}_S^\infty}\int u(\pi^*,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r})+2\gamma^{\hat{T}}c\leq\inf_{\mu\in\bar{\mathcal{C}}_S}\int u(\pi^*,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r})+2\gamma^{\hat{T}}c.\tag{14}$$

Here, the first inequality holds from (12). The second inequality holds because $\bar{\mathcal{C}}_S\subseteq\bar{\mathcal{C}}_S^\infty$.

Note that by construction, π^* can be any S-robust strategy. Furthermore, $\pi_{s,t}^*$ and $\mu_{s,t}^*$ are stationary, that is, they do depend on t. Hence, we have $\mu^*\in\bar{\mathcal{C}}_S$. Therefore,

$$\sup_{\pi\in\Pi^{HR}}\inf_{\mu\in\bar{\mathcal{C}}_S}\int u(\pi,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r})\leq\sup_{\pi\in\Pi^{HR}}\int u(\pi,\mathbf{p},\mathbf{r})\,d\mu^*(\mathbf{p},\mathbf{r}).$$

This, combined with Equation (14) leads to

$$\sup_{\pi\in\Pi^{HR}}\inf_{\mu\in\bar{\mathcal{C}}_S}\int u(\pi,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r})\leq\inf_{\mu\in\bar{\mathcal{C}}_S}\int u(\pi^*,\mathbf{p},\mathbf{r})\,d\mu(\mathbf{p},\mathbf{r})+2\gamma^{\hat{T}}c.$$

Since $\hat{T}$ can be arbitrarily large, this shows that the S-robust strategy π^* is a distributionally robust strategy with respect to $\bar{\mathcal{C}}_S$. □

Before concluding this section, we briefly compare the stationary model and the non-stationary model. These two formulations model different setups: if the system, more specifically the distribution of uncertain parameters, evolves with time, then non-stationary model is more appropriate; while if the system is static, then stationary model is preferable. For any given strategy, the worst expected performance under the non-stationary model provides a lower bound to that of the stationary model, due to $\bar{\mathcal{C}}_S\subseteq\bar{\mathcal{C}}_S^\infty$. Thus, one can use the non-stationary model to approximate the stationary model, when the latter is intractable (e.g., in the finite horizon case; see Nilim and El Ghaoui [25]). When horizon approaches infinity, such approximation becomes exact, as we showed in this section, the optimal solutions to both formulations coincide, and can be computed by iteratively solving a minimax problem.

5. Concluding remarks. In this paper we addressed MDPs under parameter uncertainty following the distributionally robust approach, to mitigate the conservatism of the robust MDP framework and incorporate additional *a priori* probabilistic information regarding the unknown parameters. In particular, we considered the nested-set structured parameter uncertainty to model *a priori* probabilistic information of the parameters. We proposed to find a strategy that achieves maximum expected utility under the worst admissible distribution of the parameters. Our formulation leads to an optimal strategy that is obtained through a Bellman type backward induction, and can be solved in polynomial time under mild technical conditions.

Before concluding this paper we discuss how the proposed uncertainty model relates to some standard methods to model uncertainty. As we showed, algorithmically, solving the proposed nested-set formulation can be reduced to a robust MDP with a single uncertainty set. However, the main advantage of the proposed approach is that it provides extra *modeling* flexibility. One of the main open problems in robust optimization (and robust MDPs) is uncertainty set selection. The most natural approach uses confidence intervals around the nominal parameter as the uncertainty set, and fail to incorporate extra probabilistic information. Our approach provides a principled method for choosing an uncertainty set that incorporates the available information, following the distributionally robustness framework, which has attracted much attention recently for single-stage optimization (e.g., Popescu [26], Delage and Ye [12], Calafiore and El Ghaoui [10], and Goh and Sim [18]). Indeed, as far as we know, we are the first to incorporate distributional robustness in MDPs, and multi-stage decision making in general. Interestingly, it has been observed in robust optimization that shrinking the uncertainty set often leads to better performance. Following our distributionally robustness framework, shrinking the uncertainty set is equivalent to solving a distributionally robust optimization problem with multiple nested-sets, and it is thus not surprising that when uncertain parameters are not completely adversarial, such formulation improves the practical performance.

A different approach to embedding prior information is by adopting a Bayesian perspective on the parameters of the problem; see Delage and Mannor [11] and references therein. However, a complete Bayesian prior to the model parameters may be difficult to conjure as the decision maker may not have a reliable generative model to the uncertainty. Our approach offers a middle ground between the fully Bayesian approach and the robust approach: we want the decision maker to be able to use prior information but we do not require a complete Bayesian interpretation.

Acknowledgements We thank the referees and the associate editor for providing helpful comments that led to improvements of the manuscript. The research of H. Xu was supported partially by the National University of Singapore under startup grant R-265-000-384-133. The research of S. Mannor was partially supported by the Israel Science Foundation (contract 890015).

References

[1] M. Abbad and J. Filar, *Perturbation and stability theory for Markov control problems*, IEEE Transactions on Automatic Control **37** (1992), no. 9, 1415–1420.

[2] M. Abbad, J.A. Fillar, and T.R. Bielecki, *Algorithms for singularly perturbed limiting average Markov control problems*, IEEE Transactions on Automatic Control **37** (1992), no. 9, 1421–1425.

[3] K. E. Avrachenkov, J. Filar, and M. Haviv, *Singular perturbations of Markov chains and decision processes*, Handbook of Markov Decision Processes: Methods and Applications (E. A. Feinberg and A. Shwartz, eds.), 2002.

[4] A. Bagnell, A. Ng, and J. Schneider, *Solving uncertain Markov decision problems*, Tech. Report CMU-RI-TR-01-25, Carnegie Mellon University, August 2001.

[5] J. Baron, *Thinking and deciding*, Cambridge University Press, 2000.

[6] A. Ben-Tal and A. Nemirovski, *Robust solutions of uncertain linear programs*, Operations Research Letters **25** (1999), no. 1, 1–13.

[7] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*, Athena Scientific, 1996.

[8] D. Blackwell and M. Girshick, *Theory of games and statistical decisions*, New York: John Wiley & Sons Inc, 1954.

[9] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[10] G. Calafiore and L. El Ghaoui, *On distributionally robust chance-constrained linear programs*, Journal of Optimization Theory and Applications **130** (2006), no. 1, 1–22.

[11] E. Delage and S. Mannor, *Percentile optimization for Markov decision processes with parameter uncertainty*, Operations Research **58** (2010), no. 1, 203–213.

[12] E. Delage and Y. Ye, *Distributionally robust optimization under moment uncertainty with applications to data-driven problems*, Operations Research **58** (2010), no. 3, 596–612.

[13] F. Delebecque and J. P. Quadrat, *Optimal control of Markov chains admitting strong and weak interactions*, Automatica **17** (1981), no. 2, 281–296.

[14] J. Dupacová, *The minimax approach to stochastic programming and an illustrative application*, Stochastics **20** (1987), 73–88.

[15] A. Dvoretzky, A. Wald, and J. Wolfowitz, *Elimination of randomization in certain statistical decision procedures and zero-sum two-person games*, Annals of Mathematical Statistics **22** (1951), no. 1, 1–21.

[16] L. G. Epstein and M. Schneider, *Learning under ambiguity*, Review of Economic Studies **74** (2007), no. 4, 1275–1303.

[17] I. Gilboa and D. Schmeidler, *Maxmin expected utility with a non-unique prior*, Journal of Mathematical Economics **18** (1989), no. 2, 141–153.

[18] J. Goh and M. Sim, *Distributionally robust optimization and its tractable approximations*, Operations Research **58** (2010), no. 4, 902–917.

[19] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric algorithms and combinatorial optimization*, Springer, Heidelberg, 1988.

[20] G. N. Iyengar, *Robust dynamic programming*, Mathematics of Operations Research **30** (2005), no. 2, 257–280.

[21] P. Kall, *Stochastic programming with recourse: Upper bounds and moment problems, a review*, Advances in Mathematical Optimization, Academie-Verlag, Berlin, 1988.

[22] S. Karlin, *The theory of infinite games*, Annals of Mathematics **58** (1953), no. 2, 371–401.

[23] D. Kelsey, *Maxmin expected utility and weight of evidence*, Oxford Economic Papers **46** (1994), no. 3, 425–444.

[24] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis, *Bias and variance approximation in value function estimates*, Management Science **53** (2007), no. 2, 308–322.

[25] A. Nilim and L. El Ghaoui, *Robust control of Markov decision processes with uncertain transition matrices*, Operations Research **53** (2005), no. 5, 780–798.

[26] I. Popescu, *Robust mean-covariance solutions for stochastic optimization*, Operations Research **55** (2007), no. 1, 98–112.

[27] M. L. Puterman, *Markov decision processes*, John Wiley & Sons, New York, 1994.

[28] R.T. Rockafellar, *Convex analysis*, Princeton University Press, Princeton, N.J., 1970.

[29] H. Scarf, *A min-max solution of an inventory problem*, Studies in Mathematical Theory of Inventory and Production, Stanford University Press, 1958, pp. 201–209.

[30] A. Shapiro, *Worst-case distribution analysis of stochastic programs*, Mathematical Programming **107** (2006), no. 1, 91–96.

[31] L. S. Shapley, *Stochastic games*, Proceedings of the National Academy of Sciences of USA **39** (1953), no. 10, 1095–1100.

[32] M. Sion, *On general minimax theorems*, Pacific Journals of Mathematics **8** (1958), no. 1, 171–176.

[33] C. C. White III and H. K. El Deib, *Markov decision processes with imprecise transition probabilities*, Operations Research **42** (1992), no. 4, 739–748.