Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

A Robust Approach to Markov Decision Problems with Uncertain Transition Probabilities [⋆]

Ioannis Ch. Paschalidis [*] **Seong-Cheol Kang** [**]

[*] *Center for Information & Systems Eng., Boston University, Brookline, MA 02446, USA (yannisp@bu.edu, http://ionia.bu.edu/)*
[**] *Boston University, Brookline, MA 02446, USA (jsckang@bu.edu)*

Abstract: This paper considers a discrete-time infinite horizon discounted cost Markov decision problem in which the transition probability vector for each state-control pair is uncertain. A popular approach to this problem has been to find a policy that performs best in the worst-case scenario. A policy obtained in this manner, however, tends to be conservative. We construct a robust formulation for the problem, which produces a less conservative policy. We characterize the performance of the robust formulation via the probability that the optimal cost of a random instance of the problem is at most that of the robust formulation. A congestion-dependent pricing problem for network services is examined as a numerical example.

1. INTRODUCTION

A *Markov Decision Problem (MDP)* is a stochastic sequential decision making problem, whose defining characteristic is the Markov property: given the current state, transitions to a new state in the future are independent of all the past states. The probability distribution of those transitions is often assumed to be precisely known. The objective is then to find a decision policy (i.e., a decision rule) that optimizes a certain cost criterion for the problem.

The validity of the assumption that the transition probabilities are precisely known, however, has been debated. Some argue that in many real-world applications the transition probabilities may have to be estimated from historical data or observations. As such, they are subject to estimation errors, making them *uncertain* (or *ambiguous*).

When the transition probabilities are uncertain, one could choose to use some representative values for them, ignoring the uncertainty. This approach has been deemed unacceptable because an optimal policy of an MDP is typically sensitive to the transition probabilities. An alternative approach is to take all possible scenarios for the transition probabilities into consideration and to seek a policy that performs best in the worst-case scenario. Indeed, the literature has focused on this worst-case approach: Satia and Lave [1973], White and Eldeib [1994], Nilim and El Ghaoui [2005], Iyengar [2005].

To explain the motivation for our work, let us step back and think of general constrained optimization problems in which problem data are uncertain. For those problems, the classical robust optimization approach aims to find a solution that is immune to data uncertainty (i.e., a solution with guaranteed feasibility). See, for example, Ben-Tal and Nemirovski [1998] that studied various classes of constrained optimization problems from this perspective.

Noting that the classical robust optimization approach tends to produce an ultra conservative solution when applications can tolerate a small chance of infeasibility, Bertsimas and Sim [2004] and Paschalidis and Kang [2005, 2006] considered a “relaxed” robust optimization approach to linear programming problems with data uncertainty. The goal of this approach is to produce an improved solution with a certain probabilistic guarantees of feasibility. In particular, Paschalidis and Kang [2005, 2006] showed that if one exploits distributional information on the uncertain data in this approach, the resulting solution becomes more cost-effective.

The worst-case approach to an MDP with uncertain transition probabilities is, in spirit, the same as the classical robust optimization approach to a constrained optimization problem with uncertain data. Therefore, it is reasonable to suspect that an optimal policy from the worst-case approach could also be too conservative in some cases. This motivates us to extend the ideas of Paschalidis and Kang [2005, 2006] to the MDP in an attempt to find a less conservative policy. In this paper, we consider a formulation, referred to as the *robust MDP*, that produces such a policy. To characterize the performance of this robust formulation, we examine the probability that the optimal cost of a random instance of the MDP is at most that of the robust formulation. To the best of our knowledge, this line of analysis has not been pursued in the literature.

This paper is organized as follows. In Section 2, we describe the MDP under consideration and provide background information. We then define the robust MDP and present a probabilistic characterization of its performance in Section 3. To illustrate our robust approach numerically, we consider a congestion-dependent pricing problem for network services in Section 4. We conclude in Section 5.

[⋆] Research partially supported by the NSF under grants DMI-0330171, CNS-0435312, ECS-0426453, EFRI-0735974, and by the DOE under grant DE-FG52-06NA27490.

978-3-902661-00-5/08/$20.00 © 2008 IFAC

10.3182/20080706-5-KR-1001.0701

Notation: We use boldface letters to denote vectors. All vectors are assumed to be column vectors. $\mathbf{x}'$ represents the transpose of the vector $\mathbf{x}$. The vector of all zeros is denoted by $\mathbf{0}$. $P[A]$ means the probability of the event A, and $E[X]$ denotes the mean of the random variable X.

2. PROBLEM SETTING

We consider a discrete-time infinite horizon discounted cost MDP with finite state space S and control space U. For notational simplicity, we assume that U is state-independent, i.e., each state has the identical control space U. When the system is at state $i \in S$ and control $u \in U$ is taken, the stationary cost $c(i,u)$ is incurred, which we assume nonnegative and bounded. The system then makes a transition to state j with probability $p_{ij}(u) \triangleq P[j\,|\,i,u]$. At the new state j, the process is repeated. Let $\mathbf{p}_i(u) = (p_{ij}(u))_{j\in S}$ be the transition probability vector associated with the state-control pair (i,u). A stationary policy π is a mapping from S to U, and let Π be the set of all (allowable) stationary policies. Let $0 < \alpha < 1$ be the discount factor.

In the framework of standard MDPs, it is assumed that the transition probability vectors $\mathbf{p}_i(u), \forall\, i \in S, u \in U$, are precisely known. Let us denote by ω the collection of the transition probability vectors, i.e., $\omega \triangleq \{\mathbf{p}_i(u)\}_{i\in S, u\in U}$. Given the initial state $i_0 = i$, the cost of policy π is calculated as

$$V_\omega^\pi(i) = E_\omega\Big[\sum_{t=0}^{\infty} \alpha^t c(i_t, \pi(i_t)) \mid i_0 = i\Big],$$

where i_t is the state at epoch t and the expectation is taken with respect to ω. The objective is then to find an optimal policy π^* that minimizes $V_\omega^\pi(i)$ for all $i \in S$, i.e.,

$$V_\omega^*(i) = \min_{\pi\in\Pi} V_\omega^\pi(i), \quad \forall\, i \in S. \tag{1}$$

When ω represents the collection of the nominal transition probability vectors, we refer to (1) as the *nominal MDP* and use the notation $V_N^*(i)$ and π_N instead of $V_\omega^*(i)$ and π^*, respectively.

It is well known that the optimal value function V_ω^* satisfies Bellman equations and that V_ω^* and π^* can be determined by value or policy iteration (see, for instance, Puterman [1994], Bertsekas [2005, 2007]). We also note that one can solve (1) through the linear programming formulation

$$\max \sum_{i\in S} V(i) \tag{2}$$
$$\text{s.t. } V(i) \le c(i,u) + \alpha \sum_{j\in S} p_{ij}(u) V(j), \quad \forall\, i \in S, u \in U,$$

whose optimal solution $V(i)$ is equal to $V_\omega^*(i)$ for all $i \in S$.

The uncertainty in the transition probabilities can be modeled by assuming that $\mathbf{p}_i(u)$ belongs to some bounded set $\Omega_i(u)$. For instance, $\Omega_i(u)$ can be described as $\Omega_i(u) \triangleq \{\mathbf{p} \mid \underline{\mathbf{p}} \le \mathbf{p} \le \overline{\mathbf{p}},\ \mathbf{p} \in \Delta_{|S|}\}$, where $\overline{\mathbf{p}} \ge \underline{\mathbf{p}} \ge \mathbf{0}$ and $\Delta_{|S|}$ is the probability simplex in $\mathbb{R}^{|S|}$. Let $\Omega \triangleq \times\Omega_i(u)$, i.e., the Cartesian product of $\Omega_i(u)$.

When the transition probabilities are uncertain, the worst-case approach considered in the literature seeks a policy that minimizes the worst possible cost. Such a policy, denoted by π_F, is obtained through the following *classical robust MDP* or *"fat" MDP*:

$$V_F^*(i) = \min_{\pi\in\Pi} \max_{\omega\in\Omega} E_\omega\Big[\sum_{t=0}^{\infty} \alpha^t c(i_t, \pi(i_t)) \mid i_0 = i\Big], \quad \forall\, i \in S. \tag{3}$$

We refer to π_F as the *fat policy.* Nilim and El Ghaoui [2005] showed that $V_F^*(i), \forall\, i \in S$, satisfy the following set of equations, which we call the *fat Bellman equations*:

$$V_F^*(i) = \min_{u\in U}\Big\{c(i,u) + \alpha \max_{\mathbf{p}_i(u)\in\Omega_i(u)} \sum_{j\in S} p_{ij}(u) V_F^*(j)\Big\}, \ \forall\, i, \tag{4}$$

with π_F being found through

$$\pi_F(i) = \operatorname*{argmin}_{u\in U}\Big\{c(i,u) + \alpha \max_{\mathbf{p}_i(u)\in\Omega_i(u)} \sum_{j\in S} p_{ij}(u) V_F^*(j)\Big\}, \forall\, i.$$

Nilim and El Ghaoui [2005] proposed a value iteration algorithm for solving (4). (Iyengar [2005] independently proved the validity of (4) and developed value and policy iteration algorithms.)

One might be tempted to use the linear programming approach (cf. (2)) for solving the fat MDP (3) by formulating it as

$$\max \sum_{i\in S} V(i)$$
$$\text{s.t. } V(i) \le c(i,u) + \alpha \max_{\mathbf{p}_i(u)\in\Omega_i(u)} \sum_{j\in S} p_{ij}(u) V(j), \quad \forall\, i, u.$$

Unfortunately, this formulation is not a convex problem: the max operator in the constraints makes the feasible set nonconvex. So it is unlikely that efficient exact solution algorithms exist.

Let $V_\omega^{\pi_F}(i)$ be the cost of the fat policy for the initial state $i_0 = i$ when state transitions occur according to a given ω, i.e.,

$$V_\omega^{\pi_F}(i) = E_\omega\Big[\sum_{t=0}^{\infty} \alpha^t c(i_t, \pi_F(i_t)) \mid i_0 = i\Big].$$

Lemma 1. For any $\omega \in \Omega$, $V_\omega^*(i) \le V_\omega^{\pi_F}(i) \le V_F^*(i)$ for all $i \in S$.

Proof. The first inequality follows from the fact that π_F is a suboptimal policy of the MDP with ω. The second inequality holds because

$$V_\omega^{\pi_F}(i) = E_\omega\Big[\sum_{t=0}^{\infty} \alpha^t c(i_t, \pi_F(i_t)) \mid i_0 = i\Big]$$
$$\le \max_{\omega\in\Omega} E_\omega\Big[\sum_{t=0}^{\infty} \alpha^t c(i_t, \pi_F(i_t)) \mid i_0 = i\Big] = V_F^*(i).$$

■

Lemma 1 shows that for any $\omega \in \Omega$, the optimal cost $V_\omega^*(i)$ is guaranteed to be no greater than the *fat cost* $V_F^*(i)$. Put differently, $V_F^*(i)$ is an *a priori* upper bound on $V_\omega^*(i)$ that cannot be determined until ω is realized. In this sense, the fat MDP is conservative.

3. THE ROBUST MARKOV DECISION PROBLEM

The rationale for considering the fat MDP (3) is to protect against the case where a certain set of transition probability vectors causes a high cost. However, if such a case happens rarely, the use of the fat policy would be unnecessarily conservative. Thus, one would be interested in finding a policy that is less conservative than the fat policy, but permitting a small possibility of the policy being a bad one.

In order to obtain a less conservative policy than the fat policy, we restrict the uncertainty sets $\Omega_i(u)$. Let $\mathscr{R}_i(u) \subseteq \Omega_i(u)$ and $\mathscr{R} \triangleq \times\mathscr{R}_i(u)$. We define the *robust MDP* as

$$V_R^*(i) = \min_{\pi\in\Pi}\max_{\omega\in\mathscr{R}} E_\omega\Big[\sum_{t=0}^{\infty}\alpha^t c(i_t,\pi(i_t)) \mid i_0 = i\Big], \quad \forall\, i \in S. \tag{5}$$

Let π_R denote an optimal policy of the robust MDP, which will be referred to as the *robust policy*. Similarly to the fat MDP, $V_R^*(i)$, $\forall\, i \in S$, satisfy the following *robust Bellman equations*:

$$V_R^*(i) = \min_{u\in U}\Big\{c(i,u) + \alpha \max_{\mathbf{p}_i(u)\in\mathscr{R}_i(u)} \sum_{j\in S} p_{ij}(u)V_R^*(j)\Big\}, \ \forall\, i, \tag{6}$$

with

$$\pi_R(i) = \operatorname*{argmin}_{u\in U}\Big\{c(i,u) + \alpha \max_{\mathbf{p}_i(u)\in\mathscr{R}_i(u)} \sum_{j\in S} p_{ij}(u)V_R^*(j)\Big\}, \forall\, i.$$

One can use a value iteration algorithm (or a policy iteration algorithm) to determine $V_R^*(i)$ and $\pi_R(i)$ for all $i \in S$. The following lemma formalizes the argument that the robust policy π_R is less conservative than the fat policy π_F.

Lemma 2. $V_R^*(i) \le V_F^*(i)$ for all $i \in S$.

Proof.

$$\begin{aligned} V_R^*(i) &= \min_{\pi\in\Pi}\max_{\omega\in\mathscr{R}} E_\omega\Big[\sum_{t=0}^{\infty}\alpha^t c(i_t,\pi(i_t)) \mid i_0 = i\Big] \\ &\le \max_{\omega\in\mathscr{R}} E_\omega\Big[\sum_{t=0}^{\infty}\alpha^t c(i_t,\pi_F(i_t)) \mid i_0 = i\Big] \\ &\le \max_{\omega\in\Omega} E_\omega\Big[\sum_{t=0}^{\infty}\alpha^t c(i_t,\pi_F(i_t)) \mid i_0 = i\Big] = V_F^*(i). \end{aligned}$$

■

Having defined the robust MDP, we now characterize its performance by comparing its optimal cost with the optimal cost of a random instance of the MDP. Specifically, we are interested in the probability $P\big[V_\omega^*(i) \le V_R^*(i)\big]$ for a randomly selected $\omega \in \Omega$. To that end, let us consider the cost of the robust policy, $V_\omega^{\pi_R}(i)$, for the MDP with ω when the initial state is $i_0 = i$, i.e.,

$$V_\omega^{\pi_R}(i) = E_\omega\Big[\sum_{t=0}^{\infty}\alpha^t c(i_t,\pi_R(i_t)) \mid i_0 = i\Big].$$

Lemma 3. For any $\omega \in \Omega$, $V_\omega^*(i) \le V_\omega^{\pi_R}(i)$ for all $i \in S$. Moreover, if $\omega \in \mathscr{R}$, then $V_\omega^{\pi_R}(i) \le V_R^*(i)$ for all $i \in S$.

Proof. The first part follows from the fact that π_R is a suboptimal policy for the MDP with ω. For the second part, if $\omega \in \mathscr{R}$,

$$\begin{aligned} V_\omega^{\pi_R}(i) &= E_\omega\Big[\sum_{t=0}^{\infty}\alpha^t c(i_t,\pi_R(i_t)) \mid i_0 = i\Big] \\ &\le \max_{\omega\in\mathscr{R}} E_\omega\Big[\sum_{t=0}^{\infty}\alpha^t c(i_t,\pi_R(i_t)) \mid i_0 = i\Big] = V_R^*(i). \end{aligned}$$

■

It follows from Lemma 3 that

$$P\big[V_\omega^*(i) \le V_R^*(i)\big] \ge P\big[V_\omega^{\pi_R}(i) \le V_R^*(i)\big]. \tag{7}$$

Consider the probability of the complement of $V_\omega^{\pi_R}(i) \le V_R^*(i)$.

$$\begin{aligned} &P\big[V_\omega^{\pi_R}(i) > V_R^*(i)\big] \\ &\quad = P\big[V_\omega^{\pi_R}(i) > V_R^*(i) \mid \omega\in\mathscr{R}\big]P\big[\omega\in\mathscr{R}\big] \\ &\qquad + P\big[V_\omega^{\pi_R}(i) > V_R^*(i) \mid \omega\notin\mathscr{R}\big]P\big[\omega\notin\mathscr{R}\big] \\ &\quad = P\big[V_\omega^{\pi_R}(i) > V_R^*(i) \mid \omega\notin\mathscr{R}\big]P\big[\omega\notin\mathscr{R}\big], \end{aligned} \tag{8}$$

where the second equality follows from Lemma 3. Let $\mathbf{p}_i(\pi_R(i)) \in \omega$ be the transition probability vector for the state-control pair $(i, \pi_R(i))$. Since $V_\omega^{\pi_R}(i)$ and $\mathbf{p}_i(\pi_R(i))$ satisfy $V_\omega^{\pi_R}(i) = c(i,\pi_R(i)) + \alpha\sum_{j\in S} p_{ij}(\pi_R(i))V_\omega^{\pi_R}(j)$, we can write the first probability in (8) as

$$\begin{aligned} &P\big[V_\omega^{\pi_R}(i) > V_R^*(i) \mid \omega\notin\mathscr{R}\big] \\ &= P\Big[c(i,\pi_R(i)) + \alpha\sum_{j\in S} p_{ij}(\pi_R(i))V_\omega^{\pi_R}(j) > V_R^*(i) \mid \omega\notin\mathscr{R}\Big] \\ &= P\Big[\sum_{j\in S} p_{ij}(\pi_R(i))V_\omega^{\pi_R}(j) > C(i) \mid \omega\notin\mathscr{R}\Big], \end{aligned} \tag{9}$$

where $C(i) = \frac{1}{\alpha}\big\{V_R^*(i) - c(i,\pi_R(i))\big\}$.

The $V_\omega^{\pi_R}(i)$ in (9) cannot be computed until ω is realized. To rid $V_\omega^{\pi_R}(i)$ of their dependency on a particular ω, we calculate the worst cost of the policy π_R for all $\omega \notin \mathscr{R}$ as follows:

$$V^{\pi_R}(i) = \max_{\omega\notin\mathscr{R}} E_\omega\Big[\sum_{t=0}^{\infty}\alpha^t c(i_t,\pi_R(i_t)) \mid i_0 = i\Big], \quad \forall\, i \in S.$$

The $V^{\pi_R}(i)$ can be obtained through the following set of equations: for all $i \in S$

$$V^{\pi_R}(i) = c(i,\pi_R(i)) + \alpha \max_{\mathbf{p}_i(\pi_R(i))\in\omega\notin\mathscr{R}} \sum_{j\in S} p_{ij}(\pi_R(i))V^{\pi_R}(j). \tag{10}$$

It may not be easy to compute $V^{\pi_R}(i)$ because the requirement of $\mathbf{p}_i(\pi_R(i)) \in \omega \notin \mathscr{R}$ could make the maximization problem in (10) complicated. In that case, one may use $\widehat{V}^{\pi_R}(i)$ instead of $V^{\pi_R}(i)$, which are the solution of[1]

$$\begin{aligned} \widehat{V}^{\pi_R}(i) &= c(i,\pi_R(i)) \\ &\quad + \alpha \max_{\mathbf{p}_i(\pi_R(i))\in\Omega_i(\pi_R(i))} \sum_{j\in S} p_{ij}(\pi_R(i))\widehat{V}^{\pi_R}(j), \quad \forall\, i \in S. \end{aligned}$$

Replacing $V_\omega^{\pi_R}(i)$ in (9) with $V^{\pi_R}(i)$, we obtain

$$\begin{aligned} &P\Big[\textstyle\sum_{j\in S} p_{ij}(\pi_R(i))V_\omega^{\pi_R}(j) > C(i) \mid \omega\notin\mathscr{R}\Big] \\ &\quad \le P\Big[\textstyle\sum_{j\in S} p_{ij}(\pi_R(i))V^{\pi_R}(j) > C(i) \mid \omega\notin\mathscr{R}\Big] \\ &\quad \le P\big[\mathbf{V}'\mathbf{p}_i(\pi_R(i)) \ge C(i) \mid \omega\notin\mathscr{R}\big], \end{aligned} \tag{11}$$

where $\mathbf{V}$ is the vector whose components are the $V^{\pi_R}(i)$.

[1] The use of $\widehat{V}^{\pi_R}(i)$ could make the analysis weaker.

Putting (8), (9), and (11) together and using Markov's inequality, we obtain for $\theta \geq 0$

$$\begin{aligned} &P\big[V_\omega^{\pi_R}(i) > V_R^*(i)\big] \\ &\quad \leq P\big[\mathbf{V}'\mathbf{p}_i(\pi_R(i)) \geq C(i) \mid \omega \notin \mathscr{R}\big] P\big[\omega \notin \mathscr{R}\big] \\ &\quad \leq e^{-\theta C(i)} E\big[e^{\theta \mathbf{V}'\mathbf{p}_i(\pi_R(i))} \mid \omega \notin \mathscr{R}\big] P\big[\omega \notin \mathscr{R}\big] \\ &\quad = \exp\big[-\theta C(i) + \Lambda_{\mathbf{p}_i(\pi_R(i)) \in \omega \notin \mathscr{R}}(\theta \mathbf{V})\big] P\big[\omega \notin \mathscr{R}\big], \end{aligned}$$

where $\Lambda_{\mathbf{p}_i(\pi_R(i)) \in \omega \notin \mathscr{R}}(\theta \mathbf{V}) \triangleq \log E\big[e^{\theta \mathbf{V}'\mathbf{p}_i(\pi_R(i))} \mid \omega \notin \mathscr{R}\big]$. Optimizing over θ, we arrive at the following proposition.

Proposition 4. It holds that

$$\begin{aligned} &P\big[V_\omega^{\pi_R}(i) > V_R^*(i)\big] \\ &\quad \leq \exp\Big[\inf_{\theta \geq 0} \Big\{-\theta C(i) + \Lambda_{\mathbf{p}_i(\pi_R(i)) \in \omega \notin \mathscr{R}}(\theta \mathbf{V})\Big\}\Big] P\big[\omega \notin \mathscr{R}\big]. \end{aligned} \tag{12}$$

Let ϵ be the value of the right hand side of (12). From (7), we then have $P\big[V_\omega^*(i) \leq V_R^*(i)\big] \geq 1 - \epsilon$. In other words, no matter what the transition probabilities are, $V_\omega^*(i)$ is no greater than $V_R^*(i)$ with probability at least $1 - \epsilon$. In general, computing the probability bound in (12) exactly would pose computational challenges. Sometimes, however, ω is induced by a few parameters (as is the case in the example in Section 4). In this case, the computational challenges could be mild.

4. AN EXAMPLE

To numerically demonstrate the discussions of Section 3, we consider a revenue maximization problem for congestion-dependent pricing for certain network services. This problem, which will be described below, is a simplified version of the one discussed extensively in Paschalidis and Tsitsiklis [2000] (with the known transition probabilities).

Consider a service provider who provides access to a communication network or some other form of on-line services. Service requests, say "calls", arrive according to a Poisson process and stay connected to the network for a time interval that is exponentially distributed with rate μ. Connection time intervals of calls are independent and are also independent of interarrival times of calls. The service provider has a total amount R of some resource, say "bandwidth". Each incoming call requires r units of bandwidth and is only accepted if that bandwidth is available. Otherwise, the call is rejected and lost. When a call arrives and is accepted, it pays a fee of u. The service provider can change the value of u at the times when a call arrives or when a call departs. We assume that there is a known demand function $\lambda(u)$, which determines the arrival rate of calls as a function of u. We further assume that there exists a fee $u_{\max}$ beyond which the demand $\lambda(u)$ becomes zero.

In this setting, the goal is to determine an optimal policy of setting the value of u as a function of the number of calls in the system, so that the following long-term (infinite horizon) total discounted revenue is maximized:

$$\lim_{T \to \infty} E\Big[\int_0^T e^{-\alpha t} \lambda(u(t)) u(t) dt\Big],$$

where $u(t)$ is the fee at time t and α the discount factor.

Let us denote the maximum number of calls that the system can admit by $K \triangleq \lfloor R/r \rfloor$. We define the state space as $S = \{1, \ldots, K\}$ and the control space as $U = [0, u_{\max}]$. The above continuous-time MDP can be converted to a discrete-time one through uniformization, whose Bellman equations are given by

$$V(i) = \frac{1}{\alpha + \nu} \max_{u \in U} \Big\{g(i, u) + \nu \sum_{j \in S} p_{ij}(u) V(j)\Big\}, \ \forall i, \tag{13}$$

where $V(i)$ is the maximum long-term revenue when there are i calls in the system, $g(i, u)$ is the "per-stage" revenue for the state-control pair (i, u), and ν is the uniformization constant to be determined later. (We refer the reader to Bertsekas [2007] about the conversion of continuous-time MDPs to discrete-time counterparts by uniformization.) The per-stage revenue is given by

$$g(i, u) = \begin{cases} \lambda(u) u & \text{if } i \neq K, \\ 0 & \text{if } i = K. \end{cases} \tag{14}$$

The transition probabilities are

$$p_{ij}(u) = \begin{cases} \lambda(u)/\nu & \text{if } i \neq K \text{ and } j = i + 1, \\ i\mu/\nu & \text{if } i \neq 0 \text{ and } j = i - 1, \\ 1 - \frac{\lambda(u)}{\nu} - \frac{i\mu}{\nu} & \text{if } 1 \leq i = j \leq K - 1, \\ 1 - \lambda(u)/\nu & \text{if } i = j = 0, \\ 1 - i\mu/\nu & \text{if } i = j = K, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

We consider a linear demand function with an uncertain "y-intercept". Let $\lambda(u) = \lambda_0 - \lambda_1 u$, and assume that λ_0 is uniformly distributed over the interval $[\overline{\lambda}_0 - \hat{\lambda}_0, \overline{\lambda}_0 + \hat{\lambda}_0]$, where $\overline{\lambda}_0 > \hat{\lambda}_0 > 0$. Hence for any fee u, the arrival rate of calls belongs to the interval $[\overline{\lambda}_0 - \hat{\lambda}_0 - \lambda_1 u, \overline{\lambda}_0 + \hat{\lambda}_0 - \lambda_1 u]$. Note that the uniformization constant can be set to $\nu = \overline{\lambda}_0 + \hat{\lambda}_0 + K\mu$.

In the nominal MDP, λ_0 is set to $\overline{\lambda}_0$. Inserting the per-stage revenue (14) and the transition probabilities (15) into (13), we have

$$\begin{aligned} V_N^*(0) &= \frac{1}{\alpha + \nu} \max_{0 \leq u \leq u_{\max}} \Big\{(\overline{\lambda}_0 - \lambda_1 u) u \\ &\quad + (\overline{\lambda}_0 - \lambda_1 u) V_N^*(1) + \big(\nu - (\overline{\lambda}_0 - \lambda_1 u)\big) V_N^*(0)\Big\}, \\ V_N^*(i) &= \frac{1}{\alpha + \nu} \max_{0 \leq u \leq u_{\max}} \Big\{(\overline{\lambda}_0 - \lambda_1 u) u \\ &\quad + (\overline{\lambda}_0 - \lambda_1 u) V_N^*(i + 1) + i\mu V_N^*(i - 1) \\ &\quad + \big(\nu - (\overline{\lambda}_0 - \lambda_1 u) - i\mu\big) V_N^*(i)\Big\}, \ 1 \leq i \leq K-1, \\ V_N^*(K) &= \frac{1}{\alpha + \nu} \max_{0 \leq u \leq u_{\max}} \Big\{i\mu V_N^*(K-1) \\ &\quad + (\nu - i\mu) V_N^*(K)\Big\} \\ &= \frac{1}{\alpha + \nu} \Big\{i\mu V_N^*(K-1) + (\nu - i\mu) V_N^*(K)\Big\}. \end{aligned}$$

In the fat MDP, λ_0 can take any value from its range. This introduces uncertainty in the transition probabilities because some components of $\mathbf{p}_i(u)$ are functions of λ_0. Specifically, it leads to $\mathbf{p}_i(u) \in \Omega_i(u) = \big\{\mathbf{p}(\lambda_0) \mid \lambda_0 \in [\overline{\lambda}_0 - \hat{\lambda}_0, \overline{\lambda}_0 + \hat{\lambda}_0]\big\}$, where if $1 \leq i \leq K-1$

$$p_j(\lambda_0) = \begin{cases} (\lambda_0 - \lambda_1 u)/\nu & \text{if } j = i+1, \\ i\mu/\nu & \text{if } j = i-1, \\ 1 - (\lambda_0 - \lambda_1 u)/\nu - i\mu/\nu & \text{if } j = i, \\ 0 & \text{otherwise,} \end{cases}$$

and if $i = 0$

$$p_j(\lambda_0) = \begin{cases} (\lambda_0 - \lambda_1 u)/\nu & \text{if } j = i+1, \\ 1 - (\lambda_0 - \lambda_1 u)/\nu & \text{if } j = i, \\ 0 & \text{otherwise,} \end{cases}$$

and if $i = K$

$$p_j(\lambda_0) = \begin{cases} i\mu/\nu & \text{if } j = i-1, \\ 1 - i\mu/\nu & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

We note that it is also possible to model $\Omega_i(u)$ as $\Omega_i(u) = \big\{\mathbf{p} \mid \underline{\mathbf{p}} \le \mathbf{p} \le \overline{\mathbf{p}},\, \mathbf{p} \in \Delta_K\big\}$, for appropriately defined $\underline{\mathbf{p}}$ and $\overline{\mathbf{p}}$.

The Bellman equations (13) are then modified to account for the uncertainty in the transition probabilities as follows (cf. (4)): for all $i \in S$

$$V(i) = \frac{1}{\alpha+\nu} \max_{u \in U} \Big\{ \min_{\mathbf{p}_i(u) \in \Omega_i(u)} \Big[g(i,u) + \nu \sum_{j \in S} p_{ij}(u) V(j) \Big] \Big\}. \tag{16}$$

Note that since the per-stage revenue $g(i,u)$ is also a function of uncertain λ_0, it is included in the inner minimization. It can be seen that the Bellman equations (16) become

$$\begin{aligned} V_F^*(0) &= \frac{1}{\alpha+\nu} \max_{0 \le u \le u_{\max}} \Big\{ \min_{\lambda_0 \in [\overline{\lambda}_0 - \hat{\lambda}_0, \overline{\lambda}_0 + \hat{\lambda}_0]} \Big[(\lambda_0 - \lambda_1 u)u \\ &\quad + (\lambda_0 - \lambda_1 u) V_F^*(1) + \big(\nu - (\lambda_0 - \lambda_1 u)\big) V_F^*(0) \Big] \Big\}, \\ V_F^*(i) &= \frac{1}{\alpha+\nu} \max_{0 \le u \le u_{\max}} \Big\{ \min_{\lambda_0 \in [\overline{\lambda}_0 - \hat{\lambda}_0, \overline{\lambda}_0 + \hat{\lambda}_0]} \Big[(\lambda_0 - \lambda_1 u)u \\ &\quad + (\lambda_0 - \lambda_1 u) V_F^*(i+1) + i\mu V_F^*(i-1) \\ &\quad + \big(\nu - (\lambda_0 - \lambda_1 u) - i\mu\big) V_F^*(i) \Big] \Big\}, \ 1 \le i \le K-1, \\ V_F^*(K) &= \frac{1}{\alpha+\nu} \max_{0 \le u \le u_{\max}} \Big\{ \min_{\lambda_0 \in [\overline{\lambda}_0 - \hat{\lambda}_0, \overline{\lambda}_0 + \hat{\lambda}_0]} \Big[i\mu V_F^*(K-1) \\ &\quad + (\nu - i\mu) V_F^*(K) \Big] \Big\} \\ &= \frac{1}{\alpha+\nu} \Big\{ i\mu V_F^*(K-1) + (\nu - i\mu) V_F^*(K) \Big\}. \end{aligned}$$

To construct the robust MDP, we introduce a parameter $0 < \beta < 1$, which reduces the range of λ_0 to $[\overline{\lambda}_0 - \beta\hat{\lambda}_0, \overline{\lambda}_0 + \beta\hat{\lambda}_0]$. Consequently, $\mathbf{p}_i(u)$ belongs to the set $\mathscr{R}_i(u) = \big\{\mathbf{p}(\lambda_0) \mid \lambda_0 \in [\overline{\lambda}_0 - \beta\hat{\lambda}_0, \overline{\lambda}_0 + \beta\hat{\lambda}_0]\big\}$, where $p_j(\lambda_0)$ is defined as before. The Bellman equations for the robust MDP are then given by (cf. (6) and (16))

$$\begin{aligned} V_R^*(0) &= \frac{1}{\alpha+\nu} \max_{0 \le u \le u_{\max}} \Big\{ \min_{\lambda_0 \in [\overline{\lambda}_0 - \beta\hat{\lambda}_0, \overline{\lambda}_0 + \beta\hat{\lambda}_0]} \Big[(\lambda_0 - \lambda_1 u)u \\ &\quad + (\lambda_0 - \lambda_1 u) V_R^*(1) + \big(\nu - (\lambda_0 - \lambda_1 u)\big) V_R^*(0) \Big] \Big\}, \\ V_R^*(i) &= \frac{1}{\alpha+\nu} \max_{0 \le u \le u_{\max}} \Big\{ \min_{\lambda_0 \in [\overline{\lambda}_0 - \beta\hat{\lambda}_0, \overline{\lambda}_0 + \beta\hat{\lambda}_0]} \Big[(\lambda_0 - \lambda_1 u)u \\ &\quad + (\lambda_0 - \lambda_1 u) V_R^*(i+1) + i\mu V_R^*(i-1) \\ &\quad + \big(\nu - (\lambda_0 - \lambda_1 u) - i\mu\big) V_R^*(i) \Big] \Big\}, \ 1 \le i \le K-1, \\ V_R^*(K) &= \frac{1}{\alpha+\nu} \max_{0 \le u \le u_{\max}} \Big\{ \min_{\lambda_0 \in [\overline{\lambda}_0 - \beta\hat{\lambda}_0, \overline{\lambda}_0 + \beta\hat{\lambda}_0]} \Big[i\mu V_R^*(K-1) \\ &\quad + (\nu - i\mu) V_R^*(K) \Big] \Big\} \\ &= \frac{1}{\alpha+\nu} \Big\{ i\mu V_R^*(K-1) + (\nu - i\mu) V_R^*(K) \Big\}. \end{aligned}$$

For computational tests, we set $R = 30$, $r = 2$, $\mu = 1$, $\lambda_0 \in [60 - 10, 60 + 10]$ (i.e., $\overline{\lambda}_0 = 60$ and $\hat{\lambda}_0 = 10$), $\lambda_1 = 5$, $\alpha = 0.9$, and $\beta = 0.5$. We also set $u_{\max} = (\overline{\lambda}_0 + \hat{\lambda}_0)/\lambda_1 = 14$. We discretize the continuous control space $U = [0, u_{\max}]$ into 50 controls. We solve the nominal, fat, robust MDPs by value iteration.

Table 1 shows the maximum long-term revenues for each problem, and the optimal fee policy for each problem is shown in Table 2. When there are 15 calls in the system (i.e., when the system is full), the optimal fee can be set to any value because any new arriving calls are rejected and do not contribute to the total revenue. (As shown above, the Bellman equations for $V_N^*(15)$, $V_F^*(15)$, and $V_R^*(15)$ do not involve the maximization over $0 \le u \le u_{\max}$.)

Table 1. Maximum long-term revenues

i	$V_N^*(i)$	$V_R^*(i)$	$V_F^*(i)$
0	157.28	137.07	117.48
1	155.92	136.01	116.68
2	154.47	134.86	115.81
3	152.90	133.62	114.87
4	151.21	132.28	113.84
5	149.39	130.82	112.71
6	147.40	129.22	111.47
7	145.23	127.46	110.09
8	142.85	125.51	108.55
9	140.21	123.34	106.81
10	137.26	120.90	104.84
11	133.94	118.11	102.57
12	130.13	114.89	99.91
13	125.70	111.09	96.73
14	120.36	106.46	92.77
15	113.55	100.43	87.52

Table 2. Optimal fees

i	$\pi_N(i)$	$\pi_R(i)$	$\pi_F(i)$
0	6.571	6.000	5.429
1	6.857	6.000	5.429
2	6.857	6.000	5.429
3	6.857	6.286	5.429
4	6.857	6.286	5.429
5	6.857	6.286	5.714
6	7.143	6.286	5.714
7	7.143	6.571	5.714
8	7.429	6.571	6.000
9	7.429	6.857	6.000
10	7.714	6.857	6.000
11	8.000	7.143	6.286
12	8.286	7.429	6.571
13	8.571	7.714	6.857
14	9.429	8.571	7.714
15	-	-	-

Note that there is a one-to-one correspondence between the value of λ_0 and ω in this example. In other words, once λ_0 is known, all the transition probabilities are fixed (although they are still functions of u). Hence in terms of notation, we can use λ_0 in place of ω. We are now interested in an empirical estimate of $P\big[V^*_{\lambda_0}(0) \geq V^*_R(0)\big]$, i.e., the probability that the maximum long-term revenue for a randomly chosen λ_0 when no calls in the system initially is at least $V^*_R(0)$. To that end, we consider the probability

$$\begin{aligned} &P\big[V^{\pi_R}_{\lambda_0}(0) < V^*_R(0)\big] \\ &\quad = P\big[V^{\pi_R}_{\lambda_0}(0) < V^*_R(0) \mid \lambda_0 \notin [\overline{\lambda}_0 - \beta\hat{\lambda}_0,\, \overline{\lambda}_0 + \beta\hat{\lambda}_0]\big] \\ &\qquad \times P\big[\lambda_0 \notin [\overline{\lambda}_0 - \beta\hat{\lambda}_0,\, \overline{\lambda}_0 + \beta\hat{\lambda}_0]\big]. \end{aligned}$$

(See (8) and the discussion preceding it. The inequality sign is reversed because the example is a maximization problem.) We uniformly generated 1000 instances of λ_0 from $[\overline{\lambda}_0 - \hat{\lambda}_0, \overline{\lambda}_0 + \hat{\lambda}_0]$. Among them, 500 instances satisfied $\lambda_0 \notin [\overline{\lambda}_0 - \beta\hat{\lambda}_0,\, \overline{\lambda}_0 + \beta\hat{\lambda}_0]$. Out of those 500 instances of λ_0, 247 instances resulted in $V^{\pi_R}_{\lambda_0}(0) < V^*_R(0)$. Hence empirically $P\big[V^{\pi_R}_{\lambda_0}(0) < V^*_R(0)\big] = 0.247$. We thus conclude that with probability about 0.753, $V^*_R(0)$ can serve as a lower bound on an unknown $V^*_{\lambda_0}(0)$ (cf. the discussion below (12)). In fact this “probability of confidence” would be higher than 0.753 because even if $V^{\pi_R}_{\lambda_0}(0) < V^*_R(0)$, it is possible that $V^*_{\lambda_0}(0) \geq V^*_R(0)$.

Setting β to a larger value will increase $P\big[V^*_{\lambda_0}(0) \geq V^*_R(0)\big]$ by decreasing the value of $V^*_R(0)$. By taking this trade-off into consideration, one can adjust the value of β to obtain an appropriately robust MDP. In Figure 1, the empirical estimates of $P\big[V^*_{\lambda_0}(0) \geq V^*_R(0)\big]$ are plotted for various values for β. The solid line in the middle is obtained when λ_0 is uniformly distributed in $[\overline{\lambda}_0 - \hat{\lambda}_0, \overline{\lambda}_0 + \hat{\lambda}_0]$, which we assumed earlier. When the probability distribution of λ_0 is triangle[2] and reverse-triangle[3], the dashed line in the top and the dash-dotted line in the bottom represent the empirical estimates of $P\big[V^*_{\lambda_0}(0) \geq V^*_R(0)\big]$, respectively. As the figure shows, for a given value of β, the probability of confidence (i.e., $P\big[V^*_{\lambda_0}(0) \geq V^*_R(0)\big]$) increases as the realizations of λ_0 tend to be close to its mean value $\overline{\lambda}_0$ (in other words, as the variance of λ_0 decreases).

5. CONCLUSION

The transition probabilities of an infinite horizon discounted cost MDP can be uncertain due to estimation errors. If the estimation errors are not negligible, the uncertainty (or ambiguity) in the transition probabilities should be factored in when solving the MDP. In this case, a policy that performs best in the worst-case scenario has been sought in the literature. We have considered a robust formulation for the MDP to obtain a less conservative policy than the one from the worst-case approach. By comparing the optimal cost of the robust formulation with that of a random instance of the MDP, we have characterized the performance of the robust formulation probabilistically. Extending the results of this paper to an infinite horizon average cost MDP and to an approximate MDP will be considered in the future.

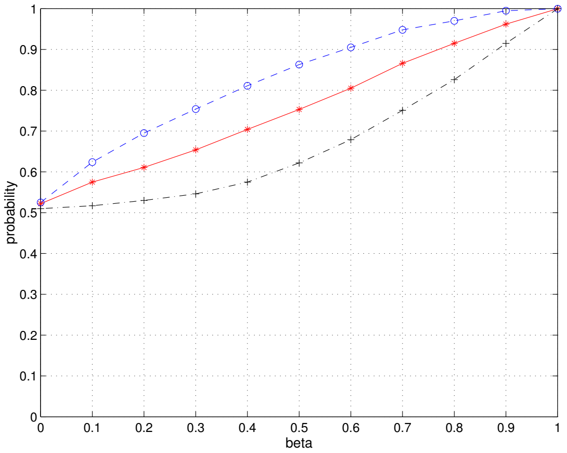

Fig. 1. Empirical estimates of $P\big[V^*_{\lambda_0}(0) \geq V^*_R(0)\big]$

REFERENCES

A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, Belmont, 3rd edition, 2005.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, 3rd edition, 2007.

D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

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

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

I. Ch. Paschalidis and S.-C. Kang. Robust linear optimization: On the benefits of distributional information and applications in inventory control. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 4416–4421, Seville, Spain, 2005.

I. Ch. Paschalidis and S.-C. Kang. On the benefits of distributional information in robust linear optimization. In *Proceedings of the 5th IFAC Symposium on Robust Control Design*, Toulouse, France, 2006.

I. Ch. Paschalidis and J. N. Tsitsiklis. Congestion–dependent pricing of network services. *IEEE/ACM Transactions on Networking*, 8(2):171–184, 2000.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, 1994.

J. K. Satia and R. E. Lave. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

C. C. White and H. K. Eldeib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, 1994.

[2] The density function of the triangle distribution is

$$f_{\lambda_0}(\lambda) = \begin{cases} \lambda/\hat{\lambda}_0^2 - \left(\overline{\lambda}_0 - \hat{\lambda}_0\right)/\hat{\lambda}_0^2 & \text{if } \overline{\lambda}_0 - \hat{\lambda}_0 \leq \lambda \leq \overline{\lambda}_0, \\ -\lambda/\hat{\lambda}_0^2 + \left(\overline{\lambda}_0 + \hat{\lambda}_0\right)/\hat{\lambda}_0^2 & \text{if } \overline{\lambda}_0 \leq \lambda \leq \overline{\lambda}_0 + \hat{\lambda}_0. \end{cases}$$

[3] The density function of the reverse-triangle distribution is

$$f_{\lambda_0}(\lambda) = \begin{cases} -\lambda/\hat{\lambda}_0^2 + \overline{\lambda}_0/\hat{\lambda}_0^2 & \text{if } \overline{\lambda}_0 - \hat{\lambda}_0 \leq \lambda \leq \overline{\lambda}_0, \\ \lambda/\hat{\lambda}_0^2 - \overline{\lambda}_0/\hat{\lambda}_0^2 & \text{if } \overline{\lambda}_0 \leq \lambda \leq \overline{\lambda}_0 + \hat{\lambda}_0. \end{cases}$$