A Robust Constrained Markov Decision Process Model for Admission Control in a Single Server Queue

Erim Kardes
Department of Industrial Engineering
Ozyegin University
Istanbul, Turkey

Abstract

This paper presents a robust optimization approach for discounted constrained Markov decision processes with payoff uncertainty. It is assumed that the decision-maker has no distributional information on the unknown payoffs. Two types of uncertainty sets, convex hulls and intervals are considered. Interval uncertainty sets are parametrized allowing a subset of the payoffs to vary within intervals, to reduce the conservativeness of robust policies. We provide formulations to calculate robust optimal policies and present an application to admission control in a discrete single server queueing problem. It is numerically shown that when the queue manager cannot observe the exact rewards from admissions, mixed threshold type of policies are not necessarily robust optimal.

Keywords
Constrained MDP, robust optimization, markov decision process

1. Introduction

Consider a queueing system where a queue manager (admission controller) controls the arrival rate of the customers. In general, the controller may not know the exact reward associated with the admission of a customer. In case the manager cannot observe the exact rewards, what type of policies could she adopt to keep the rewards generated in the system above a certain threshold? Another example is from telecommunications: A particular problem encountered in telecommunication systems is that of minimizing the delay of arrivals subject to lower-bound constraints on rewards obtained from admissions. This research is motivated by the fact that rewards and/or even delay costs may not be known exactly, especially if they are to be estimated from data. Furthermore, given threshold constraints on rewards or delays, small enough changes in these parameters could result in sub-optimal and/or infeasible policies, if optimal policies are computed by disregarding possible changes in such parameters. In this paper, we present a method, namely a robust optimization approach, that could be used to cope with incomplete information on rewards associated with the admission of customers.

Dynamic decision problems taking into account multiple performance measures are abun- dant in the literature. In particular, telecommunications literature is ripe with models that capture the transmission of different types of information at the same time, such as messages, file transfers, voice and etc. Hence, in such models, a controller deals with multiple performance measures and their trade-offs. Constrained Markov decision processes (CMDPs) with no payoff uncertainty (exact payoffs) have been used extensively in the literature to model sequential decision making problems where such trade-offs exist. Unlike in a classical finite state/action Markov decision process, a decision-maker in a CMDP receives more than one type of pay- off throughout the process and hence is concerned with optimizing an objective function while keeping other objectives below certain thresholds. In this research, we are motivated by the two examples given in the preceding paragraph. Such problems are typically solved using CMDPs. Therefore, this paper first addresses payoff uncertainty in a general discrete-time finite state finite action CMDP. Then, the new methodology is applied in a discrete-time admission control problem with uncertain rewards, where delay is minimized subject to a lower-bound constraint on rewards generated in the system. We note that the general approach developed here is by no means only applicable to problems arising in telecommunications. CMDPs are used in other applications as well and hence a robust optimization approach to payoff uncertainty in CMDPs could be useful in other suitable contexts, some of which are introduced in our literature review.

1.1 Literature Review

Altman (1999) provides a comprehensive text on CMDPs. Chapter 5 of Altman (1999) presents an application where, an admission and a service controller determine the probability of an ar- rival and service completion, respectively, at each time period. The model includes holding costs, costs on service, and rewards associated with arrivals. A trade-off exists between achiev- ing high throughput and low expected delays. Although there are indeed two controllers in the system, their objectives coincide and hence, the problem is modeled as if there is a single controller that minimizes the delay subject to constraints on throughputs and costs of service. The application in this paper differs from that model given by Altman (1999) by taking into account the uncertainty on rewards via robust optimization.
Robust optimization is used to address uncertainty in Markov decision processes (MDPs) with no constraints. Nilim and El Ghaoui (2005), and Iyengar (2005) adopt a robust optimiza- tion approach in MDPs with uncertain transition matrices, presenting independent proofs for the robust value iteration. Kardeş et al. (2011) introduce a robust approach to data uncer- tainty in discounted competitive MDPs. A motivation to consider robust optimization in these articles is to be able to cope with the sensitivity of optimal policies to perturbations in the data. The difference of the current paper from these articles is that it considers constrained MDPs with payoff uncertainty. Two reasons to consider payoff uncertainty in CMDPs are as follows: First, motivated by the possibility of having incomplete information on admission rewards in a queueing system, we consider only the payoff uncertainty in this paper. Second, the addition of extra constraints in an MDP impedes us from using value iteration arguments in a standard way, which are typical tools in solving unconstrained MDPs. Consequently, the analysis of CMDPs with uncertain transition matrices become even less tractable. Therefore, for mathe- matical tractability, we only consider payoff uncertainty in this paper, and note that addressing transition uncertainty in CMDPs via robust optimization methods is an open problem.

Several articles have appeared in the literature recently that study incomplete information on rewards, or in other words, imprecise rewards. McMahan et al. (2003) considers imprecise rewards in an unconstrained MDP with a controller who assumes that given his policy fixed, an adversary can choose a reward function from a discrete finite set. The adversary is allowed to use a mixed strategy, that is, a probability distribution over his finite set of cost function alternatives. The model then becomes a variant of a zero-sum stochastic game. Authors develop an efficient algorithm to solve the model presented, and apply it to a robot path planning problem. The main difference of our work from that article is that we consider constrained MDPs with imprecise rewards. Furthermore, given a fixed policy of the controller in our model, one could assume that there is an adversary who chooses the worst-case rewards for the constraints. However, the alternative set of the adversary in our model is not finite. Rather, we investigate two uncertainty sets for rewards: Convex hulls and intervals with a parameter that controls the conservativeness of the robust solution.

Regan and Boutilier (2010) provide an approach for unconstrained MDPs with imprecise rewards. That work is fundamentally different from ours in that the authors adopt a minmax regret criterion In summary, the controller minimizes the maximum regret where the maximum regret is both with respect to other possible policies that could have been adopted and to imprecise rewards. The authors argue that this approach is adopted because a typical robust optimization approach maximizes the minimum (worst-case) of all rewards (a maxmin approach), which results in conservative solutions. In the current paper, we do adopt the maxmin approach of robust optimization. However, to alleviate the conservativeness, we consider interval uncertainty sets. For interval sets, we consider that not all reward parameters are subject to change but only a subset of them. Applying an approach first introduced by Bertsimas and Sim (2004) to constrained MDPs, we are able to adjust the conservativeness of the policies using a budget for the uncertainty. We note that convex hull uncertainty sets have been used before in other contexts, for example in vehicle routing (see Sungur et al., 2008). Regan and Boutilier (2011) provide a more efficient approach to solving MDPs with imprecise rewards involving the minmax regret criterion.

Xu and Mannor (2012) investigate distributionally robust unconstrained MDPs when distri- butional information is available for both reward and transition data. Here, the distributional robustness means that the optimal strategy maximizes the expected total reward under the most adversarial admissible probability distributions. Based on parametric linear programming, Xu and Mannor (2006) consider the tradeoff between the nominal and worst case performances in unconstrained MDPs when there is payoff uncertainty. Uncertainty sets considered are state- wise independent and the set for each state is taken to be a polytope. The main differences of the current work from that paper is that, rather than

unconstrained MDPs, we treat con- strained MDPs with convex hull and interval uncertainty sets. The purpose of Xu and Mannor (2006) is to model explicitly the tradeoff between robustness and performance using a weighted sum of the performance and robustness objectives as the minimizing objective. Instead, in the current paper, we are interested in providing explicit robust formulations for constrained MDPs. Xu and Mannor (2009) consider unconstrained MDPs with uncertainty on the reward param- eters. Authors consider two related problems: minimax regret and mean-variance tradeoff of the regret.

Zadorojniy and Shwartz (2006) study a certain type of robustness of the optimal cost and the policy under changes in the constraints. Authors' definition of robustness is that a small change in some parameters requires a small change in the policy. They develop a new technique to characterize and establish this type of robustness with respect to changes in the values of the constraints. On the other hand, robustness in our work is defined as a solution that gives the best possible guaranteed value under all possible realizations of constraints. Therefore, the two lines of research are fundamentally different. Altman and Shwartz (1991) investigate the sensitivity of the optimal cost and policy in CMDPs to changes in various parameters.

Next, a sample of articles is presented that use constrained MDPs with fixed exact rewards. The results of the current paper could be used in such types of models and their variants when payoffs are not exactly known and can be represented by convex hull and/or interval uncertainty sets. Kuhn and Madanat (2005) present an infrastructure management problem that involves transition data uncertainty and contrasts the expected costs incurred when the uncertainty is ignored with those incurred when it is considered using robust optimization. The model we present in this paper can be applied to the model given by Kuhn and Madanat (2005) if an uncertainty is also assumed on agency costs for a facility in a given state with a given action being applied. Optimal pavement maintenance policies are investigated by Golabi et al. (1982) for a 7400-mile network of Arizona freeways using CMDPs, resulting in a reported 14 million dollar savings in the first year of implementation. Wessels (1980) presents an application on credit institutions' daily decisions as to how much and in what form cash should be hold. The model includes loss of interest costs, replenishment costs, and stock reduction costs. The objective is to minimize the infinite horizon expected costs subject to a constraint on cash shortages. A CMDP is presented by Kolesar (1970) which, among several other related applications, optimizes long-run bed occupancy subject to constraints on the probability of a hospital being full at any time. Rothstein (1971) presents an application where decisions have to be made each day prior to a target day as to whether airline bookings for the target day should be accepted subject to a constraint on the expected number of bookings. Rothstein (1974) presents a similar application to hotel bookings. Turgeon (1985) studies power generation as the objective. Several reservoirs exist and the decision is how much power should be generated from each, subject to constraints on the frequency with which critical reservoir levels are exceeded. Sniedovich (1979) presents an application on the amount of water to be released optimally from a reservoir at any time, subject to a reliability constraint.

2. The Model

This section sets up the background on constrained MDPs with payoff uncertainty and thus introduces a new approach. A finite state/action discrete-time constrained Markov decision process is considered. S denotes the finite set of states and A the finite set of actions. When the system is in state s, a controller chooses an action a and receives immediate payoffs c(s, a) and $\tilde{d}_k(s, a)$, $k = 1, ..., K$. Here, c(s, a) is associated with the objective function considered by the controller and $\tilde{d}_k(s, a)$ is the uncertain payoff related to the k^{th} constraint. P_{sak} denotes the probability of moving from state s to state k when the action a is chosen. If the payoffs to be received from future states are discounted, $\alpha \in (0, 1)$ represents the discount factor. β denotes a distribution over the initial states.A history h_t at time t is a sequence of previous states, actions, and the current state. A policy is denoted by $x = (x_t), t = 0, 1, ...$ If the history h_t is observed at time t, then the controller chooses an action a with probability $x_t(a|h_t)$. A stationary policy is only a function of the current state and does not depend on time t. The controller can restrict its policies to stationary policies without degrading the overall objective function value he would have obtained by using history dependent policies. A policy henceforth means stationary policy in this paper. An initial distribution β and a given policy x determine a probability distribution over the space of state-action trajectories. E_β denotes the corresponding expectation.

Let S_t and A_t denote the state and action at time t, $t = 0, 1, \ldots$. Given values of $\tilde{d}_k(s, a)$ for any state-action pair, the value of the k^{th} constraint could be calculated using the expression:

$$(1-\alpha)\sum_{t=0}^{\infty}\alpha^t E_\beta^x \tilde{d}_k(S_t, A_t),$$

where $(1-\alpha)$ is a normalization factor. Let us denote the vector of payoffs associated with the k^{th} constraint by $\tilde{d}_k$. In this paper, it is assumed that $\tilde{d}_k$ is not fixed and that it is known to belong to a given uncertainty set U_k. The uncertainty considered in $\tilde{d}_k$ is of a stationary type. Note that the same uncertainty model is also used in other models (see Delage and Mannor, 2010, Nilim and El Ghaoui, 2005). In other words, the uncertainty considered here represents modeling uncertainty and is attractive for statistical reasons because it is time-invariant.

We are interested in minimizing the total expected discounted payoffs subject to constraints that must be satisfied no matter what the actual realization of $\tilde{d}_k \in U_k$ is. Note that the same notion of feasibility is also studied in the robust control literature. Formally, the controller's robust optimization problem is the following:

$$\min_{x\in X} \quad (1-\alpha)\sum_{t=1}^{\infty}\alpha^{t-1}E_\beta^x c(S_t, A_t)$$

$$\text{s.t.} \quad (1-\alpha)\sum_{t=1}^{\infty}\alpha^{t-1}E_\beta^x \tilde{d}_k(S_t, A_t) \le v_k, \quad \forall \tilde{d}_k \in U_k$$

Based on occupation measures approach presented by Altman (1999), the above problem can be stated as Problem 1, as follows: Problem 1:

$$\min_{\rho} \quad \mathbf{c}^T\boldsymbol{\rho}$$

$$\text{s.t.} \quad \tilde{D}\boldsymbol{\rho} \le \mathbf{v}, \quad \forall \tilde{D} \in U$$

$$\boldsymbol{\rho} \in Q$$

where Q is the set of occupation measures, $\tilde{D} \in U$, v is a given vector. For the discounted performance criterion, Q is given by the following set (see Altman, 1999):

$$\sum_{s\in S}\sum_{a\in A}\rho(s,a)(I_k(s) - \alpha P_{sak}) = (1-\alpha)\beta(k), \quad \forall k \in S$$

$$\rho(s,a) \ge 0 \forall s, a,$$

where $I_k(s) = 1$ if $k = s$, 0 otherwise. Row k of the matrix $\tilde{D}$, denoted by $\tilde{d}_k$ above, contains the uncertain costs of each state-action pair for constraint k of the MDP. That is, $\tilde{d}_k$ is a vector that contains the costs $\tilde{d}_k(s, a)$ associated with constraint k and state-action pair (s, a).

3. Characterization of Robust Optimal Policies

3.1 Interval Uncertainty Sets

In this section, we introduce scaled deviations for cost parameters. Our approach closely follows the approach in Bertsimas and Sim (2004), applied in the context of constrained MDPs. We assume $\tilde{d}_k(s, a)$ could deviate within the interval $[\bar{d}_k(s, a) - \hat{d}_k(s, a), \bar{d}_k(s, a) + \hat{d}_k(s, a)]$. Define the scaled deviation of a cost parameter $\tilde{d}_k(s, a)$ from its expected (nominal) value $\bar{d}_k(s, a)$ as $\eta_k(s, a) = (\tilde{d}_k(s, a) - \bar{d}_k(s, a))/\hat{d}_k(s, a)$. Note that $\eta_k(s, a) \in [-1, 1], \forall k, s, a$.

Let

$$U_k = \left\{ \tilde{\mathbf{d}}_k \;\middle|\; \tilde{d}_k(s,a) \in \left[\bar{d}_k(s,a) - \hat{d}_k(s,a) \;,\; \bar{d}_k(s,a) + \hat{d}_k(s,a)\right], \; \sum_{(s,a)} |\eta_k(s,a)| \le \Gamma_k \right\}$$

where Γ_k is the total number of payoffs that the decision-maker assumes inexact. We assume that the set of uncertain parameters in constraint k is $S \times A$, i.e., any state-action pair payoff is subject to change.

The formulation that follows could easily be modified if only a subset of state-action pairs is subject to change. Using this uncertainty, for the constraints of problem (1) to hold, we impose

$$\max_{\tilde{\mathbf{d}}_k} \quad \sum_{(s,a)\in S\times A} \tilde{d}_k(s,a)\rho(s,a) \quad \leq v_k$$
$$\text{s.t.} \qquad\qquad \tilde{\mathbf{d}}_k \in U_k$$

Since occupation measures are always nonnegative, the above maximization is equivalent to the following:

$$\sum_{(s,a)\in SxA} \bar{d}_k(s,a)\rho(s,a) \quad +$$
$$\max_{\{H\cup\{(s',a')\}|H\subseteq S\times A,|H|=\lfloor\Gamma\rfloor,(s',a')\in(S\times A)\setminus H\}} \left\{ \sum_{(s,a)\in H} \hat{d}_k(s,a)\rho(s,a) + (\Gamma_k - \lfloor\Gamma_k\rfloor)\hat{d}_k(s',a')\rho(s',a') \right\}$$

Note that, given ρ^*, the objective of the above max problem equals the objective of the following primal LP:

$$\max_{z_k} \quad \sum_{(s,a)} \hat{d}_k(s,a)\rho^*(s,a)z_k(s,a) \qquad (5)$$
$$\text{s.t.} \qquad \sum_{(s,a)} z_k(s,a) \leq \Gamma_k$$
$$0 \leq z_k(s,a) \leq 1, \quad \forall(s,a)$$

The dual of this primal is:

$$\min_{w,v} \qquad \sum_{(s,a)} y_k(s,a) + w_k\Gamma_k$$
$$\text{s.t.} \quad w_k + y_k(s,a) \geq \hat{d}_k(s,a)\rho^*(s,a) \quad \forall(s,a)$$
$$w_k \geq 0, \;\; y_k(s,a) \geq 0 \quad \forall(s,a), \forall k$$

Therefore, the following is proved:

Proposition 1. In the case of interval uncertainty, weak duality ensures that problem (1) is equivalent to

$$\min_{\rho,w,y} \qquad \boldsymbol{c}^T\boldsymbol{\rho}$$
$$s.t. \quad \sum_{(s,a)} \bar{d}_k(s,a)\rho(s,a) + \sum_{(s,a)} y_k(s,a) + w_k\Gamma_k \leq v_k, \quad \forall k$$
$$w_k + y_k(s,a) \geq \hat{d}_k(s,a)\rho(s,a) \quad \forall(s,a), \;\forall k \qquad (6)$$
$$\boldsymbol{\rho} \in Q$$
$$w_k \geq 0, \;\; y_k(s,a) \geq 0 \quad \forall(s,a), \;\forall k$$

4. Application

This section presents an admission control application with an interval uncertainty model of section 3.2 for the unknown rewards. We consider a discrete-time single-server queue with a fixed buffer size M. State space of the system is X = {0, 1, ..., M}. If the system is not empty, a service completion occurs at the end of a time slot with a fixed probability p. A controller probabilistically admits one customer at a time into the system at the beginning of a time slot. At each state except state M, the controller has two (pure) actions. Not admitting any customers (alternative 1) or admitting customers with probability $\bar{a} < 1$ (alternative 2). Note that the admission controller can use a mixed strategy (randomization) at any state except M, and hence the probability of admission lies within $[0, \bar{a}]$.No admission is allowed in state M, hence only the first alternative is allowed in that state. If admission probability in a state is a, then the transition probabilities are as follows:

$$P_{xay} = \begin{cases} (1-a)p, & 1 \le x \le M, \quad y = x-1 \\ a(1-p), & 0 \le x \le M-1, \quad y = x+1 \\ ap + (1-a)(1-p) & 1 \le x \le M-1, \quad y = x \\ 1-p, & x = M, \quad y = M \\ 1-a, & x = 0, \quad y = 0 \end{cases}$$

A fixed holding cost c(x) is incurred when the state is x. Holding cost in the empty state is 0. Little's law states that the expected waiting time is proportional to the expected queue length. Hence the cost c is also related to the delay in the system. It is natural to assume that c is convex and increasing. We assume that the expected rewards are not exactly known to the controller and lies within an interval for each state. Let the unknown expected reward associated with an admission of a customer be denoted by $\tilde{r}$. That is, $\tilde{r} \in [\bar{r}-\hat{r}, \bar{r}+\hat{r}]$. Following the notation of sections 2 and 3, we take $\bar{d}(x, 2) = \bar{r}$ for $x < M$, $\bar{d}(x, 1) = 0$ for $x \in X$, and $\hat{d}(x, 2) = \hat{r} > 0$ for $x < M$, $\hat{d}(x, 1) = 0$ for $x \in X$. Note that if the controller chooses his pure strategy $\bar{a}$ in state x, then the expected reward from an admission is $\tilde{r}\bar{a}$.

4.1 Numerical Results

M , the buffer size, is taken to be 100. The initial state is the empty state, that is, $\beta(0) = 1$, $\beta(s) = 0$, $s > 0$. The discount factor α is 0.99. Γ is as indicated in the following figures and denotes the number of uncertain parameters that the controller assumes inexact to protect himself from overly conservative solutions. The mid-point of the expected reward for an admis- sion is taken to be 1, i.e., $\bar{r} = 1$. Γ and the half-length deviations from the expected reward, $\hat{r}$, are as indicated in the following figures. Lowerbound on rewards generated in the system is v = 0.8. Holding cost function is c(x) = x. p = 0.5, $\bar{a} = 0.8$. Nominal solution refers to a solution that disregards uncertainty, i.e., the solution obtained when $\hat{r} = 0$, $\Gamma = 0$.

In the following, the subscript k is dropped as we have a single constraint on rewards. For the above parameters, Figure 1 illustrates the percentage gain in rewards from using robust policies as opposed to nominal policies under worst-case realization of the data with respect to nominal policies. Percentage increase values in delay from using robust policies are also depicted. In Figure 1, "d" represents percentage increase in rewards generated in the system by using robust optimal policies as opposed to using nominal optimal policies, given that the data is realized in a worst-case manner with respect to the nominal optimal policies. It is depicted for different values of Γ and $\hat{d}$. x-axis represents Γ and numbers 0.1, 0.3, 0.6 represent the $\hat{d}$ value used. Similarly, "c" represents the percentage increase in delay when robust optimal policies are used as opposed to the nominal optimal policies.

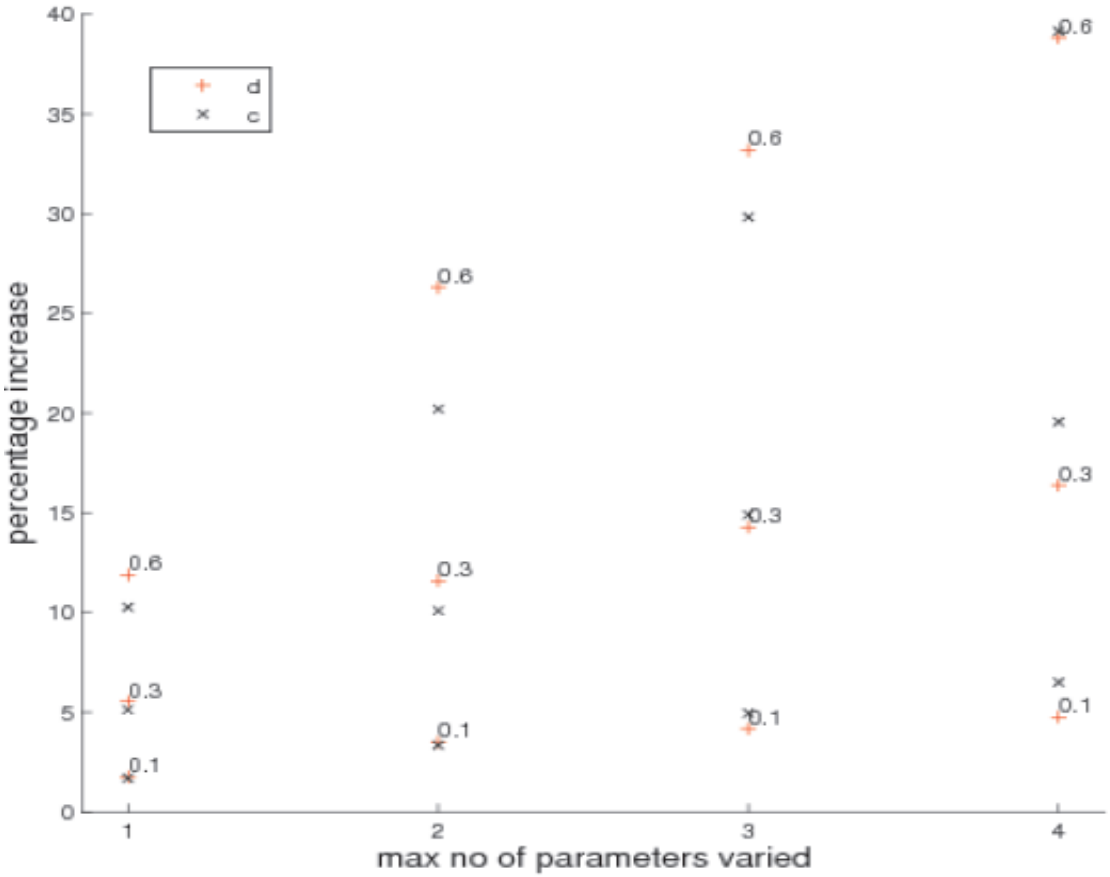

Figure 1: Max no of parameters varied vs percentage increase in reward and delay

Note that, by using robust optimal policies, the percentage increase in rewards obtained are higher than the percentage increase in delay, when the controller protects himself by considering parameters to vary in up to three states and when the possible changes in the parameters are relatively large. For example, when two parameters are considered, a half length of 0.3 and 0.6 in the reward uncertainty results in higher percentage increases in overall total reward than in delay (approximately 11.6% and 10.1%, respectively). The reason for the picture is as follows: In this example, a linear holding cost with no uncertainty is used whereas, naturally, the uncertainty set on rewards is the same in all states. The linear holding cost function is increasing over the states with a 45-degree slope, whereas rewards are the same over all states. To account for the uncertainty in rewards and to achieve the threshold on rewards, robust optimal policies in general prescribe admission of customers in higher states where the nominal policies would normally not admit any customers. Therefore, robust optimal policies provide protection against possible variations in rewards obtained from customers and the price for this protection is an increase in delay costs. Note that if the slope of the holding cost function were lower, one could expect the relative percentage in increase in rewards to outweigh the increase in delay when Γ is larger. We note that even with a 45-degree slope in the holding cost function, we are able to see a relative benefit by using robust optimal solutions, when the uncertainty on rewards is large and when the controller protects himself from the uncertainty by assuming that rewards could vary in a relatively small number of states.

4.2 Threshold Policies

For the nominal model, it is known that mixed threshold policies are optimal (see Altman, 1999, chapter 5). Our numerical experience with the optimal solution of Problem (6) returned by CPLEX suggests that the solution is not necessarily of a mixed threshold type. This brings up the question of whether mixed threshold type of policies are robust optimal at all. In this section, we perform numerical analyses to show that this is not

necessarily the case. The following is the definition of mixed threshold policies, with the notation tailored to our model, where $x(\bar{a} \mid s)$ denotes the probability that policy "x" assigns to alternative $\bar{a}$ in state "s".

Definition. A policy x is a mixed threshold policy for some threshold L if

$$x(\bar{a}|s) = \begin{cases} 1 & \text{if} \quad s < L \\ q & \text{if} \quad s = L. \\ 0 & \text{if} \quad s > L \end{cases}$$

A controller who employs this policy admits customers if the number in the system is less than L, while he does so with probability q if there are L customers, and does not admit if the number in the system exceeds L. Note that holding costs are increasing over the states and reward from an admission is the same over all states. Therefore, one could argue to admit customers until the constraint on the overall reward is first satisfied and expect simultaneously to achieve the minimum holding cost. Following this idea, we implemented a procedure, the sketch of which is given in Algorithm 1. R(0) below represents the reward achieved when the process starts in state 0; similarly C(0) is the holding cost in state 0. v is the lower bound value for the constraint on total accumulated discounted rewards, and ε is some tolerance.

Algorithm 1:Fix an initial pure threshold policy x.

While $\clubsuit R(0) - v_J > \varepsilon$ **do**

Find R(0) and C(0)

If R(0) > v then

decrease the probability of admission in the last state that admits with a positive probability; adjust x accordingly ; compute the occupation measures;
assign Γ of the states with the highest occupation measures the worst case reward parameters (i.e. $\bar{d} - \hat{d}$).

Else

increase the probability of admission in the state where no admission is chosen; adjust x accordingly; compute occupation measures; assign Γ of the states with the highest occupation measures the worst case reward parameters (i.e. $\bar{d} - \hat{d}$).

End

In this procedure, for every fixed policy $\textbf{x}$, occupation measures are calculated. Following the robust optimization paradigm, worst-case reward parameters are assigned to those states with highest occupation measure values. That is, Γ of the states with the highest occupation measures are assigned the worst case reward parameters $\bar{d} - \hat{d}$, for the second ($\bar{a}$) alternative of the controller. Note that given a discount factor α and a policy $\textbf{x}$, the discounted overall total reward in state s, denoted by $R(s)$, can be calculated by solving the following linear set of equalities

$$R(s) = \sum_{a \in A} x(a|s) d(s,a) + \alpha \sum_{k=0}^{M} x(a|s) P_{sak} R(k), \quad s = 1, ..., M,$$

where P_{sak} is the transition probability from state s to k under alternative a. Similarly, one could obtain the discounted overall total holding cost C(s) in state s. Table 1 depicts the optimal holding costs given by the above procedure and those returned by CPLEX after solving Problem (6).

Table 1: Percentage increase in holding costs by using mixed threshold policies

Γ	$\hat{d}$	holding cost (HC) by Algorithm 1	robust optimal HC	% difference
1	0.1	16.32	14.92	9.37
1	0.3	18.78	15.42	21.75
1	0.6	21.24	16.17	31.29
2	0.1	16.74	15.16	10.38
2	0.3	19.89	16.15	23.13
3	0.1	17.03	15.40	10.59
3	0.3	20.26	16.86	20.17

The percentage differences between the two values show that these instances are counterexamples to the claim that there always exist mixed threshold type of policies that are robust optimal. In fact, our numerical experience indicates that, compared with the nominal policies, robust optimal policies typically randomize between admission and no admission in a larger number of states, before prescribing no admission with certainty. In most instances, for a variety of values for the parameters of our model, robust optimal solutions decrease the probability of admission gradually over some number of states, before stopping admissions into the system with probability 1. Intuitively, one could think of such a randomization as a protective measure taken by the controller in the presence of uncertainty on rewards.

5. Conclusions and Future Research

Motivated by an admission control problem where a queue manager cannot observe the exact rewards associated with admissions, this paper introduces robust constrained Markov decision process with payoff uncertainty. For convex hull and interval uncertainty sets, linear optimization formulations that characterize robust optimal policies are given. These formulations can be used in a number of contexts that require the use of constrained Markov decision processes, when payoffs are not exactly known but vary in given interval or convex hull uncertainty sets. In this paper, an application to admission control is presented, where a queue manager admits customers into a single server queue but cannot observe the exact rewards obtained from admissions. A procedure is presented which can be used to compute robust feasible mixed threshold policies. Using this procedure, it is numerically shown that when the queue manager cannot observe the exact rewards from admissions, mixed threshold type of policies are not necessarily robust optimal. A future research direction would be to extend the results in this paper using the average reward performance criterion for CMDPs rather than the discounted one. On the application side, it would be interesting to apply the formulations provided in this paper within the context of bandit models. Project selection can be an interesting application area, where a decision-maker must choose a project to work on at any time from a set of projects. The reward from working on a specific project for one time period may not be known exactly and can vary within some upper and lower bounds. A robust approach can be adopted in such a case. It would also be an interesting extension to investigate how a robust approach performs under risk neutral and risk averse decision maker attitudes.

References

Altman, E., Shwartz, A. Sensitivity of constrained markov decision problems, *Annals of Operations Research*, vol. 32, pp. 1-22, 1991.

Altman, E. *Constrained Markov Decision Processes*. Chapman and Hall. 1999.

Bertsimas, D., Sim, M. The price of robustness. *Operations Research*, vol. 52, 35-53, 2004

Delage, E., Mannor, S. Percentile optimization for markov decision processes with parameter uncertainty. *Operations Research*, Vol. 58, 203-213, 2010.

Golabi, K., Kulkarni, R. B., Way, G. B. A statewide pavement management system. *Interfaces*, Vol. 12, 5-21, 1982.

Iyengar, G. Robust dynamic programming. *Mathematics of Operations Research*, Vol. 30, pp. 1-21, 2005.

Kardes, E., Ordonez, F., Hall, R. W. Discounted robust stochastic games and an application to queueing control. *Operations Research*, Vol. 59, pp. 365-382, 2011.

Kolesar, P. A markovian model for hospital admission scheduling. *Management Science*, Vol. 16., 1970.

Kuhn, K. D., Madanat, S. M. Model uncertainty and the management of a system of infrastructure facilities. *Transportation Research Part C: Emerging Technologies*, Vol. 13, 391-404. 2005.

Nilim, A., El Ghaoui, L.. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, Vol. 53,780-798, 2005.
Regan, K., Boutilier, C. Robust policy computation in reward-uncertain MDPs using nondominated policies. *In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010
Regan, K. Boutilier, C. Robust online optimization of reward-uncertain MDPs. *In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. 2011.
Rothstein, M.. An airline overbooking model. *Transportation Science*, Vol., 180-192. 1971.
Rothstein, M. Hotel overbooking as a markovian sequential decision process. *Decision Sciences*, Vol. 5, 389-404. 1974.
Sniedovich, M. Reliability-constrained reservoir control problems, i: Methodological issues. *Water Resources Research*, Vol. 15. 1979.
Sungur, I., Ordonez, F., Dessouky, M. M. A robust optimization approach for the capacitated vehicle routing problem with demand uncertainty. *IIE Transactions*, Vol. 40, 509-523. 2008.
Turgeon, A. The weekly scheduling of hydroelectric power plants subject to reliability constraints. Technical report. Hydro-Quebec, Canada. 1985.
Wessels, J. Markov decision processes; implementation aspects. Technical report, Memorandum COSOR 80-14, Department of Mathematics, Eindhoven. 1980.
Xu, H., Mannor, S. The robustness-performance tradeoff in markov decision processes. *In Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*. 2006.
Xu, H., Mannor, S. Parametric regret in uncertain markov decision processes. *In Proceedings of the 48th IEEE Conference on Decision and Control.* 2009.
Xu, H., Mannor, S. Distributionally robust markov decision processes. *Mathematics of Operations Research,* Forthcoming. 2012.
Zadorojniy, A., Shwartz, A. Robustness of policies in constrained markov decision processes. *IEEE Transactions on Automatic Control*, Vol 51, 635-638, 2006.

Biography

Erim Kardes received his Ph.D. in Industrial and Systems Engineering from the University of Southern California in 2007. His research interests are in the areas of probability models with optimization applications, Markov decision processes, stochastic games, robust optimization, and their applications in optimization and equilibrium in queues. He works as an assistant professor at Ozyegin University, Department of Industrial Engineering.