OPERATIONS RESEARCH
Vol. 58, No. 1, January–February 2010, pp. 203–213
ISSN 0030-364X | EISSN 1526-5463 | 10 | 5801 | 0203

informs

DOI 10.1287/opre.1080.0685
© 2010 INFORMS

Percentile Optimization for Markov Decision Processes with Parameter Uncertainty

Erick Delage
Department of Management Science, HEC Montréal, Montréal, Quebec H3T 2A7, Canada,
erick.delage@hec.ca

Shie Mannor
Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec H3A 2A7, Canada,
shie.mannor@mcgill.ca

Markov decision processes are an effective tool in modeling decision making in uncertain dynamic environments. Because the parameters of these models typically are estimated from data or learned from experience, it is not surprising that the actual performance of a chosen strategy often differs significantly from the designer's initial expectations due to unavoidable modeling ambiguity. In this paper, we present a set of percentile criteria that are conceptually natural and representative of the trade-off between optimistic and pessimistic views of the question. We study the use of these criteria under different forms of uncertainty for both the rewards and the transitions. Some forms are shown to be efficiently solvable and others highly intractable. In each case, we outline solution concepts that take parametric uncertainty into account in the process of decision making.

Subject classifications: Markov decision processes; parameter uncertainty; finite state; stochastic model applications; stochastic programming; value at risk; chance-constrained optimization.
Area of review: Stochastic Models.
History: Received November 2006; revisions received May 2007, January 2008; accepted April 2008. Published online in *Articles in Advance* August 17, 2009.

1. Introduction

Markov decision processes (MDPs) are an effective tool in modeling decision making in uncertain dynamic environments (e.g., Puterman 1994). Because the parameters of these models typically are either estimated from data or learned from experience, it is not surprising that, in some applications, unavoidable modeling uncertainty often causes the long-term performance of a strategy to differ significantly from the model's predictions (refer to experiments by Mannor et al. 2007). Let us consider a concrete problem in which one needs to deal with inherent model uncertainty. A factory owner wants to design a replacement policy for a line of machines. This problem is known to be well modeled with an MDP, with states representing reachable aging phases and actions describing different repair or replacement alternatives. Although the parameters used in such a model typically can be estimated from historical data (experienced repair costs and decreases in production due to failures), one can rarely fully resolve them. For example, there is inherent uncertainty in future fluctuations for the cost of new equipment. Moreover, one often does not have access to enough historical data to adequately assess the probability of a machine breaking down at a given aging stage. One should expect significant improvements from incorporating this uncertainty in the performance evaluation of a given repair policy. This example illustrates the need for criteria that address parameter uncertainty in general and specifically in the MDPs (e.g., Ben-Tal and Nemirovski 1998, Silver 1963, Martin 1967, Satia and Lave 1973, Dearden et al. 1999).

To date, most efforts have focused on the study of robust MDPs (e.g., Nilim and El Ghaoui 2005, Iyengar 2005, Givan et al. 2000, Bagnell et al. 2001). In this context, under the assumption that parameters lie in a given uncertainty set, one considers a dynamic game against nature as equivalent to choosing the best strategy for the worst-case scenario. Under mild conditions (namely, the convexity of the uncertainty sets), the robust formulation of the problem of parameter uncertainty becomes tractable. Unfortunately, as will be demonstrated in §5, the robust MDP approach often generates overly conservative strategies. A similar conclusion can be drawn in the context of the H_∞ robust control formulation, as in van der Schaft (1996), which considers uncertainty in terms of bounded perturbations in the system. Previous work also studies parameter uncertainty in the form of perturbations of the underlying Markov chain, but it focuses more on understanding the long-term dynamics of the system rather than the performance of policies (see Avrachenkov et al. 2002).

In this paper, we offer a more practical way of handling uncertainty in the parameters. Following recent work by Mannor et al. (2007) that studied the effect of parameter uncertainty on the mean and variance of the value function of Markov processes with fixed policy, we will

INFORMS holds copyright to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

consider the parameters as random variables and study the Bayesian point of view on the question of decision making. In fact, it will be shown that this framework can lead to a performance measure called the *percentile criterion*, which is both conceptually natural and representative of the trade-off between optimistic and pessimistic strategies when facing parameter uncertainty. Unlike robust methods, our approach does not require the assumption that parameters lie in a bounded uncertainty set but instead will attempt to reason directly about the effect of this uncertainty on the total cumulative reward itself. Note that Filar et al. (1995) introduced the percentile criterion as a risk-adjusted performance measure for "average reward" MDPs. However, their study did not address the question of parameter uncertainty.

The chance-constrained criterion that is widely studied for single-period optimization problems (e.g., Charnes and Cooper 1959, Prékopa 1995) will be generalized in §2 to infinite-horizon MDPs. Although general chance constraints are suspected to be "severely computationally intractable" (Nemirovski and Shapiro 2006), this paper will detail the spectrum of computational difficulties related to solving the chance-constrained criterion. In §3, we demonstrate that under the assumption that the transitions are known and the rewards are normally distributed, the chance-constrained MDP can be solved using a deterministic "second-order cone" program (c.f., Lobo et al. 1998), for which a solution can be found in polynomial time. However, we will then show that although the normality assumption on rewards can be softened, there still exist forms of uncertainty for which exact optimization of the percentile criterion is NP-hard. We then address in §4 the question of uncertainty in the transitions of the Markov chain and present an approximation method for finding an optimal policy of the chance-constrained MDP. In §5, we illustrate how this criterion outperforms the nominal and robust criterion on instances of the machine replacement problem with either reward or transitions uncertainty. Section 6 contains concluding remarks.

2. Background

In the context of an MDP with parameter uncertainty, one can either be "careless" and disregard parameter uncertainty during decision making, or be "pessimistic" by planning in order to be protected from worst-case scenario. The purpose of our research is to focus on a "tempered" attitude that will realistically trade between the two conflicting views. Next, we present these three attitudes in mathematical terms.

2.1. The Nominal MDP Problem

We consider an infinite-horizon MDP described as follows: a finite state space S with $|S|$ states, a finite action space A with $|A|$ actions, a transition probability matrix $P \in \mathbb{R}^{|S|\times|A|\times|S|}$ with $P(s,a,s') = \mathbb{P}(s' \mid s,a)$, an initial distribution on states q, and a reward vector $r \in \mathbb{R}^{|S|}$. Although our analysis will strictly consider the case in which the reward depends only on the current state, the results presented in this work can easily be extended to a reward function of the form $r(s,a,s')$. In the context of an infinite-horizon MDP, one can choose to apply a mixed policy π, which is a mapping from the set of states S to the probability simplex over the available actions. For reasons of tractability, we will limit our attention to the set of stationary Markov policies, which is denoted by Υ. When considering an infinite horizon, an optimal discounted reward stationary policy π is the solution to the following optimization problem:

$$\underset{\pi\in\Upsilon}{\text{maximize}}\ \mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t r(x_t)\,\middle|\, x_0 \propto q, \pi\right),$$

where $\alpha \in [0,1)$ is the discount factor. This problem is known to be easily solvable using value iteration (e.g., Bertsekas and Tsitsiklis 1996). However, it does not take into account any uncertainty in the choice of the parameters P and r. In practice, this uncertainty is unavoidable.

In Mannor et al. (2007), the authors address this issue by investigating the effect of random $\tilde{r}$ and $\tilde{P}$ on a new nominal problem:

$$\underset{\pi\in\Upsilon}{\text{maximize}}\ \mathbb{E}_{\tilde{r},\tilde{P}}\left(\mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t \tilde{r}(x_t)\,\middle|\, x_0 \propto q, \pi, \tilde{P}\right)\right).$$

This problem maximizes the expected return over both the trajectories of x and the random variables $\tilde{r}$ and $\tilde{P}$. Because of the nonlinear effect of $\tilde{P}$ on the expected return, the authors argue that evaluating the objective of this problem for a given policy is already difficult. Most importantly, their experiments demonstrate that the common approach consisting of using the most likely (or expected) parameters in the nominal problem leads to a strong bias in the performance of the chosen policy. These results underline the difficulty in handling parameter uncertainty by simply formulating risk-adjusted utility functions, such as in Howard and Matheson (1972). In this paper, we will consider efficient techniques to take the uncertain $\tilde{r}$ and $\tilde{P}$ into account in the decision making.

2.2. The Robust MDP Problem

The most common approach to account for uncertainty in the parameters of an optimization problem is to use robust optimization. This framework assumes that the uncertain parameters are constrained to lie in a given complete set (hopefully convex) and optimize the worst-case scenario over this set. In the case of discounted reward MDP, where the rewards r_t for each time step and the transition matrix P are known to lie in a set R and P, respectively, the robust problem thus becomes

$$\underset{\pi\in\Upsilon}{\text{maximize}}\ \min_{P\in\mathsf{P},\, r_0\in\mathsf{R},\, r_1\in\mathsf{R},\ldots} \mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t r_t(x_t)\,\middle|\, x_0 \propto q, \pi\right). \quad (1)$$

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

There are two types of reward uncertainty that are of interest. In the first type, termed *fixed uncertainty*, the reward vector is drawn once and remains fixed for all time-steps. In the second type, termed *repeated uncertainty*, the reward is independently drawn from the feasible set at each time-step. It is a well-known fact that in both cases, under the assumption of no transition uncertainty, the optimal policy π^* for problem (1) is the same (see Bertsekas and Tsitsiklis 1996) and can be found efficiently. The same is true if one disregards reward uncertainty and wants to solve the robust problem under transition uncertainty (see Nilim and El Ghaoui 2005).

2.3. The Chance-Constrained MDP Problem

Consider a Bayesian setup in which the random reward vector $\tilde{r}$ and random transition matrix $\tilde{P}$ are known to be independent and have joint probability distribution functions $f(\tilde{r})$ and $f(\tilde{P})$, respectively. In such a scenario, unless the distributions are supported over a "small" bounded subset of their domain, formulating problem (1) with $\mathsf{R} = \{r \mid f(r) \neq 0\}$ and $\mathsf{P} = \{P \mid f(P) \neq 0\}$ is no longer pertinent (e.g., if $\tilde{r} \propto \mathcal{N}(\mu_{\tilde{r}}, \Theta_{\tilde{r}})$, then $\mathsf{R} = \mathbb{R}^{|S|}$ and (1) is $-\infty$). Even if the optimization is performed over a restricted bounded subset (e.g., ellipsoids representing a 95% confidence), there is no clear method to select this uncertainty set because the real concern is the level of confidence in the total cumulative reward and not in the individual parameters. Instead, it is more relevant to express the risk-adjusted discounted performance of an uncertain MDP in the following *chance-constrained* form:

$$\underset{y\in\mathbb{R},\,\pi\in\Upsilon}{\text{maximize}} \quad y \tag{2a}$$

$$\text{subject to} \quad \mathbb{P}_{\tilde{r},\tilde{P}}\left(\mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t\tilde{r}_t(x_t)\,\middle|\, x_0\propto q,\,\pi\right)\geqslant y\right) \geqslant 1-\epsilon, \tag{2b}$$

where the probability $\mathbb{P}_{\tilde{r},\tilde{P}}$ is the probability of drawing the reward vector $\tilde{r}_t$ for each time-step independently from $f(\tilde{r})$ and the transition matrix $\tilde{P}$ from $f(\tilde{P})$, and where $\mathbb{E}_x(\cdot \mid x_0 \propto q, \pi)$ is the expectation of the trajectory given a concrete realization of each $\tilde{r}_t$ and $\tilde{P}$, a policy π, and a distribution q for the initial state x_0. For a given policy π, the above chance-constrained problem gives us a $1-\epsilon$ guarantee that π will perform better than y^*, the optimal value of problem (2b), under the distribution of $\tilde{r}$ and $\tilde{P}$. Note that, when $\epsilon = 0$, problem (2b) and problem (1) are equivalent; thus, ϵ measures the risk of the policy doing worse than y^*. The performance measure we use is related to risk-sensitive criteria often used in finance (value-at-risk). However, in finance, one is usually interested in the risk of a single trajectory. We focus on the risk of the expected performance similarly to the robust optimization approach of Givan et al. (2000), Bagnell et al. (2001), and Nilim and El Ghaoui (2005).

Section 3 will focus on uncertainty in the reward parameters. Later, in §4, parameter uncertainty will be addressed. Although we do limit ourselves to presenting the details from a Bayesian point of view to preserve the clarity of our derivations, a frequentist approach to the percentile criterion does follow naturally and is summarized in the print appendix. This work focuses on fixed parameter uncertainty (i.e., uncertainty due to the modeling, although in the system the parameters are actually fixed). Similar methods can be derived for the problem of repeated uncertainty.

2.4. Notation

In the remainder of the paper, the following notation is used. $\mathbf{1}_K$ is the vector of all ones in $\mathbb{R}^K$. For clarity, $Q_{(i,j)}$ will refer to the ith row, jth column term of a matrix Q. Also, for the sake of simpler linear manipulations, we will present a policy π under its matrix form $\Pi \in \mathbb{R}^{|S|\times|S|\times|A|}$, such that $\Pi_{(s_1,s_2,a)} = \pi(s_1,a)\mathbb{1}\{s_1 = s_2\}$, and when this three-dimensional matrix will be multiplied to another matrix $Q \in \mathbb{R}^{|S|\times|A|\times K}$ it will refer to a matrix multiplication carried along $\mathbb{R}^{|S|\times(|S||A|)} \times \mathbb{R}^{(|S||A|)\times K}$, such that $(\Pi Q)_{(i,j)} = \sum_{(k,a)}\Pi_{(i,k,a)}Q_{(k,a,j)}$. Note that this formulation explicitly denotes the linear relation between the decision variable Π and the inferred transition probability P_π, such that $(\Pi P)_{(i,j)} = (P_\pi)_{(i,j)} = \mathbb{P}(s' = j \mid s = i, a = \pi(i))$.

3. Decision Making Under Uncertain Reward Parameters

First, the problem of reward uncertainty is addressed for a common family of distribution functions, the multivariate Gaussian distribution $\tilde{r} \propto \mathcal{N}(\mu_{\tilde{r}}, \Theta_{\tilde{r}})$. Under the assumption of Gaussian rewards, solving the percentile MDP is not considerably harder than solving the nominal MDP. We later briefly describe how the Gaussian reward assumption can be relaxed although there exist distributions over the parameters for which the percentile problem becomes intractable.

3.1. Reward Uncertainty with Gaussian Distribution

The Gaussian assumption is standard in many applications because it allows modeling correlation between the reward obtained in different states. Also, in the Bayesian framework it is common to assume that $\Theta_{\tilde{r}}$ is known and use a Gaussian prior, with parameters (μ_0, Θ_0), over $\mu_{\tilde{r}}$. Then, based on new independent samples $\{r_1, r_2, \ldots, r_m\}$ from the distribution $f(\tilde{r})$, one can obtain an analytical posterior over $\mu_{\tilde{r}}$, which has the same Gaussian shape with parameters (see Gelman et al. 2003 for more details):

$$\mu_1 = \Theta_1\left(\Theta_0^{-1}\mu_0 + \Theta_{\tilde{r}}^{-1}\sum_{i=1}^{m} r_i\right), \quad \Theta_1 = (\Theta_0^{-1} + m\Theta_{\tilde{r}}^{-1})^{-1}.$$

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

Lemma 1 (Theorem 10.4.1 of Prékopa 1995). *Suppose that $\xi \in \mathbb{R}^n$ has a multivariate Gaussian distribution. Then, the set of $x \in \mathbb{R}^n$ vectors satisfying*

$$\mathbb{P}(x^\top \xi \leqslant 0) \geqslant 1-\epsilon$$

is the same as those satisfying

$$x^\top \mu_\xi + \Phi^{-1}(1-\epsilon)\sqrt{x^\top \Theta_\xi x} \leqslant 0,$$

where $\mu_\xi = \mathbb{E}(\xi)$, Θ_ξ is the covariance matrix of the random vector ξ, ϵ is a fixed probability such that $0 \leqslant \epsilon \leqslant 1$, and Φ is the cumulative distribution function of $\mathcal{N}(0,1)$.

Lemma 1 is an important result in the field of stochastic programming. In our specific context, the lemma allows us to show that finding an optimal stationary policy for the problem of maximizing the $(1-\epsilon)$-percentile criterion under Gaussian uncertainty can be expressed as a "second-order cone" program.

Theorem 1. *For any $\epsilon \in (0, 0.5]$, the discounted reward chance-constrained problem with fixed Gaussian uncertainty in the rewards,*

$$\underset{y\in\mathbb{R},\,\pi\in\Upsilon}{\text{maximize}} \quad y \tag{3a}$$

$$\text{subject to} \quad \mathbb{P}_{\tilde r}\left(\mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t \tilde r(x_t)\,\middle|\, x_0 \propto q, \pi\right) \geqslant y\right) \geqslant 1-\epsilon, \tag{3b}$$

where the expectation is taken with respect to the random trajectory of x when following stationary policy π, and $\tilde r \propto \mathcal{N}(\mu_{\tilde r}, \Theta_{\tilde r})$, is equivalent to the convex second-order cone program

$$\underset{\rho\in\mathbb{R}^{|S|\times|A|}}{\text{maximize}} \quad \sum_a \rho_a^\top \mu_{\tilde r} - \Phi^{-1}(1-\epsilon)\left\|\sum_a \rho_a^\top \Theta_{\tilde r}^{1/2}\right\|_2 \tag{4a}$$

$$\text{subject to} \quad \sum_a \rho_a^\top = q^\top + \sum_a \alpha \rho_a^\top P_a, \tag{4b}$$

$$\rho_a^\top \geqslant 0 \quad \forall a \in A, \tag{4c}$$

where given an optimal assignment ρ^*, an optimal policy π^* to problem (3) *can be retrieved using*

$$\pi^*(s,a) = \begin{cases} \dfrac{1}{|A|} & \text{if } \sum_a \rho_a^*(s) = 0, \\ \dfrac{\rho_a^*(s)}{\sum_a \rho_a^*(s)} & \text{otherwise.} \end{cases} \tag{5}$$

Proof. We first use the fact that with fixed reward uncertainty, constraint (3b) can be expressed in the form

$$\mathbb{P}_{\tilde r}(v^\top \tilde r \geqslant y) \geqslant 1-\epsilon, \tag{6a}$$

$$q^\top \sum_{t=0}^{\infty}(\alpha \Pi P)^t = v^\top. \tag{6b}$$

Using a change of variable that is commonly used in the MDP litterature (see Puterman 1994), constraint (6b) is equivalent to

$$v^\top = q^\top + \alpha \sum_a \rho_a^\top P_a, \tag{7a}$$

$$v^\top = \sum_{a\in A} \rho_a^\top \quad \rho_a^\top \geqslant 0, \ \forall a \in A, \tag{7b}$$

where $\rho_a \in \mathbb{R}^{|S|}$. From feasible point (v, ρ), an equivalent pair (v, Π) feasible according to constraint (6b) can be retrieved using

$$\Pi(s,s',a) = \begin{cases} 0 & \text{if } v(s') = 0 \\ \dfrac{\rho_a(s')}{v(s')}\mathbb{I}\{s = s'\} & \text{otherwise.} \end{cases} \tag{8}$$

Given that $\epsilon \leqslant 0.5$, one can use Lemma 1 to convert constraint (6a) into an equivalent deterministic convex constraint. Theorem 1 follows naturally. □

3.2. Complexity of the Solution

It is important to note that second-order cone programming (SOCP) is a well-developed field of optimization for which a number of polynomial-time algorithms have been proposed. We refer the reader to Lobo et al. (1998) for background on the subject and algorithms for solving this family of problems.[1] Based on a primal-dual interior point method presented in Lobo et al. (1998), we can show the following.

Theorem 2. *Given an N states, M actions MDP with fixed Gaussian uncertainty in the reward vector, chance-constrained problem* (3) *can be solved in time $O(M^{7/2}N^{7/2})$.*

Proof. Based on the work presented in Lobo et al. (1998), solving an SOCP to any precision is bounded above by $O(\sqrt{K}(k^2\sum_{i=1}^K k_i + k^3))$, where K is the number of constraints, k is the number of variables, and k_i is the size of the vector in the norm operator of constraint i. These results lead to a bound of

$$O\left(\sqrt{MN+N+1}(M^2N^2N + M^3N^3)\right) = O(M^{7/2}N^{7/2})$$

for problem (4) and consequently for problem (3) because the transformation from one problem to the other does not depend on the size of the MDP. □

Note that following Calafiore and El Ghaoui (2006), it is possible to reduce the Gaussian reward assumption while preserving tractability of the percentile problem. An example of such a reduction can be referred to as the *Q-radial distribution assumption*. The random vector $\tilde r$ is said to have a Q-radial distribution if it can be defined as $\tilde r = Q\widetilde{w} + \mu_{\tilde r}$, where $\mu_{\tilde r} = \mathbb{E}(\tilde r)$, $Q \in \mathbb{R}^{|S|\times k}$ for some $k \leqslant |S|$, and $\widetilde{w} \in \mathbb{R}^k$ is a random vector having probability density $f(\widetilde{w})$ that depends only on the norm of $\widetilde{w}$ (i.e., $f(\widetilde{w}) =$

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

$g(\|\widetilde{w}\|_2)$). Theorem 1 can naturally be extended for radial distributions.

Unfortunately, one can also show that some uncertainty models on the reward parameters actually lead to intractable forms for percentile problem (3).

THEOREM 3. *Solving the chance-constrained MDP problem* (3) *with general uncertainty in the reward parameters is NP-hard.*

A detailed proof of this theorem, where we show that the NP-complete 3SAT problem can be reduced to solving problem (3) for an MDP with discrete reward uncertainty, is presented in the online appendix.

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

4. Decision Making Under Uncertain Transition Parameters

We now focus on the problem of transition parameter uncertainty. This type of uncertainty is present in applications where one does not have a physical model of the dynamics of the system. In this case, P must be estimated from experimentation and is therefore inherently uncertain. Because the Bayesian framework allows us to formulate a distribution over $\widetilde{P}$, we consider a chance-constrained MDP problem with transition uncertainty:

$$\underset{y\in\mathbb{R},\,\pi\in\Upsilon}{\text{maximize}} \quad y \tag{9a}$$

$$\text{subject to} \quad \mathbb{P}_{\widetilde{P}}\left(\mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t r_t(x_t)\,\middle|\, x_0\propto q,\pi\right)\geqslant y\right) \geqslant 1-\epsilon, \tag{9b}$$

where the probability $\mathbb{P}_{\widetilde{P}}$ is the probability of drawing the transition matrix $\widetilde{P}$ from a distribution $f(\widetilde{P})$, and where $\mathbb{E}_x(\cdot \mid x_0\propto q,\pi)$ is the expectation of the trajectory given a concrete realization of $\widetilde{P}$, deterministic rewards r, a policy π, and a distribution of the initial state q. As was the case for reward uncertainty, this problem is hard to solve in general. However, in §4.3 we use the Dirichlet prior to suggest a method that generates a near optimal policy given a sufficient number of samples drawn from $\widetilde{P}$.

4.1. Computational Complexity of Uncertainty in the Transition Parameters

Finding an optimal policy, according to the chance-constrained problem, for an uncertain MDP is NP-hard even if there is no uncertainty in the reward parameters.

COROLLARY 1. *Solving chance-constrained MDP problem* (9) *for general uncertainty in the transition parameters is NP-hard.*

Following similar lines as for proving Theorem 3, given an instance of the NP-complete 3SAT problem, one can easily construct in polynomial time an MDP with discrete transition uncertainty. Solving problem (9) for this uncertain MDP is equivalent to determining if the 3SAT instance is satisfiable. A sketch of this proof is included in the online appendix.

4.2. The Dirichlet Prior on Transition Probability

Because we cannot expect to solve chance-constrained problem (9) for a general distribution, for each state-action pair (i,a), we will use independent Dirichlet priors to model the uncertainty in the parameters of $\widetilde{P}_{(i,a)}(j)$, the probability of observing a transition to state j out of state i when taking action a. This assumption is very convenient for describing prior knowledge about transition parameters due to the fact that, after gathering new transition observations, one can easily evaluate a posterior distribution over these parameters. More specifically, for a vector of transition parameters $\tilde{p}=(\tilde{p}_1,\ldots,\tilde{p}_N)$, the Dirichlet distribution over $\tilde{p}$ follows the density function $f(p)=(1/Z(\beta))\prod_{j=1}^N p_j^{\beta_j-1}$, where β are modeling parameters for the Dirichlet prior and $Z(\beta)$ is a normalization factor. Given a set of observed transition observations $\{j^{(1)},j^{(2)},\ldots,j^{(M)}\}$ from the multinomial distribution $f(j\mid p)=p_j$, one can analytically resolve the posterior distribution over $\tilde{p}$. This distribution conveniently takes the same Dirichlet form $f(p\mid j^{(1)},j^{(2)},\ldots,j^{(M)})=(1/Z(\beta,M_1,\ldots,M_N))\prod_{j=1}^N p_j^{\beta_j+M_j-1}$, where M_j is the number of times that a transition to j was observed. It is also known that the covariance between different terms of $\tilde{p}$ is (see Gelman et al. 2003 for details):

$$\Theta_{(j,k)}=-\frac{(\beta_k+M_k)(\beta_j+M_j)}{(\beta_0+M)^2(\beta_0+M+1)},$$

$$\Theta_{(j,j)}=\frac{(\beta_j+M_j)(\beta_0+M-\beta_j-M_j)}{(\beta_0+M)^2(\beta_0+M+1)},$$

where $\beta_0=\sum_j\beta_j$ and $M=\sum_j M_j$.

4.3. Expected Return Approximation Using a Dirichlet Prior

Even with the Dirichlet assumption we are confronted with the following difficulty in solving percentile problem (9). Unlike in the case of reward uncertainty (where under fixed reward uncertainty and known transitions parameters, $\mathbb{E}_{\tilde{r},x}(\sum_{t=0}^{\infty}\alpha^t\tilde{r}(x_t)\mid x_0\propto q,\pi)=q^{\mathsf{T}}(I-\alpha\Pi P)^{-1}\mathbb{E}(\tilde{r})$ and the optimal policy can be found using the nominal problem), finding a policy that simply minimizes the expected return $\mathbb{E}_{\widetilde{P},x}(\sum_{t=0}^{\infty}\alpha^t r(x_t)\mid x_0\propto q,\pi)$ under transition uncertainty $\widetilde{P}$ is already hard. More specifically, the expected return can be expressed as

$$\mathbb{E}_{\widetilde{P},x}\left(\sum_{t=0}^{\infty}\alpha^t r(x_t)\,\middle|\, x_0\propto q,\pi\right)$$

$$
\begin{aligned}
&= \mathbb{E}_{\tilde{P}}\left(\mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t r(x_t)\,\middle|\, x_0 \propto q, \pi\right)\right)\\
&= \mathbb{E}_{\tilde{P}}\left(q^{\mathsf{T}}(I-\alpha\Pi\tilde{P})^{-1}r\right)\\
&= \mathbb{E}_{\tilde{P}}\left(q^{\mathsf{T}}(I-\alpha\Pi(\mathbb{E}(\tilde{P})+\Delta\tilde{P}))^{-1}r\right)\\
&= \mathbb{E}_{\tilde{P}}\left(q^{\mathsf{T}}((X^{\pi})^{-1}-(X^{\pi})^{-1}\alpha X^{\pi}\Pi\Delta\tilde{P})^{-1}r\right)\\
&= \mathbb{E}_{\tilde{P}}(q^{\mathsf{T}}(I-\alpha X^{\pi}\Pi\Delta\tilde{P})^{-1}X^{\pi}r)\\
&= \mathbb{E}_{\tilde{P}}\left(q^{\mathsf{T}}\sum_{k=0}^{\infty}\alpha^k(X^{\pi}\Pi\Delta\tilde{P})^k X^{\pi}r\right),
\end{aligned}
$$

where $\Delta\tilde{P}=\tilde{P}-\mathbb{E}(\tilde{P})$ and $X^{\pi}=(I-\alpha\Pi\mathbb{E}(\tilde{P}))^{-1}$. The matrix X^{π} is always well defined because $\tilde{P}$ is modeled with the Dirichlet distribution, thus ensuring that $\mathbb{E}(\tilde{P})$ is a valid transition matrix and that $I-\alpha\Pi\mathbb{E}(\tilde{P})$ is nonsingular. $\mathbb{E}_{\tilde{P},x}(\sum_{t=0}^{\infty}\alpha^t r(x_t)\mid x_0\propto q,\pi)$ therefore depends on all the moments of the uncertainty in $\tilde{P}$. Following similar lines as in Mannor et al. (2007), we focus on finding a stationary policy that performs well according to the second-order approximation of the expected return. We expect the norm of higher-order moments of $\Delta\tilde{P}$ to decay with the number of observed transitions:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{P},x}\left(\sum_{t=0}^{\infty}\alpha^t r(x_t)\,\middle|\, x_0\propto q,\pi,\tilde{P}\right)\\
&= q^{\mathsf{T}}X^{\pi}r+\alpha q^{\mathsf{T}}X^{\pi}\Pi\mathbb{E}(\Delta\tilde{P})X^{\pi}r+\alpha^2 q^{\mathsf{T}}X^{\pi}\\
&\quad\cdot\Pi\mathbb{E}(\Delta\tilde{P}X^{\pi}\Pi\Delta\tilde{P})X^{\pi}r+L_{\exp}\\
&\approx q^{\mathsf{T}}X^{\pi}r+\alpha^2 q^{\mathsf{T}}X^{\pi}\Pi Q X^{\pi}r,
\end{aligned}
$$

where $L_{\exp}=\sum_{k=3}^{\infty}\alpha^k q^{\mathsf{T}}\mathbb{E}((X^{\pi}\Pi\Delta\tilde{P})^k)X^{\pi}r$, and where $Q\in\mathbb{R}^{|S|\times|A|\times|S|}$, such that

$$
\begin{aligned}
Q_{(i,a,j)} &= (\mathbb{E}(\Delta\tilde{P}X^{\pi}\Pi\Delta\tilde{P}))_{(i,a,j)}\\
&= \sum_{k,l,a'}(X^{\pi}\Pi)_{(k,l,a')}\mathbb{E}(\Delta\tilde{P}_{(i,a,k)}\Delta\tilde{P}_{(l,a',j)})\\
&= \sum_{k} X^{\pi}_{(k,i)}\pi_{(i,a)}\mathbb{E}(\Delta\tilde{P}_{(i,a,k)}\Delta\tilde{P}_{(i,a,j)})\\
&= \pi_{(i,a)}\Theta^{(i,a)}_{(j,\cdot)}X^{\pi}_{(\cdot,i)}.
\end{aligned}
$$

This is under the assumption that the rows of $\tilde{P}$ are independent from each other and using $\Theta^{(i,a)}$ to represent the covariance between the terms of the transition vector from state i with action a. We are now interested in the second-order approximation of $\mathbb{E}_{\tilde{P},x}(\sum_{t=0}^{\infty}\alpha^t r(x_t)\mid x_0\propto q,\pi,\tilde{P})$.

Definition 1. $\mathbb{F}(\pi)$ is the second-order approximation of the expected return under transition uncertainty, such that

$$\mathbb{F}(\pi)=q^{\mathsf{T}}X^{\pi}r+\alpha^2 q^{\mathsf{T}}X^{\pi}\Pi Q X^{\pi}r.$$

Remark 1. One should note that the approximation $\mathbb{F}(\pi)$ depends on the first two moments of random matrix $\tilde{P}$. It can therefore efficiently be evaluated for any policy. Although $\mathbb{F}(\pi)$ is still nonconvex in π, in practice, global optimization techniques will lead to useful solutions as presented in §5.2.

Before studying the usefulness of minimizing $\mathbb{F}(\pi)$, we will first introduce the definition of $(1-\epsilon)$-percentile performance for a policy in this context and present a lemma that constrains the range of possible solutions for any chance-constrained problem.

Definition 2. For a fixed policy π, $\mathcal{Y}(\pi,\epsilon)$, the $(1-\epsilon)$-percentile performance of policy π under transition uncertainty $\tilde{P}$, is the solution to

$$
\begin{aligned}
\mathcal{Y}(\pi,\epsilon) = \underset{y\in\mathbb{R}}{\text{maximize}}\quad & y\\
\text{subject to}\quad & \mathbb{P}_{\tilde{P}}\left(\mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t r_t(x_t)\,\middle|\, x_0\propto q,\pi\right)\geqslant y\right)\\
& \geqslant 1-\epsilon,
\end{aligned}
$$

where the probability $\mathbb{P}_{\tilde{P}}$ is the probability of drawing $\tilde{P}$ from the posterior Dirichlet distribution, and where the expectation is taken with respect to the random trajectory of x when following stationary policy π given a concrete realization of $\tilde{P}$.

Lemma 2. *Given any random variable $\tilde{z}$ with mean μ and standard deviation σ, then the optimal value y^* of the optimization problem*

$$
\begin{aligned}
&\underset{y\in\mathbb{R}}{\text{maximize}}\quad y\\
&\text{subject to}\quad \mathbb{P}(\tilde{z}\geqslant y)\geqslant 1-\epsilon
\end{aligned}
$$

is assured to be in the range

$$y^*\in\left[\mu-\sigma/\sqrt{\epsilon},\,\mu+\sigma/\sqrt{1-\epsilon}\right].$$

The proof is given in the online appendix. One can now derive the following theorem.

Theorem 4. *Given state transition observations*

$$\left\{(s_1,a_1,s_1'),\ldots,(s_M,a_M,s_M')\right\}$$

and suppose that $M^*=\min_{i,a}\sum_j M_j^{(i,a)}$, the minimum number of transitions observed from any state using any action, and $\epsilon\in(0,0.5]$, policy

$$\hat{\pi}=\arg\max_{\pi}\mathbb{F}(\pi) \tag{10}$$

is $O(1/\sqrt{\epsilon M^*})$ optimal with respect to the chance-constrained MDP problem

$$\underset{\pi\in\Upsilon}{\text{maximize}}\ \mathcal{Y}(\pi,\epsilon). \tag{11}$$

Proof. Using Lemma 2 with $\tilde{z}$ replaced by $\tilde{g}_{\tilde{P}}(\pi)=\mathbb{E}_x(\sum_{t=0}^{\infty}\alpha^t r(x_t)\mid x_0\propto q,\pi,\tilde{P})$, one can easily show that for any policy π,

$$
\begin{aligned}
&\mathcal{Y}_{\tilde{P}}(\pi,\epsilon)-\mathbb{F}(\pi)\\
&\quad\leqslant\mathbb{E}_{\tilde{P}}(\tilde{g}_{\tilde{P}}(\pi))+\frac{1}{\sqrt{1-\epsilon}}\sqrt{\mathbb{E}_{\tilde{P}}(\tilde{g}_{\tilde{P}}(\pi)^2)-\mathbb{E}_{\tilde{P}}(\tilde{g}_{\tilde{P}}(\pi))^2}-\mathbb{F}(\pi)\\
&\quad= L_{\exp}(\pi)+\sqrt{\frac{L_{\mathrm{var}}(\pi)}{1-\epsilon}}
\end{aligned}
$$

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

and

$$\begin{aligned}
&\mathcal{Y}_{\tilde{P}}(\pi,\epsilon)-\mathbb{F}(\pi)\\
&\quad\geqslant \mathbb{E}_{\tilde{P}}(\tilde{g}_{\tilde{P}}(\pi))-\frac{1}{\sqrt{\epsilon}}\sqrt{\mathbb{E}_{\tilde{P}}(\tilde{g}_{\tilde{P}}(\pi)^2)-\mathbb{E}_{\tilde{P}}(\tilde{g}_{\tilde{P}}(\pi))^2}-\mathbb{F}(\pi)\\
&\quad= L_{\exp}(\pi)-\sqrt{\frac{L_{\mathrm{var}}(\pi)}{\epsilon}}\ ,
\end{aligned}$$

where

$$\begin{aligned}
L_{\exp}(\pi)&=\sum_{k=3}^{\infty}\alpha^k q^{\mathsf{T}}\mathbb{E}((X^{\pi}\Pi\Delta\tilde{P})^k)X^{\pi}r=O\left(\frac{1}{(M^*)^2}\right),\\
L_{\mathrm{var}}(\pi)&=\mathbb{E}_{\Delta\tilde{P}}(\mathbb{E}(\tilde{y}_{\pi}\mid\Delta\tilde{P})^2)-\mathbb{E}(\tilde{y}_{\pi})^2\\
&=\mathbb{E}\left(\left(q^{\mathsf{T}}\sum_{k=0}^{\infty}\alpha^k(X^{\pi}\Pi\Delta\tilde{P})^kX^{\pi}r\right)^2\right)-\mathbb{E}(\tilde{y}_{\pi})^2\\
&=\sum_{k,\,l:\,k+l\geqslant 0}\mathbb{E}\big(\alpha^{k+l}q^{\mathsf{T}}(X^{\pi}\Pi\Delta\tilde{P})^kX^{\pi}r\\
&\qquad\qquad\cdot q^{\mathsf{T}}(X^{\pi}\Pi\Delta\tilde{P})^lX^{\pi}r\big)-\mathbb{E}(\tilde{y}_{\pi})^2\\
&=\sum_{k,\,l:\,k+l\geqslant 2}\mathbb{E}\big(\alpha^{k+l}q^{\mathsf{T}}(X^{\pi}\Pi\Delta\tilde{P})^kX^{\pi}r\\
&\qquad\qquad\cdot q^{\mathsf{T}}(X^{\pi}\Pi\Delta\tilde{P})^lX^{\pi}r\big)=O\left(\frac{1}{M^*}\right),
\end{aligned}$$

where the bounds $O(1/(M^*)^2)$ and $O(1/M^*)$ were derived from the rate of decay for each moment of a Dirichlet distribution (see Wilks 1962 for details on these moments).

This gives us a bound between the optimal $(1-\epsilon)$-percentile performance obtained from policy $\pi^*=\arg\max_{\pi}\mathcal{Y}_{\tilde{P}}(\pi,\epsilon)$ and $\hat{\pi}$ returned by problem (10):

$$\begin{aligned}
&\mathcal{Y}_{\tilde{P}}(\pi^*,\epsilon)-\mathcal{Y}_{\tilde{P}}(\hat{\pi},\epsilon)\\
&\quad=\mathcal{Y}_{\tilde{P}}(\pi^*,\epsilon)-\mathbb{F}(\pi^*)+\mathbb{F}(\pi^*)-\mathcal{Y}_{\tilde{P}}(\hat{\pi},\epsilon)\\
&\quad\leqslant\mathcal{Y}_{\tilde{P}}(\pi^*,\epsilon)-\mathbb{F}(\pi^*)+\mathbb{F}(\hat{\pi})-\mathcal{Y}_{\tilde{P}}(\hat{\pi},\epsilon)\\
&\quad\leqslant L_{\exp}(\pi^*)+\frac{\sqrt{L_{\mathrm{var}}(\pi^*)}}{\sqrt{1-\epsilon}}-L_{\exp}(\hat{\pi})+\frac{\sqrt{L_{\mathrm{var}}(\hat{\pi})}}{\sqrt{\epsilon}}\\
&\quad=O\left(\frac{1}{\sqrt{\epsilon M^*}}\right).\quad\square
\end{aligned}$$

4.4. Improving the Bound with Action Elimination

In some instances of MDPs with transition uncertainty, it might be the case that little observations were made from state-action pairs that were observed to have low return. Unfortunately, although most likely neither the true optimal percentile policy nor the approximate one put positive weight on these state-action pairs, Theorem 4 states that our confidence in the approximate policy should depend on this reduced number of transition observations from the given pairs. We apply the idea of action elimination, proposed by MacQueen (1966) in the context of the nominal MDP, to the percentile optimization framework to relax this dependence.

Definition 3. Let $\mathcal{B}$ be an arbitrary set of undesirable state-action pairs such that for any state i there exists an action a for which $(i,a)\notin\mathcal{B}$. Let $\mathcal{B}^c$ be the complement of $\mathcal{B}$ with respect to $S\times A$.

To prevent the dependence of the proposed bound on the state-action pairs in $\mathcal{B}$, we propose a simple test that will allow us to redefine M^* in Theorem 4 as $M^{**}=\min_{(i,a)\in\mathcal{B}^c}\sum_j M_j^{(i,a)}$.

Definition 4. Considering $\Upsilon_{\mathcal{B}^c}$ to be the set of stationary policies that have support strictly on state-action pairs in $\mathcal{B}^c$, let

$$Q_+(i,a;\mathcal{B}^c)=\sup_{\substack{\pi\in\Upsilon_{\mathcal{B}^c}\\ P\in\ \mathrm{support}(\tilde{P})}}\mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t r(x_t)\,\middle|\,x_0=i,\pi,a_0=a\right)$$

be the highest achievable expected return given that one starts in state i, and takes action a before following a policy in $\Upsilon_{\mathcal{B}^c}$.

Similarly, let

$$Q_-(i,a;\mathcal{B}^c)=\inf_{\substack{\pi\in\Upsilon_{\mathcal{B}^c}\\ P\in\mathrm{support}(\tilde{P})}}\mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t r(x_t)\,\middle|\,x_0=i,\pi,a_0=a\right)$$

be the lowest achievable expected return given that one starts in state i, and takes action a before following a policy in $\Upsilon_{\mathcal{B}^c}$.

Both of these limits are finite, using the fact that the expected return is always bounded above by $1/(1-\alpha)$ times the largest achievable reward and below by $1/(1-\alpha)$ times the smallest achievable one.

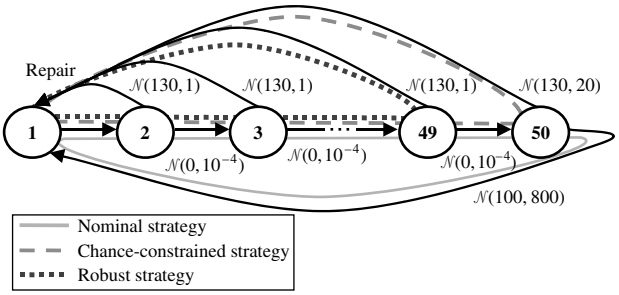
Corollary 2. *Suppose that a set of state-action pairs $\mathcal{B}$ according to Definition* 3 *and a slack parameter $\lambda\geqslant 0$ satisfy the condition*

$$Q_+(i,a;\mathcal{B}^c)\leqslant\max_{b:\,(i,b)\in\mathcal{B}^c}Q_-(i,b;\mathcal{B}^c)+\lambda\quad\forall\,(i,a)\in\mathcal{B}.$$

Then, for state-transition observations $\{(s_1,a_1,s_1'),\ldots,(s_M,a_M,s_M')\}$ and $\epsilon\in(0,0.5]$, the policy obtained solving problem (10) *is $O(1/\sqrt{\epsilon M^{**}}+\lambda/(1-\alpha))$ optimal according to problem* (11)*, where $M^{**}=\min_{(i,a)\in\mathcal{B}^c}\sum_j M_j^{(i,a)}$.*

Note that $Q_+(i,a;\mathcal{B}^c)$ and $Q_-(i,a;\mathcal{B}^c)$ can be computed to a sufficient level of accuracy for all (i,a) pairs using backup operations similar to what was presented in Nilim and El Ghaoui (2005). The proof is presented in the online appendix and relies mostly on applying Theorem 4 on a version of the MDP that does not possess the state-action pairs in $\mathcal{B}$. As a final remark on this result, Corollary 2 can easily be extended to a probabilistic setting where Q_+ and Q_- are defined in terms of high-probability bounds.

Figure 1. Instance of a machine replacement problem with fixed uncertainty in the rewards.



Note. The optimal paths followed for three strategies are shown.

5. Experiments

We have chosen the machine replacement problem as an application for our methods. Consider the repair cost that is incurred by a factory that holds a high number of machines, given that each of these machines is modeled with the same underlying MDP for which parameters are not known with certainty. In such a setting, it would be natural to apply a repair policy uniformly on all the machines with the hope that, with probability higher than $1-\epsilon$, this policy will have a low maintenance cost on average. This is specifically what the percentile criterion quantifies. We now present two instances of this problem with either reward or transition parameter uncertainty. Note that we have selected simple instances of this problem to present clearly how our method compares to the nominal and the robust approaches described in §2. In fact, our methods remain computationally tractable with machine replacement problems of more than 1,000 states.

5.1. Machine Replacement as an MDP with Gaussian Rewards

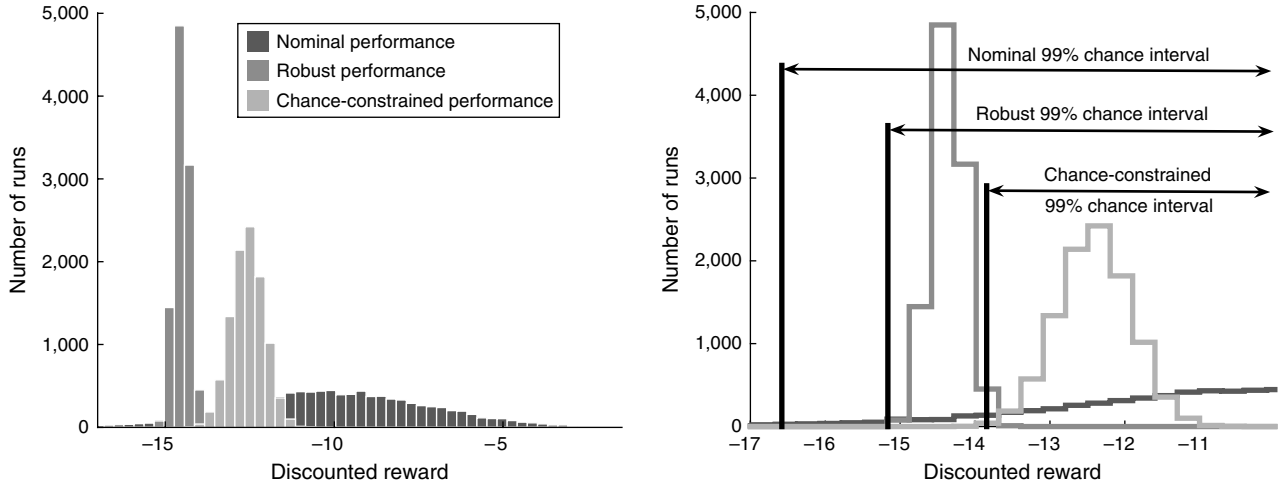
In our experiment with Gaussian reward MDP, we used a simple version of the machine replacement problem with 50 states, 2 actions, deterministic transitions, a discount factor of 0.8, and fixed Gaussian uncertainty in the rewards (see Figure 1). Our model develops as follows: after the policy is chosen by the agent, the environment is created according to a predefined joint Gaussian distribution over the rewards, and the policy is applied on this environment, which is solely deterministic thereafter. For each of the first 49 steps, repairs have a cost independently distributed as $\mathcal{N}(130,1)$. The 50th state of the machine's life was designed to be a more risky state: Not repairing incurs a highly uncertain cost $\mathcal{N}(100,800)$, while repairing is a more secure but still uncertain option $\mathcal{N}(130,20)$.

The performance of policies obtained using nominal, robust and 99% chance-constrained problem formulations[2] is presented in Figure 2. These results describe what one would typically expect from the three solution concepts. While the nominal strategy, blind to any form of risk, finds no advantage in ever repairing, the robust strategy ends up following a highly conservative policy (repairing the machine in state #49 to avoid state #50). On the other hand, the 99% chance-constrained optimal strategy handles the risk more efficiently by waiting until state #50 to apply a mixed strategy that repairs with 90% probability. This strategy performed better than its robust alternative while preserving small variance in performance over the 10,000 different sampled environments.

5.2. Machine Replacement as an MDP with Dirichlet Prior on Transitions

In this experiment, we use a version of the machine replacement problem with 10 states, 4 actions, a discount

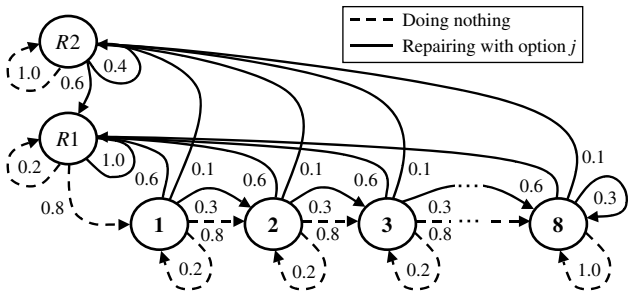
Figure 2. Performance comparisons between nominal, robust, and chance-constrained policies on 10,000 runs of the machine replacement problem.



Note. The right figure focuses on the interval $[-17,-10]$.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

Figure 3. Instance of a machine replacement problem with Dirichlet uncertainty in the transition parameters.



Notes. The graph presents the expected transition probabilities for the two types of actions (repairing, or not) after observing M transitions from each state. In our experiments, three repair options are available, all three leading to slightly perturbed version of the Dirichlet model presented here.

factor of 0.8, a uniform initial state distribution and transition uncertainty modeled with Dirichlet distribution. States 1 to 8 describe the normal aging of the machine, while states $R1$ and $R2$ represent two possible stages of repairs: $R1$ for normal repairs on the machine costing 2, and $R2$ for a harder one with a cost of 10. Letting the machine reach the age of 8 is penalized with a cost of 20. In each of these states, one has access to three repair services for the machine. We designed a Dirichlet model for transitions occurring when no repairs are done. In the case of each of the three repair options, for simplicity we used slightly perturbed versions of a reference Dirichlet model that is presented in Figure 3. In this figure, the expected transition parameters are presented given that M transitions were observed from each state. The parameter M acts as a control for the amount of transition uncertainty present in the model.

We applied three solution methods to this decision problem. First, the nominal problem was formulated using the expected transition probabilities from the Dirichlet distribution. Then, we applied the robust method presented in §2.2. As mentioned earlier, it is unclear how to state the robust MDP problem when using probabilistic models for parameter uncertainty. Here, we chose to evaluate the 90% percentile performance of policies and therefore built a 90% confidence box in $\mathbb{R}^{|S|\times|A|\times|S|}$ for the random vector $\tilde{P}$. (Using 10,000 samples drawn from $\tilde{P}$ and a given γ ratio, for each parameter $P_{(i,a,j)}$ we chose $A_{(i,a,j)}$ and $B_{(i,a,j)}$ so that they included a ratio of γ of the random samples. A search over γ was done to find the minimal γ that led to a box $A_{(i,a,j)} \leqslant P_{(i,a,j)} \leqslant B_{(i,a,j)}$ containing 90% of the samples drawn from $\tilde{P}$. We do not discuss the validity of this method because it is purely illustrative of the difficulties involved in the choice of a 90% uncertainty set for $\tilde{P}$.) Finally, we used the second-order approximation performance measure presented in §4.3 to find an optimal policy for this machine replacement problem. To do so, we were required to solve a nonconvex optimization problem using a gradient descent algorithm (applied on $-\mathbb{F}(\pi)$). The gradient of $\mathbb{F}(\pi)$ was found to be

$$\begin{aligned}\frac{\partial \mathbb{F}(\pi)}{\partial \pi_{(i,a)}} = \sum_{k,l} \big(& q_k r_l + \alpha^2 q_k (\Pi_{(l,\cdot,\cdot)} Q X^\pi r) \\ & + \alpha^2 (q^\top X^\pi \Pi Q_{(\cdot,\cdot,k)}) r_l \big) \frac{\partial X^\pi_{(k,l)}}{\partial \pi_{(i,a)}} \\ & + \alpha^2 (q^\top X^\pi_{(\cdot,i)})(Q_{(i,a,\cdot)} X^\pi r) \\ & + \alpha^2 \sum_{k,a',l} (q^\top X^\pi_{(\cdot,k)})(X^\pi_{(l,\cdot)} r) \pi_{(k,a')} \frac{\partial Q_{(k,a',l)}}{\partial \pi_{(i,a)}},\end{aligned}$$

where

$$\begin{aligned}\frac{\partial Q_{(k,a',l)}}{\partial \pi_{(i,a)}} &= \mathbb{1}\{i = k \wedge a = a'\} \Theta^{(i,a)}_{(l,\cdot)} X^\pi_{(\cdot,i)} \\ &\quad + \pi_{(k,a')} \sum_r \Theta^{(k,a')}_{(l,r)} \frac{\partial X^\pi_{(r,k)}}{\partial \pi_{(i,a)}}, \\ \frac{\partial X^\pi_{(k,l)}}{\partial \pi_{(i,a)}} &= \alpha X^\pi_{(k,i)} (P_{(i,a,\cdot)} X^\pi_{(\cdot,l)}).\end{aligned}$$

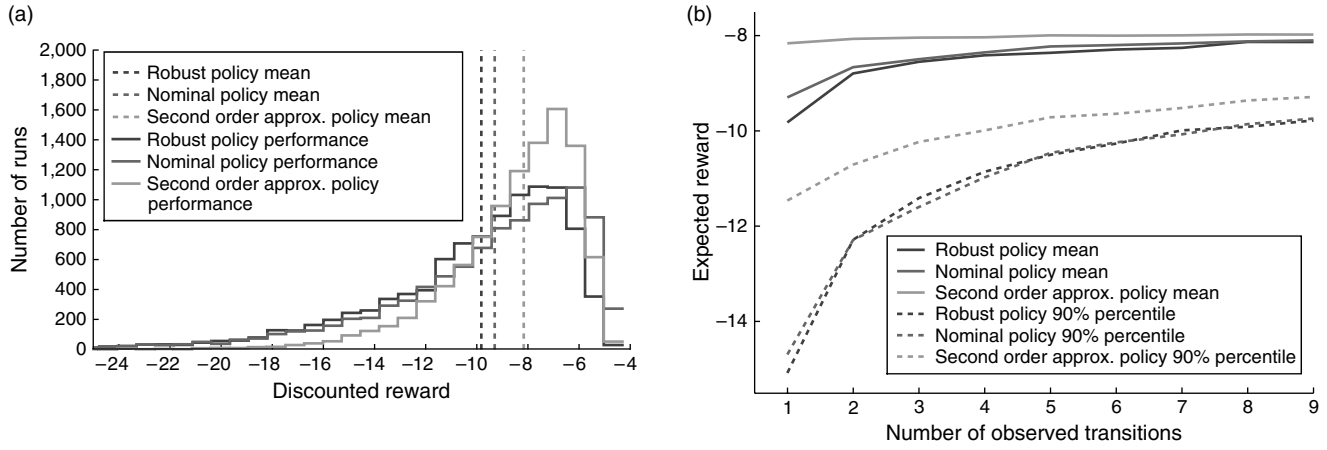
Although gradient descent techniques provide no guarantees of reaching a global optimum, by taking as initial point the policy returned by the nominal problem, we were assured to find a policy that performs better than the nominal one with respect to $\mathbb{F}(\pi)$.[3] Figure 4(a) shows the histogram of expected discounted rewards obtained using the different methods on 10,000 instances of the described uncertain machine replacement problem (with $M = 1$). We also indicated the mean and the 90% percentile of the different methods. It is interesting to see that although the second-order approximation method and the nominal method do not directly address the percentile criterion, the 90% percentile performance actually outperforms the policy obtained using the robust method for large parameter uncertainty. When having a look at the different policies returned by the methods, we realize that the robust policy again acts very conservatively by applying repairs too early. On the other hand, the nominal strategy does not make any use of the fact that three repair options are available. The second-order approximation method returns a policy that, for example, uses a mixed strategy over the repair options in states $R1$ (i.e., heavy repair state) to reduce the variance of transition probabilities and, indirectly, the overall expected cost. In Figure 4(b), we show how these results evolve with the number of observed transitions (quantified by M in the Dirichlet model). As expected, when more transitions are observed, the second-order approximation policy slowly converges to the nominal policy, due to the vanishing second term of $\mathbb{F}(\pi)$.

6. Conclusion

In this paper, we present a chance-constrained formulation for MDPs with uncertain parameters. We show that, although some of its instances are intractable to solve, some instances of this problem can be efficiently solved using

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

Figure 4. (a) presents a performance comparison between nominal, robust, and chance-constrained policies on 10,000 runs of the machine replacement problem with $M = 1$. (b) presents the effect of decreasing the uncertainty in the transitions on the mean and the 90% percentile performances of the different methods.



second-order cone programming. In fact, our experiments demonstrate that, given a preferred level of risk, the proposed criterion compares favorably with policies derived using a nominal model or a robust approach. We believe that many important problems that are usually addressed using standard MDP models should be revisited and better resolved using our proposed models for parameter uncertainty (e.g., machine replacement, inventory management, some queueing control problems, etc.). Finally, we consider the chance-constrained formulation to be an important step toward the optimization of data-driven MDPs. Given that the MDP's parameters are estimated based on data, this formulation naturally enables the decision maker to account for parameter uncertainty.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Appendix. The Frequentist Approach

Interestingly, the percentile criterion can also be reformulated under the frequentist perspective. In this context, one makes no prior assumption on the parameters $\tilde{r}$ and $\tilde{P}$ but instead bases his analysis solely on realized instances of these variables. When estimating the reward associated with each state of the MDP, based on the central limit theorem, one can typically approximate his uncertainty using the Gaussian distribution. It is easy to show that given enough noisy measurements of $\tilde{r}$, Theorem 1 can be applied to this context.

In the case of the transition probabilities, one assumes that for each state-action pair (i, a), there exists an underlying multinomial distribution $P_{(i,a)}(j)$ describing the transitions of the system. Given enough examples of transitions from state i using action a, one typically builds an estimate $\hat{P}_{(i,a)}(j)$ based on the frequencies of transitions. One must now consider the uncertainty related to mean estimation from samples $\Delta\tilde{P}_{(i,\cdot)} = P_{(i,\cdot)} - \hat{P}_{(i,\cdot)}$ for which mean and covariance can be approximated using the central limit theorem. Because of the nature of the multinomial distribution, one can show that third and higher moments of $\Delta\tilde{P}$ decrease in magnitude with the number of observed transitions. Thus, the algorithm and performance bounds presented in Theorem 4 extend naturally to the frequentist framework. We encourage interested readers to find more insights on this problem in Mannor et al. (2007).

We would like to briefly outline an alternate frequentist approach for dealing with reward uncertainty. Given that the two first moments of $\tilde{r}$ are estimated, based on the sampling, to be close to $(\mu_{\tilde{r}}, \Theta_{\tilde{r}})$ with high probability, a rigorous interpretation of the percentile criterion (called distributionally robust) can enforce the chance constraint to be met over the set $\mathcal{F}_{\mu_{\tilde{r}},\Theta_{\tilde{r}}}$ of all possible distributions with such moments. The concept of distributionally robust solutions is commonly applied in the field of stochastic optimization (see Shapiro and Kleywegt 2002). Using Theorem 3.1 from Calafiore and El Ghaoui (2006), Theorem 1 can naturally be extended to this case.

COROLLARY 3. *Given that $\tilde{r}$ is drawn from a distribution in the set $\mathcal{F}_{\mu_{\tilde{r}},\Theta_{\tilde{r}}}$, Theorem* 1 *holds with chance constraint* (3b) *replaced with the distributionally robust chance constraint*

$$\inf_{f_{\tilde{r}}\in\mathcal{F}_{\mu_{\tilde{r}},\Theta_{\tilde{r}}}} \mathbb{P}_{\tilde{r}}\left(\mathbb{E}_x\left(\sum_{t=0}^{\infty}\alpha^t\tilde{r}(x_t)\,\middle|\, x_0\propto q,\,\pi\right)\geqslant y\right)\geqslant 1-\epsilon,$$

and objective (4ca) replaced with

$$\underset{\rho\in\mathbb{R}^{|S|\times|A|}}{\text{maximize}}\ \sum_a \rho_a^{\mathsf{T}}\mu_{\tilde{r}} - \sqrt{\frac{1-\epsilon}{\epsilon}}\left\|\left[\sum_a \rho_a^{\mathsf{T}}\Theta_{\tilde{r}}^{\frac{1}{2}}\right]\right\|_2.$$

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

Thus, for any $\epsilon \in (0, 1)$, the distributionally robust version of the discounted reward chance-constrained MDP problem (3) can be solved using an equivalent second-order cone problem.

Endnotes

1. In our implementation, we used a toolbox developed for Matlab: "CVX: Matlab Software for Disciplined Convex Programming" (Grant and Boyd 2009).
2. Implementation details: the robust problem was solved using the method presented in §2.2, setting the 99% confidence ellipsoid of the random cost vector as the uncertainty set. Also, all second-order cone programming was implemented in Matlab using the CVX software available online at http://www.stanford.edu/∼boyd/cvx/.
3. Implementation details: Matlab's optimization toolbox was used to solve this nonlinear optimization problem.

Acknowledgments

The authors acknowledge the Fonds québécois de la recherche sur la nature et les technologies for their financial support, and they thank Constantine Caramanis and Huan Xu for helpful discussions.

References

Avrachenkov, K. E., J. A. Filar, M. Haviv. 2002. Singular perturbations of Markov chains and decision processes. E. Feinberg, A. Shwartz, eds. *Handbook of Markov Decision Processes*: *Methods and Applications*. Kluwer, Boston, 113–152.

Bagnell, J., A. Y. Ng, J. Schneider. 2001. Solving uncertain Markov decision problems. Technical report CMU-RI-TR-01-25, Robotics Institute, Carnegie Mellon University, Pittsburgh.

Ben-Tal, A., A. Nemirovski. 1998. Robust convex optimization. *Math. Oper. Res.* **23**(4) 769–805.

Bertsekas, D. P., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Cambridge, MA.

Calafiore, G., L. El Ghaoui. 2006. On distributionally robust chance-constrained linear programs. *Optim. Theory Appl.* **130**(1) 1–22.

Charnes, A., W. W. Cooper. 1959. Chance constrained programming. *Management Sci.* **6** 73–79.

Dearden, R., N. Friedman, D. Andre. 1999. Model based Bayesian exploration. *Proc. 15th Conf. Uncertainty in Artificial Intelligence*, Stockholm, 150–159.

Filar, J. A., D. Krass, K. W. Ross. 1995. Percentile performance criteria for limiting average Markov control problems. *IEEE Trans. Automatic Control* **40** 2–10.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Rubin. 2003. *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL.

Givan, R., S. M. Leach, T. Dean. 2000. Bounded-parameter Markov decision processes. *Artificial Intelligence* **122**(1–2) 71–109.

Grant, M., S. Boyd. 2009. CVX: Matlab software for disciplined convex programming (Web page and software). February. http://www.stanford.edu/boyd/cvx.

Howard, R., J. Matheson. 1972. Risk-sensitive Markov decision processes. *Management Sci.* **18**(7) 356–369.

Iyengar, G. 2005. Robust dynamic programming. *Math. Oper. Res.* **30**(2) 257–280.

Lobo, M. S., L. Vandenberghe, S. Boyd, H. Lebret. 1998. Applications of second order cone programming. *Linear Algebra and Its Appl.* **284** 193–228.

MacQueen, J. 1966. A modified dynamic programming method for Markov decision problems. *J. Math. Anal. Appl.* **14** 28–43.

Mannor, S., D. Simester, P. Sun, J. N. Tsitsiklis. 2007. Bias and variance approximation in value function estimates. *Management Sci.* **53**(2) 308–322.

Martin, J. J. 1967. *Bayesian Decision Problems and Markov Chains*. John Wiley & Sons, New York.

Nemirovski, A., A. Shapiro. 2006. Convex approximations of chance constrained programs. *SIAM J. Optim.* **17**(4) 969–996.

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

Prékopa, A. 1995. *Stochastic Programming*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

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

Satia, J. K., R. L. Lave. 1973. Markov decision processes with imprecise transition probabilities. *Oper. Res.* **21**(3) 755–763.

Shapiro, A., A. J. Kleywegt. 2002. Minimax analysis of stochastic problems. *Optim. Methods Software* **17** 523–542.

Silver, E. A. 1963. Markovian decision processes with uncertain transition probabilities or rewards. Technical report 1, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

van der Schaft, A. J. 1996. *L2-Gain and Passivity Techniques in Non-Linear Control*. Springer-Verlag, New York, Inc., Secaucus, NJ.

Wilks, S. S. 1962. *Mathematical Statistics*. John Wiley & Sons, New York.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.