OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Percentile Optimization for Markov Decision Processes with Parameter Uncertainty

Erick Delage
Department of Electrical Engineering, Stanford University, Stanford, California 94305, edelage@stanford.edu, www.stanford.edu/∼edelage

Shie Mannor
Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada, shie.mannor@mcgill.ca, www.ece.mcgill.ca/∼shie

Markov decision processes are an effective tool in modeling decision-making in uncertain dynamic environments. Since the parameters of these models are typically estimated from data or learned from experience, it is not surprising that the actual performance of a chosen strategy often significantly differs 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 point of views on 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: Dynamic programming: Markov, finite state, Probability: stochastic model applications, Programming: stochastic.
Area of review: Stochastic Models.
History: Submitted November 15 2006.

1. Introduction

Markov decision processes (MDPs) are an effective tool in modeling decision-making in uncertain dynamic environments (*e.g.*, Putterman (1994)). Since the parameters of these models are typically 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 significantly differ from the model's predictions (refer to experiments by Mannor et al. (2006)). Let us consider a concrete problem where 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 a MDP with states representing reachable aging phases and actions describing different repair or replacement alternatives. Although the parameters used in such a model can typically 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. Also one often doesn't 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 (2002), 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 Section 5, the robust MDP approach often generates overly conservative strategies. Similar conclusion can be drawn in the context of the H_∞ robust control formulation, as in van der Schaft (1999), 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. (2006) that studied the effect of parameter uncertainty on the mean and variance of the value function of Markov processes with fixed policy, we will 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 the robust methods, our approach will 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 Section 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 Section 3 we will demonstrate that under the assumption that the transitions are known and that 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 Section 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 Section 5, we will illustrate how this criterion outperforms the nominal and robust criterion on instances of the machine replacement problem with either reward or transitions uncertainty.

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 trades 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 Markov decision process described as followed: 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 $P(s,a,s') = \mathbb{P}(s'|s,a)$, an initial distribution on states q, and a reward vector $r \in \mathbb{R}^{|S|}$. Although our analysis will strictly consider the case where the reward only depends 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 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:

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

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. (2006), the authors address this issue by investigating the effect of random $\tilde{r}$ and $\tilde{P}$ on a new nominal problem

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

This problem maximizes the expected return over both the trajectories of x and the random variables $\tilde{r}$ and $\tilde{P}$. Because of the non-linear 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:

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

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 $\mathsf{R}=\{r|f(r)\neq 0\}$ and $\mathsf{P}=\{P|f(P)\neq 0\}$ 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 since the real concern is the level of confidence in the total cumulative reward and not in the individual parameters. Instead, it is much more relevant to express the risk adjusted discounted performance of an uncertain MDP in the following **chance constrained** form:

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

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 $f(\tilde{r}_t)$ and the transition matrix $\tilde{P}$ from $f(\tilde{P})$, and where $\mathbb{E}_x(\cdot|x_0 \propto q, \pi)$ is the expectation of the trajectory given a concrete realization of $\tilde{r}$ 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 (2), under the distribution of $\tilde{r}$ and $\tilde{P}$. Note that, when $\epsilon = 0$, Problem (2) 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), Nilim and El Ghaoui (2005).

Section 3 will focus on uncertainty in the reward parameters. Later, in Section 4, parameter uncertainty will be addressed. Although we do limit ourselves to presenting the details from a Bayesian point of view in order to preserve the clarity of our derivations, a frequentist approach to the percentile criterion do follow naturally and is summarized in the paper 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 i-th row, j-th 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 $(\Pi P)_{(i,j)} = (P_\pi)_{(i,j)} = \mathbb{P}(s' = j|s = i, a = \pi(i))$.

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 as 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 $\{r_1, r_2, ..., r_m\}$ 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):

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

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

$$\mathbb{P}(x^\mathsf{T}\xi \leq 0) \geq 1 - \epsilon$$

is the same as those satisfying

$$x^\mathsf{T}\mu_\xi + \Phi^{-1}(1-\epsilon)\sqrt{x^\mathsf{T}\Theta_\xi x} \leq 0 \ ,$$

where $\mu_\xi = \mathbb{E}(\xi)$, Θ_ξ *is the covariance matrix of the random vector* ξ, ϵ *is a fixed probability such that* $0 \leq \epsilon \leq 1$, *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*

$$\begin{aligned} &\underset{y\in\mathbb{R},\pi\in\Upsilon}{\text{maximize}} && y && \text{(3a)}\\ &\text{subject to} && \mathbb{P}_{\tilde{r}}(\mathbb{E}_x(\textstyle\sum_{t=0}^{\infty}\alpha^t\tilde{r}(x_t)|x_0 \propto q,\pi) \geq y) \geq 1-\epsilon \ , && \text{(3b)}\end{aligned}$$

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*

$$\begin{aligned} &\underset{\rho\in\mathbb{R}^{|S|\times|A|}}{\text{maximize}} && \textstyle\sum_a \rho_a^\mathsf{T}\mu_{\tilde{r}} - \Phi^{-1}(1-\epsilon)\|\sum_a \rho_a^\mathsf{T}\Theta_{\tilde{r}}^{\frac{1}{2}}\|_2 && \text{(4a)}\\ &\text{subject to} && \textstyle\sum_a \rho_a^\mathsf{T} = q^\mathsf{T} + \sum_a \alpha\rho_a^\mathsf{T}P_a && \text{(4b)}\\ &&& \rho_a^\mathsf{T} \geq 0 \ , \quad \forall \ a\in A \ , && \text{(4c)}\end{aligned}$$

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

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

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

$$\mathbb{P}_{\tilde{r}}(\upsilon^\mathsf{T}\tilde{r} \geq y) \geq 1-\epsilon \tag{6a}$$

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

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

$$\upsilon^\mathsf{T} = q^\mathsf{T} + \alpha\textstyle\sum_a \rho_a^\mathsf{T}P_a \tag{7a}$$

$$\upsilon^\mathsf{T} = \textstyle\sum_{a\in A}\rho_a^\mathsf{T} \ , \ \rho_a^\mathsf{T} \geq 0 \ , \quad \forall \ a \in A \ , \tag{7b}$$

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

$$\Pi(s,s',a) = \begin{cases} 0 & \text{if } \upsilon(s') = 0 \\ \frac{\rho_a(s')}{\upsilon(s')}\mathbb{1}\{s=s'\} & \text{otherwise.} \end{cases} \tag{8}$$

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

3.2. Complexity of the solution

It is important to note that "second order cone" programming 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 $O(M^{\frac{7}{2}}N^{\frac{7}{2}})$.*

Proof Based on the work presented in Lobo et al. (1998), solving an SOCP to any precision is bounded above by $O\left(\sqrt{K}(k^2\sum_{i=1}^K k_i + k^3)\right)$, 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

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

for Problem 4 and consequently for Problem 3 since 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 $\tilde{r} = Q\tilde{w} + \mu_{\tilde{r}}$, where $\mu_{\tilde{r}} = \mathbb{E}(\tilde{r})$, $Q \in \mathbb{R}^{|S|\times k}$ for some $k \leq |S|$, and $\tilde{w} \in \mathbb{R}^k$ is a random vector having probability density $f(\tilde{w})$ that only depends on the norm of $\tilde{w}$ (*i.e.*, $f(\tilde{w}) = g(\|\tilde{w}\|_2)$). 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 is presented in the online appendix, where we show that the NP-complete 3SAT problem can be reduced to solving Problem 3 for an MDP with discrete reward uncertainty.

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. Since the Bayesian framework allows us to formulate a distribution over $\tilde{P}$, we consider a chance constrained MDP problem with transition uncertainty:

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

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

where the probability $\mathbb{P}_{\tilde{P}}$ is the probability of drawing the transition matrix $\tilde{P}$ from a distribution $f(\tilde{P})$ and where $\mathbb{E}_x(\cdot|x_0 \propto q, \pi)$ is the expectation of the trajectory given a concrete realization of $\tilde{P}$, 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 section 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 $\tilde{P}$.

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

Since 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 $\tilde{P}_{(i,a)}(j)$, 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 $\tilde{p}=(\tilde{p}_1,...,\tilde{p}_N)$, 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 $\{j^{(1)},j^{(2)},...,j^{(M)}\}$ from the multinomial distribution $f(j|p)=p_j$, one can analytically resolve the posterior distribution over $\tilde{p}$. This distribution conveniently takes the same Dirichlet form $f(p|j^{(1)},j^{(2)},...,j^{(M)})=(1/Z(\beta,M_1,...,M_N))\prod_{j=1}^N p_j^{\beta_j+M_j-1}$, 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):

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

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, $\mathbb{E}_{\tilde{r},x}(\sum_{t=0}^\infty \alpha^t\tilde{r}(x_t)|x_0\propto q,\pi)=q^\top(I-\alpha\Pi P)^{-1}\mathbb{E}(\tilde{r})$ and the optimal policy can be found using the nominal problem), finding a policy that simply minimizes the expected return $\mathbb{E}_{\tilde{P},x}(\sum_{t=0}^\infty \alpha^t r(x_t)|x_0\propto q,\pi)$ under transition uncertainty $\tilde{P}$ is already hard. More specifically, the expected return can be expressed as

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

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 since $\tilde{P}$ 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. $\mathbb{E}_{\tilde{P},x}(\sum_{t=0}^{\infty}\alpha^t r(x_t)|x_0 \propto q, \pi)$ therefore depends on all the moments of the uncertainty in $\tilde{P}$. Following similar lines as in Mannor et al. (2006), 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.

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

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

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

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

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 non-convex in π, in practice, global optimization techniques will lead to useful solutions as presented in Section 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:

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

LEMMA 2. *Given any random variable $\tilde{z}$ with mean μ and variance Θ, then the optimal value y^* of the optimization problem*

$$\underset{y\in\mathbb{R}}{\text{maximize}} \quad y \tag{11a}$$
$$\text{subject to } \mathbb{P}(\tilde{z} \ge y) \ge 1-\epsilon\ , \tag{11b}$$

is assured to be in the range $y^* \in [\mu - \frac{\Theta}{\sqrt{\epsilon}}, \mu + \frac{\Theta}{\sqrt{1-\epsilon}}]$.

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

THEOREM 4. *Given state transition observations* $\{(s_1,a_1,s_1'),...,(s_M,a_M,s_M')\}$ *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*

$$\hat{\pi} = \arg\max_{\pi} \mathbb{F}(\pi) \tag{12}$$

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

$$\underset{\pi\in\Upsilon}{\text{maximize}}\ \mathcal{Y}(\pi,\epsilon)\ \ , \tag{13}$$

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* $\tilde{P}$*.*

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

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

and

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

where

$$\begin{aligned}L_{\text{exp}}(\pi) &= \sum_{k=3}^{\infty}\alpha^k q^{\mathsf{T}}\mathbb{E}\left((X^\pi\Pi\Delta\tilde{P})^k\right)X^\pi r = o(\frac{1}{(M^*)^2})\\ L_{\text{var}}(\pi) &= \mathbb{E}_{\Delta\tilde{P}}\left(\mathbb{E}(\tilde{y}_\pi|\Delta\tilde{P})^2\right) - \mathbb{E}(\tilde{y}_\pi)^2\\ &= \mathbb{E}\left(\left(q^{\mathsf{T}}\sum_{k=0}^{\infty}\alpha^k(X^\pi\Pi\Delta\tilde{P})^k X^\pi r\right)^2\right) - \mathbb{E}(\tilde{y}_\pi)^2\\ &= \sum_{k,l:k+l\geq 0}\mathbb{E}\left(\alpha^{k+l}q^{\mathsf{T}}(X^\pi\Pi\Delta\tilde{P})^k X^\pi r\ q^{\mathsf{T}}(X^\pi\Pi\Delta\tilde{P})^l X^\pi r\right) - \mathbb{E}(\tilde{y}_\pi)^2\\ &= \sum_{k,l:k+l\geq 2}\mathbb{E}\left(\alpha^{k+l}q^{\mathsf{T}}(X^\pi\Pi\Delta\tilde{P})^k X^\pi r\ q^{\mathsf{T}}(X^\pi\Pi\Delta\tilde{P})^l X^\pi r\right) = o(\frac{1}{M^*})\ ,\end{aligned}$$

where the bounds $o(\frac{1}{(M^*)^2})$ and $o(\frac{1}{M^*})$ 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 12.

$$\begin{aligned}\mathcal{Y}_{\tilde{P}}(\pi^*,\epsilon) - \mathcal{Y}_{\tilde{P}}(\hat{\pi},\epsilon) &= \mathcal{Y}_{\tilde{P}}(\pi^*,\epsilon) - \mathbb{F}(\pi^*) + \mathbb{F}(\pi^*) - \mathcal{Y}_{\tilde{P}}(\hat{\pi},\epsilon)\\ &\leq \mathcal{Y}_{\tilde{P}}(\pi^*,\epsilon) - \mathbb{F}(\pi^*) + \mathbb{F}(\hat{\pi}) - \mathcal{Y}_{\tilde{P}}(\hat{\pi},\epsilon)\\ &\leq L_{\text{exp}}(\pi^*) + \frac{\sqrt{L_{\text{var}}(\pi^*)}}{\sqrt{1-\epsilon}} - L_{\text{exp}}(\hat{\pi}) + \frac{\sqrt{L_{\text{var}}(\hat{\pi})}}{\sqrt{\epsilon}}\\ &= o(\frac{1}{\sqrt{\epsilon M^*}})\ . \qquad \square\end{aligned}$$

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 that 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 in order 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 $S\times A$.

In order 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 policy that have support strictly on state-action pairs in $\mathcal{B}^c$, let:

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

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

Similarly, let:

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

be the lowest achievable expected return given that one starts in state i, 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 $\lambda\geq 0$ satisfy the condition:*

$$Q_+(i,a;\mathcal{B}^c)\;\leq\;\max_{b:(i,b)\in\mathcal{B}^c}Q_-(i,b;\mathcal{B}^c)+\lambda\quad\forall(i,a)\in\mathcal{B}\;\;. \tag{14}$$

Then for state transition observations $\{(s_1,a_1,s_1'),...,(s_M,a_M,s_M')\}$ and $\epsilon\in(0,0.5]$, the policy obtained solving Problem 12 is $o(1/\sqrt{\epsilon M^{**}}+\lambda/(1-\alpha))$ optimal according to Problem 13, 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 do 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 be defined in terms of high probability bounds.

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 are 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 in order to present clearly how our method compares to the nominal and the robust approaches described in Section 2. In fact, our methods remain computationally tractable with machine replacement problems of more than 1000 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 follow: 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 are presented in Figure 2.[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 10000 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 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: $R1$ being normal repairs on the machine costing 2, and $R2$ standing 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

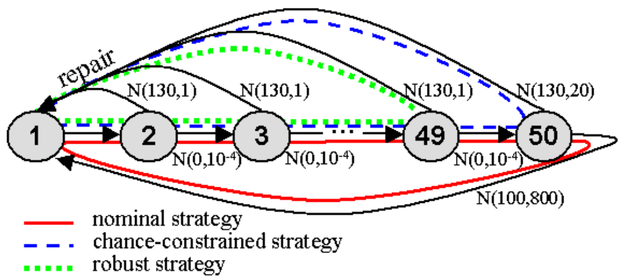


Figure 1 Instance of a machine replacement problem with fixed uncertainty in the rewards. The optimal paths followed for three strategy criterion are drawn.

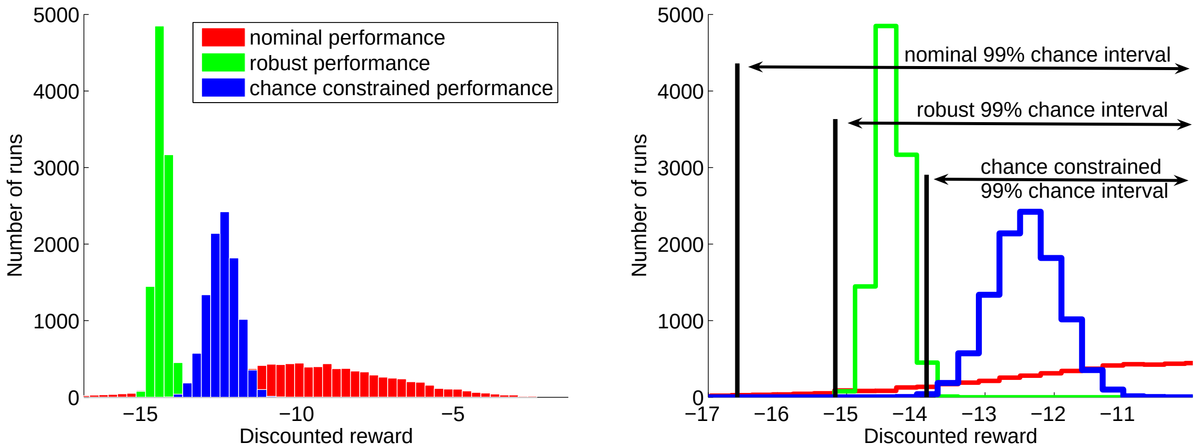


Figure 2 Performance comparisons between nominal, robust and chance constrained policies on 10000 runs of the machine replacement problem. The right figure focuses on the interval $[-17,-10]$.

method presented in Section 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 10000 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 $A_{(i,a,j)} \leq P_{(i,a,j)} \leq B_{(i,a,j)}$ containing 90% of the samples drawn from $\tilde{P}$. We do not discuss the validity of this method as it is purely illustrative of the difficulties involved in the choice of an 90% uncertainty set for $\tilde{P}$.) Finally, we used the “2nd order approximation” performance measure presented in Section 4.3 to find an optimal policy for this machine replacement problem. To do so, we were required to solve a non-convex optimization problem using a gradient descent algorithm (applied on $-\mathbb{F}(\pi)$). The gradient of $\mathbb{F}(\pi)$ was found to be

$$\frac{\partial \mathbb{F}(\pi)}{\partial \pi_{(i,a)}} = \sum_{k,l} \left(q_k r_l + \alpha^2 q_k (\Pi_{(l,\cdot,\cdot)} Q X^\pi r) + \alpha^2 (q^\mathsf{T} X^\pi \Pi Q_{(\cdot,\cdot,k)}) r_l\right) \frac{\partial X^\pi_{(k,l)}}{\partial \pi_{(i,a)}} + \alpha^2 (q^\mathsf{T} X^\pi_{(\cdot,i)})(Q_{(i,a,\cdot)} X^\pi r) + \alpha^2 \sum_{k,a',l} (q^\mathsf{T} X^\pi_{(\cdot,k)})(X^\pi_{(l,\cdot)} r)\pi_{(k,a')} \frac{\partial Q_{(k,a',l)}}{\partial \pi_{(i,a)}} \,,$$

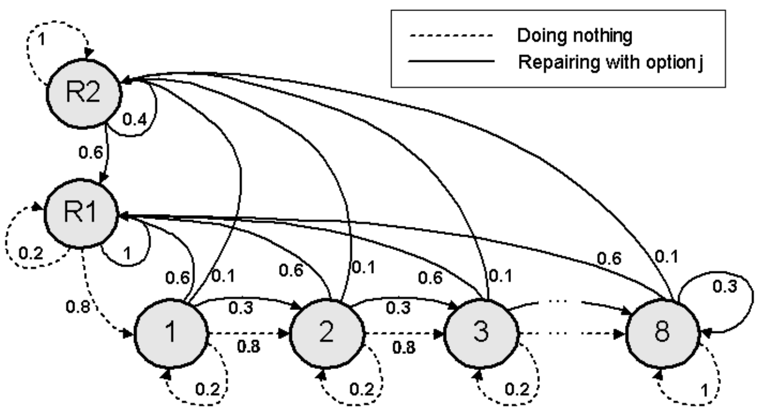
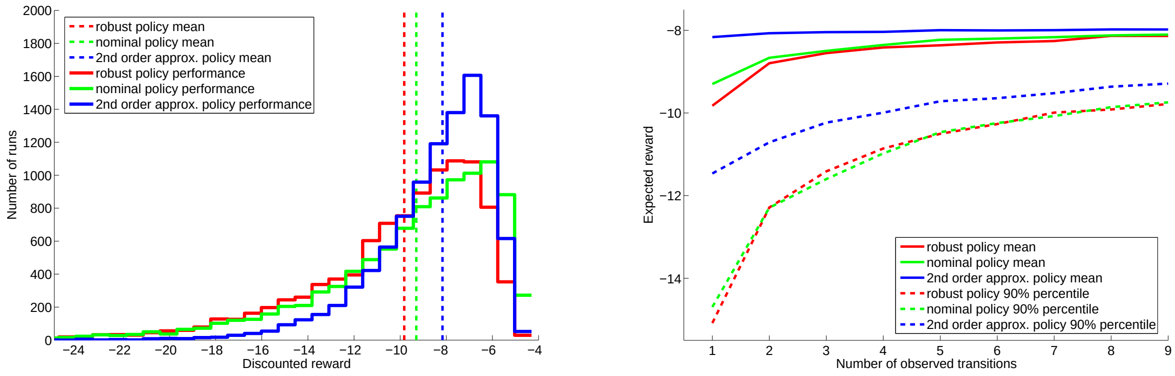


Figure 3 Instance of a machine replacement problem with Dirichlet uncertainty in the transition parameters. The graph presents the expected transition probabilities for the two types of actions (repairing, or not) after observing M transitions from each state. In our experiments, three repair options are available, all three leading to slightly perturbed version of the Dirichlet model presented here.



(a) (b)

Figure 4 (a) presents a performance comparisons between nominal, robust and chance constrained policies on 10000 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.

where

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

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 10000 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 2nd 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 3 repair options are available. The 2nd order approximation method returns a policy that for instance uses a mixed strategy over the repair options in states $R1$ (*i.e.*, heavy repair state) in order 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 2nd 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 presented a "chance constrained formulation" for MDPs with uncertain parameters. We showed that, although some of its instances are intractable to solve, some instances of this problem can be efficiently solved using second-order cone programming. In fact, our experiments demonstrated 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 towards 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.

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 $\Delta\tilde{P}_{(i,.)} = P_{(i,.)} - \hat{P}_{(i,.)}$ 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. (2006).

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*

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

and Objective (4a) *replaced with*

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

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.*

Notes

[1] In our implementation, we used a toolbox developed for Matlab: "CVX: Matlab Software for Disciplined Convex Programming" by Michael Grant *et al*.

[2] Implementation details: the robust problem was solved using the method presented in Section 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 non-linear optimization problem.

Acknowledgments

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

References

Avrachenkov, K.E., J.A. Filar, M. Haviv. 2002. A survey on singular perturbations of Markov chains and decision processes. E.Feinberg, A. Shwartz, eds., *Handbook of Markov Decision Processes : Methods and Applications*. Kluwer.

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

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

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

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

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

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

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

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Rubin. 2003. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC.

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

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

Iyengar, G. 2002. Robust dynamic programming. *Mathematics of Operations Research* .

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

MacQueen, J. 1966. A modified dynamic programming method for markov decision problems. *Journal of Mathematical Analysis and Applications* **14** 28–43.

Mannor, S., D. Simester, P. Sun, J. N. Tsitsiklis. 2006. Bias and variance in value function estimation. *Management Science,* In press .

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

Nemirovski, A., A. Shapiro. 2006. Convex approximations of chance constrained programs. *SIAM Journal on Optimization,* In press .

Nilim, A., L. El Ghaoui. 2005. Robust Markov decision processes with uncertain transition matrices. *Operations Research* .

Prékopa, A. 1995. *Stochastic Programming*. Kluwer Academic Publishers.

Putterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley.

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

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

Silver, E. A. 1963. Markovian decision processes with uncertain transition probabilities or rewards. Technical Report 1, Operations Research Center, MIT.

van der Schaft, A. J. 1999. *L_2-gain and Passivity techniques in Non-linear Control*. Springer-Verlag.

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