

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 tradeoff 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 will be shown to be efficiently solvable and others highly intractable. In each case, we will 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's consider a concrete problem where one needs 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 asses the probability of a machine breaking down at a given aging stage. One should expect significant improvements from incorporating this incertitude in the evaluation of a repair policy's performance. This example illustrates the need for criteria that address parametric uncertainty in general and specifically in the MDPs \(*e.g.*, Ben-Tal and Nemirovski \(1998\), Silver \(1963\), Martin \(1967\), Satia and Lave \(1973\)\).](#)

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 parametric uncertainty in the form of perturbations of the underlying Markov chain but these focus 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 parametric 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 both the Bayesian points of view on the question of decision-making when faced with this extra layer of uncertainty in the MDP model. In fact, it will be shown that both frameworks can lead to a performance measure called the percentile criterion, which is both conceptually natural and representative of the tradeoff between optimistic and pessimistic strategies when facing parametric 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 reduces to 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 exists 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. We close this article with encouraging results on the application of the percentile criterion to cost-effective exploration.

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 worse case scenario. The purpose of our research is to focus on a “tempered” attitude that will realistically tradeoff 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 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:

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

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

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

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

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

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 $R = \{r | f(r) \neq 0\}$ and $P = \{P | f(P) \neq 0\}$ is no longer pertinent (e.g., if $\tilde{r} \propto \mathcal{N}(\mu_{\tilde{r}}, \Sigma_{\tilde{r}})$, then $R = \mathbb{R}^{|S|}$ 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:

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

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

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 of the initial state q . 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 influence 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 initially 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 extension of the percentile criterion do follow naturally and is summarized in Appendix A. 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 . Also, $\mathbb{1}$ will refer to the indicator operator over a deterministic clause such that $\mathbb{1}\{v\} = 1 \iff v$ is true. 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 3d 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 $\tilde{r} \propto \mathcal{N}(\mu_{\tilde{r}}, \Sigma_{\tilde{r}})$. 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 reduced although there exist distributions over the parameters for which the percentile problem becomes intractable.

3.1. Reward uncertainty with Gaussian distribution

This 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 $\Sigma_{\tilde{r}}$ is known and use a Gaussian prior, with parameters (μ_0, Σ_0) , over $\mu_{\tilde{r}}$. Then, based on new independent samples $\{r_1, r_2, \dots, r_m\}$ 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):

$$\mu_1 = \Sigma_1 \left(\Sigma_0^{-1} \mu_0 + \Sigma^{-1} \sum_{i=1}^m r_i \right), \quad \Sigma_1 = (\Sigma_0^{-1} + m \Sigma^{-1})^{-1}.$$

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

$$\mathbb{P}(x^\top \xi \leq 0) \geq p$$

is the same as those satisfying

$$x^\top \mu_\xi + \Phi^{-1}(p) \sqrt{x^\top \Sigma_\xi x} \leq 0,$$

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

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

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

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

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

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

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

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

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

Proof of Theorem 1 We first use the fact that with fixed reward uncertainty Constraint (3b) can be expressed in the form

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

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

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

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

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

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:

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

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 information 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 of Theorem 2 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 computation 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. \square

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 Appendix B, 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 transition parameters uncertainty

We now focus on the problem of transition uncertainty. This type of uncertainty is especially 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 will investigate the chance constrained MDP problem with transition uncertainty

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

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

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 $\mathbb{E}_x(\cdot \mid x_0 \propto q, \pi)$ 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 Appendix C.

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 samples, one can easily evaluate a posterior distribution over these parameters. More specifically, for a vector of transition parameters $\tilde{p} = (\tilde{p}_1, \dots, \tilde{p}_N)$, 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 samples $\{j^{(1)}, j^{(2)}, \dots, j^{(M)}\}$ from the multinomial distribution $f_j(j) = p_j$, one can analytically resolve the posterior distribution over \tilde{p} . This distribution conveniently takes the same Dirichlet form $f(p|j^{(1)}, j^{(2)}, \dots, j^{(M)}) = (1/Z(\beta, M_1, \dots, M_N)) \prod_{j=1}^N p_j^{\beta_j + M_j - 1}$, 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):

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

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

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

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 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.**

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

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

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

This is under the assumption that the rows of \tilde{P} are independent from each other and using $\Sigma^{(i,a)}$ 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

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

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

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

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

$$\underset{y \in \mathbb{R}}{\text{maximize}} \quad y \tag{11a}$$

$$\text{subject to } \mathbb{P}(\tilde{z} \geq y) \geq 1 - \epsilon, \tag{11b}$$

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

The proof is given in Appendix D. One can now derive the following theorem.

THEOREM 4. *Given state transition samples $\{(s_1, a_1, s'_1), \dots, (s_M, a_M, s'_M)\}$ and suppose that $M_{i,a}^{a*} = \min_{i,a} M^{(i,a)}$, **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 $o(1/\sqrt{\epsilon M_{i^*}^{a^*}})$ optimal according to the chance constrained MDP problem

$$\begin{aligned} & \underset{y \in \mathbb{R}, \pi \in \Upsilon}{\text{maximize}} && y \end{aligned} \tag{13a}$$

$$\text{subject to } \mathbb{P}_{\tilde{P}}(\mathbb{E}_x(\sum_{t=0}^{\infty} \alpha^t r(x_t) | x_0 \propto q, \pi) \geq y) \geq 1 - \epsilon, \tag{13b}$$

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 of Theorem 4 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

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

and

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

where

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

where the bounds $o(\frac{1}{(M_{i^*}^{a^*})^2})$ and $o(\frac{1}{M_{i^*}^{a^*}})$ 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.

$$\begin{aligned} \mathcal{Y}_{\tilde{P}}(\pi^*, \epsilon) - \mathcal{Y}_{\tilde{P}}(\hat{\pi}, \epsilon) &= \mathcal{Y}_{\tilde{P}}(\pi^*, \epsilon) - \mathbb{F}(\pi^*) + \mathbb{F}(\pi^*) - \mathcal{Y}_{\tilde{P}}(\hat{\pi}, \epsilon) \\ &\leq \mathcal{Y}_{\tilde{P}}(\pi^*, \epsilon) - \mathbb{F}(\pi^*) + \mathbb{F}(\hat{\pi}) - \mathcal{Y}_{\tilde{P}}(\hat{\pi}, \epsilon) \\ &\leq L \exp(\pi^*) + \frac{\sqrt{L \text{var}(\pi^*)}}{\sqrt{1-\epsilon}} - L \exp(\hat{\pi}) + \frac{\sqrt{L \text{var}(\hat{\pi})}}{\sqrt{\epsilon}} \\ &= o\left(\frac{1}{\sqrt{\epsilon M_{i^*}^{a^*}}}\right). \quad \square \end{aligned}$$

5. Experiments

We have chosen the machine replacement problem as a good application for our methods. Let's assume that we are interested in 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 have shown to 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.

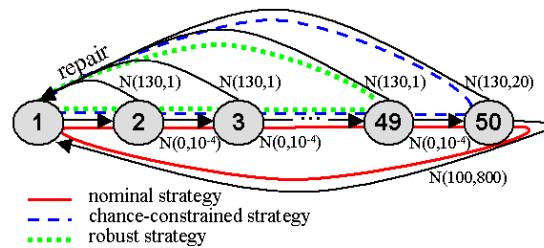


Figure 1 Instance of a machine replacement problem with fixed uncertainty in the rewards. The optimal paths followed for three strategy criterion are drawn.

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

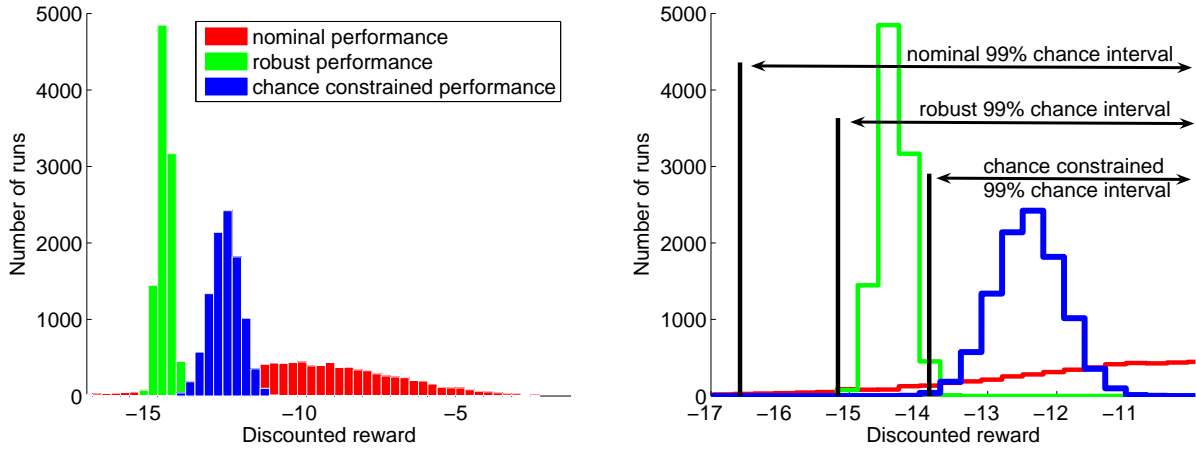


Figure 2 Performance comparisons between nominal, robust and chance constrained policies on 10000 runs of the machine replacement problem. The right figure focuses on the interval $[-17, -10]$.

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.

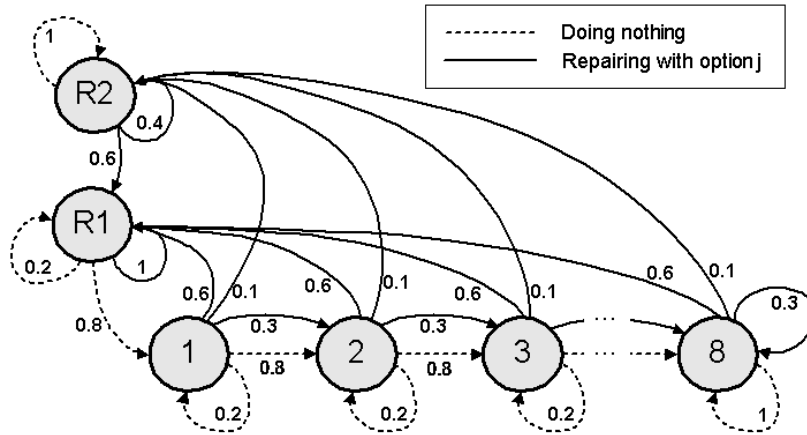


Figure 3 Instance of a machine replacement problem with Dirichlet uncertainty in the transition parameters. The graph presents the expected transition probabilities for the two types of actions (repairing, or not) after observing M transitions from each state. In our experiments, three repair options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

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 Section 2.2. As mentioned earlier, it is unclear how to state the robust MDP problem when using probabilistic models for parametric 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

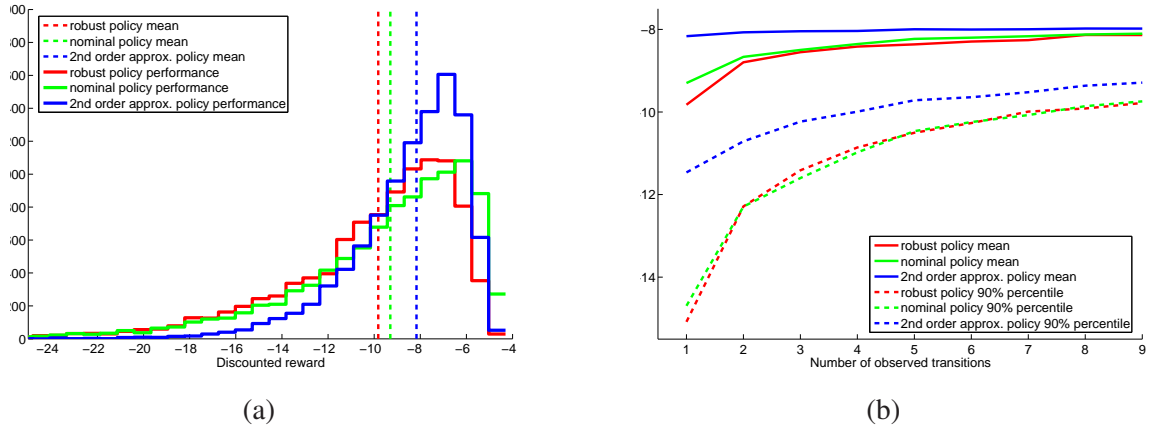


Figure 4 (a) presents a performance comparisons between nominal, robust and chance constrained policies on 10000 runs of the machine replacement problem with $M = 1$. (b) presents the effect of decreasing the uncertainty in the transitions on the mean and the 90% percentile performances of the different methods.

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

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

where

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

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 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 parametric 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 uses, in states 8 and R1, a mixed strategy over these repair options in order to reduce the transition variance 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. Cost-effective exploration with the percentile criterion

In many practical situations, one has the possibility of investing (money, time or computation efforts) in actions that will reduce one’s uncertainty in the model. This gives rise to the so-called exploration-exploitation dilemma, one of the most studied issues in reinforcement learning. In a more popular “online” version of this problem, an agent must decide at each point of time between actions with known return or actions with unknown return but with the potential of even better return. Exploration methods range from slowly converging ones such as ϵ -greedy exploration to better behaved ones, such as regret minimization (see Hazan et al. (2006)) and model based interval estimation (see Strehl and Littman (2005)) which is for instance known to lead with high-probability to near-optimal policies in polynomial time. Because the state space of these problems is typically large, we formulate the exploration problem differently. We assume that, before committing to an exploitation strategy (such as repair policies for the problems described in Sections 5.1 and 5.2), one has the option to buy observations of the reward vector (or of transitions) for any state and action pair (i, a) of the system. In this context, a valid exploration strategy needs to provide either a pair (i, a) that it wishes to observe or commit to a full exploitation strategy for the system. We believe that this framework is particularly well suited for problems of short horizon compared to the size of the state space.

In order to provide guidance in this decision, we apply the concept of value of information (see Howard (1966)) to the “percentile framework”. Given a probabilistic prior on the model parameters \tilde{r} and \tilde{P} , and a risk-sensitive measure of return $\mathcal{G}(\pi, \tilde{r}, \tilde{P})$ for stationary policies $\pi \in \Upsilon$, we define the value of sampling \tilde{r} and \tilde{P} at (i, a) as

$$\mathcal{V}(i, a) = \mathbb{E} \left(\max_{\pi'} \mathcal{G}(\pi', \tilde{r}', \tilde{P}') \right) - \max_{\pi} \mathcal{G}(\pi, \tilde{r}, \tilde{P}), \quad (14)$$

where \tilde{r}' and \tilde{P}' are the posterior distribution of \tilde{r} and \tilde{P} respectively given random reward and transition samples from state i with action a , and the expectation is taken over the prior distribution of reward and transition parameters. Intuitively, $\mathcal{V}(i, a)$ gives the expected increase in return given that one would know more about the parameters related to (i, a) . The learning strategy we propose selects $(i, a)^* = \arg \max \mathcal{V}(i, a)$ as the most cost effective location for a new observation, and decides to stop investing in uncertainty reduction when the maximum $\mathcal{V}(i, a)$ achievable is smaller than the observation cost.

In the case of uncertainty limited to the reward parameters, we can imagine the scenario where one has the option of buying noisy measurements of these rewards. Here, $\mathcal{V}(i, a)$ is the value of using an extra sample in the modeling $\tilde{r}(i, a)$. Assuming Gaussian measurement noise and a Gaussian prior to represent the uncertainty in $r(i, a)$, one can easily solve the percentile problem (see Section 3.1) to find an optimal risk-sensitive policy, the question is: is it worth buying more information about the MDP before committing to a policy of this form? Given a measurement $\hat{r}(i, a)$ and a prior distribution on $\tilde{r}(i, a) \propto \mathcal{N}(\mu_{(i,a)}, \sigma_{(i,a)}^2)$, we can evaluate the posterior distribution $\tilde{r}'(i, a) \propto \mathcal{N}(\mu'_{(i,a)}, \sigma'^2_{(i,a)})$.⁴ The value of information $\mathcal{V}(i, a)$, with $\mathcal{G}(\pi, \tilde{r})$ set as the optimal value of percentile Problem 4, can therefore be estimated using Monte Carlo methods. To reduce computation, our approach relies on computing a lower bound for $\mathcal{V}(i, a)$ by evaluating $\mathcal{V}(i, a) = \mathbb{E}(\mathcal{G}(\pi^*, \tilde{r}'_{\hat{r}(i,a)})) - \max_{\pi} \mathcal{G}(\pi, \tilde{r})$, where $\pi^* = \arg \max_{\pi} \mathcal{G}(\pi, \tilde{r})$. It turns out that this approximation for $\mathcal{V}(i, a)$ can be computed in closed-form given π^* :

$$\begin{aligned} \mathcal{V}(i, a) &= \mathbb{E}(\mathcal{G}(\pi^*, \tilde{r}'_{\hat{r}(i,a)})) - \mathcal{G}(\pi^*, \tilde{r}) \\ &= \mathbb{E} \left(\sum_a \rho_a^{*\top} \mu_{\tilde{r}'} \right) - \Phi^{-1}(\eta) \left\| \sum_a \rho_a^{*\top} \Sigma_{\tilde{r}'}^{\frac{1}{2}} \right\|_2 - \mathcal{G}(\pi^*, \tilde{r}) \\ &= \Phi^{-1}(\eta) \left(\left\| \sum_a \rho_a^{*\top} \Sigma_{\tilde{r}}^{\frac{1}{2}} \right\|_2 - \left\| \sum_a \rho_a^{*\top} \Sigma_{\tilde{r}'}^{\frac{1}{2}} \right\|_2 \right), \end{aligned}$$

since the posterior update for $\sigma(i, a)$ is independent of \hat{r} and since $\mathbb{E}(\mu_{\tilde{r}'}) = \mu_{\tilde{r}}$ for such a Gaussian model. In this framework, the η parameter for the percentile problem studied in Section 3.1 controls how conservative the policy is during the exploitation stage.

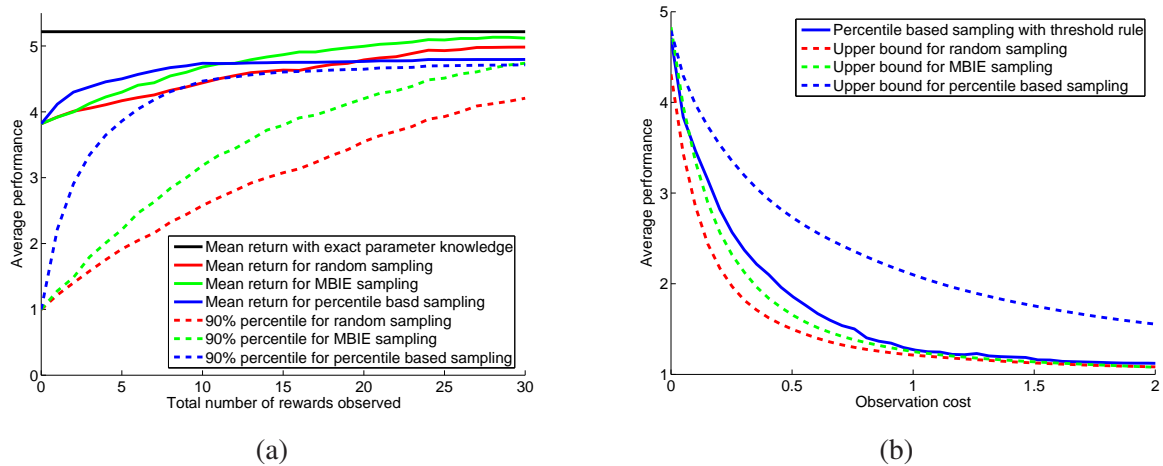


Figure 5 (a) presents the average 90% percentile and mean performances of sampling strategies on a set of 1000 random MDPs with reward uncertainty (free observations). (b) presents the average “effective” return on the MDPs used in (a) for a range of observation costs.

The following experiments compare percentile based sampling to random sampling and the model based interval estimation (MBIE) strategy⁵ on a set of 1000 randomly generated MDPs with reward uncertainty. Each model has 10 states, 2 actions, a discount factor of 0.8, initial reward uncertainty $\tilde{r}(i, a) \propto \mathcal{N}(\mu_{\tilde{r}}(i, a), 1)$ and measurement noise $\nu \propto \mathcal{N}(0, 1)$. For a given model, each (i, a) has a deterministic transition drawn uniformly from the set of states and has $\mu_{\tilde{r}}(i, a)$ drawn from $\mathcal{N}(0, 1)$. Figure 5(a) presents the average 90% percentile and average mean performances over this set of uncertain MDPs given a number of observations chosen by the different strategies (no observation cost). In each run, once a strategy ran out of observations, the posterior uncertainty \tilde{r}' was computed and used to evaluate the mean and percentile return of the strategy through optimizing the nominal problem and the percentile MDP problem given the uncertain reward \tilde{r}' . We note that the percentile strategy clearly outperforms both MBIE and random sampling for percentile return and, when restricted only to a small number of observations, even in terms of mean returns. Figure 5(b) shows the average “effective” return of our learning strategy given different observation prices on the set of randomly generated MDPs. The effective return of a run was considered to be the final exploitation percentile return from which was deducted the incurred sampling cost for exploration. Since MBIE and random sampling do not provide a stopping criterion for exploration, average total percentile cost cannot be directly evaluated for them. Instead, we computed a lower bound on this performance by selecting in each run, given the observation cost, the most profitable point to start exploitation. We see that the percentile criterion based strategy outperforms even this performance bound for both random and MBIE sampling.

7. 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, this problem can also take forms that are 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, can now 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. Moreover, using sensitivity analysis, we

showed that the formulation can be used to guide exploration and determine where additional sampling should occur in order to have the highest potential impact on optimal long-term reward.

Appendix A: 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 simple 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 $\Delta\hat{P}_{(i,\cdot)} = P_{(i,\cdot)} - \hat{P}_{(i,\cdot)}$ 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 $\Delta\hat{P}$ 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 $(\mu_{\tilde{r}}, \Sigma_{\tilde{r}})$ with high probability, a rigorous interpretation of the percentile criterion (called distributionally robust) can enforce the chance constraint to be met over the set $\mathcal{F}_{\mu_{\tilde{r}}, \Sigma_{\tilde{r}}}$ 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 2. *Given that \tilde{r} is drawn from a distribution in the set $\mathcal{F}_{\mu_{\tilde{r}}, \Sigma_{\tilde{r}}}$, Theorem 1 holds with Chance Constraint (3b) replaced with the **distributionally robust** Chance Constraint*

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

and Objective (4a) replaced with

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

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.

Appendix B: Proof of Theorem 3

We reduce the NP-complete 3SAT problem to solving Problem 3 with general reward uncertainty in the reward parameters.

3SAT Problem: Let W be a collection of disjunctive clauses $W = \{w_1, w_2, \dots, w_M\}$ on a finite set of variables $V = \{v_1, v_2, \dots, v_N\}$ such that $|w_m| = 3 \forall m \in \{1, \dots, M\}$. Let each clause be of the form $w = v_i \vee v_j \vee \bar{v}_k$, where \bar{v} is the negation of v . Is there a truth assignment for V that satisfies all the clauses in W ?

Given an instance of the 3SAT Problem, we can construct an MDP with uncertainty in the rewards such that 3SAT is satisfiable if and only if the optimal value for the chance-constrained MDP Problem 3 is greater than 0. After describing a 2-action uncertain MDP and its associated chance constraint problem in

Step 1 and 2, we will demonstrate, in Step 3, that feasible policies must necessarily be deterministic on a set of states. In Step 4, this fact will be used to build from such a feasible policy an assignment for the variables that satisfies all the clauses of the 3SAT problem. Step 5 will confirm that if the original 3SAT problem is satisfiable then the constructed chance-constrained problem has a feasible solution. The final step demonstrates that the transformation involved can be done in polynomial time with a polynomial amount of memory.

STEP 1. Let $W = \{w_1, w_2, \dots, w_M\}$ and $V = \{v_1, v_2, \dots, v_N\}$ be an instance of 3SAT, we first create a set of $N + 1$ states $S_0 = \{s_{(0,0)}, s_{(1,0)}, \dots, s_{(N,0)}\}$ with no reward and with two actions a_1, a_2 available. Then we create two sets of states $S_1 = \{s_{(0,1)}, s_{(1,1)}, \dots, s_{(N,1)}\}$ and $S_2 = \{s_{(0,2)}, s_{(1,2)}, \dots, s_{(N,2)}\}$ for which the rewards will be uncertain and, finally, we create an absorption state s_3 with no reward. The transition matrix for our uncertain MDP is described in Figure 6. Specifically, if action a_j is taken in state $s_{(i,0)}$, the system transitions to state $s_{(i,j)}$ in the next time step. All remaining states ($s \in S_1 \cup S_2$) lead to s_3 for all actions. We set the initial distribution q to be uniform over the states in S_0 , and the discount factor to any value $\alpha > 0$. It remains to describe the uncertainty over the rewards for states $S_1 \cup S_2$.

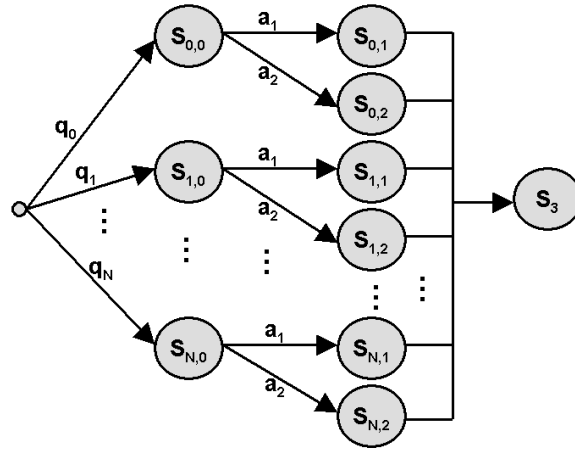


Figure 6 Structure for the MDP with uncertainty in the rewards used in the proof of Theorem 3

We create a discrete probability function for the reward matrix $\tilde{R} \in \mathbb{R}^{N \times 2}$, with $\tilde{R}_{(i,j)} = \tilde{r}(s_{(i,j)})$. To each variable v_n , $n \in \{1, \dots, N\}$, we associate two events ($E^{(v_n,1)}$ and $E^{(v_n,2)}$) that each have probability $0.25/N$ of occurring. If event $E^{(v_n,1)}$ occurs, it leads to using the reward matrix $R^{(v_n,1)}$, while event $E^{(v_n,2)}$ leads to using $R^{(v_n,2)}$. The total probability of events $\{E^{(v_1,1)}, E^{(v_1,2)}, \dots, E^{(v_N,1)}, E^{(v_N,2)}\}$ is $1/2$. Next, for each clause w_m in the original 3SAT problem, we create an event $E^{(w_m)}$ that occurs with probability $0.5/M$. Drawing $E^{(w_m)}$ leads to using the reward matrix $R^{(w_m)}$. The matrices $R^{(v_n,1)}, R^{(v_n,2)}$ and $R^{(w_m)}$ are described as follow

$$R_{(i,j)}^{(v_n,1)} = \begin{cases} -1 & \text{if } i = n \text{ and } j = 1 \\ 0 & \text{otherwise} \end{cases}, \quad R_{(i,j)}^{(v_n,2)} = \begin{cases} -1 & \text{if } i = n \text{ and } j = 2 \\ 0 & \text{otherwise} \end{cases},$$

$$R_{(0,1)}^{(w_m)} = -1, \quad R_{(0,2)}^{(w_m)} = -1, \\ R_{(i,1)}^{(w_m)} = \begin{cases} 1 & \text{if } v_i \in w_m \\ 0 & \text{otherwise} \end{cases} \quad \forall 1 \leq i \leq N, \quad R_{(i,2)}^{(w_m)} = \begin{cases} 1 & \text{if } \neg v_i \in w_m \\ 0 & \text{otherwise} \end{cases} \quad \forall 1 \leq i \leq N.$$

where $\neg v$ means negating the boolean variable. It is clear that these events form a distribution

$$\sum_{n=1}^N (\mathbb{P}(E^{(v_n,1)}) + \mathbb{P}(E^{(v_n,2)})) + \sum_{m=1}^M \mathbb{P}(E^{(w_m)}) = 1.$$

STEP 2. We will concentrate on the feasibility of

$$\mathbb{P}_{\tilde{r}} \left(\mathbb{E}_x \left(\sum_{t=0}^{\infty} \alpha^t \tilde{r}_t(x_t) \mid x_0 \propto q, \pi \right) \geq 0 \right) \geq 0.75, \quad (15)$$

which is obviously equivalent to showing that Problem 3's optimal value is higher than 0 when $\epsilon = 0.25$. With the constructed uncertain MDP, this constraint is equivalent to

$$\sum_{n=1}^N \mathbb{P}(E^{(v_n,1)}) (\mathbb{1}^{(v_n,1)} + \mathbb{1}^{(v_n,2)}) + \sum_{m=1}^M \mathbb{P}(E^{(w_m)}) \mathbb{1}^{(w_m)} \geq 0.75,$$

where we used the fact that $\mathbb{P}(E^{(v_n,2)}) = \mathbb{P}(E^{(v_n,1)})$ and where we made the following substitutions

$$\begin{aligned} \mathbb{1}^{(v_n,1)} &= \mathbb{1} \left\{ \mathbb{E}_x \left(\sum_{t=0}^{\infty} \alpha^t \tilde{r}_t(x_t) \mid q, \pi, R^{(v_n,1)} \right) \geq 0 \right\} \\ \mathbb{1}^{(v_n,2)} &= \mathbb{1} \left\{ \mathbb{E}_x \left(\sum_{t=0}^{\infty} \alpha^t \tilde{r}_t(x_t) \mid q, \pi, R^{(v_n,2)} \right) \geq 0 \right\} \\ \mathbb{1}^{(w_m)} &= \mathbb{1} \left\{ \mathbb{E}_x \left(\sum_{t=0}^{\infty} \alpha^t \tilde{r}_t(x_t) \mid q, \pi, R^{(w_m)} \right) \geq 0 \right\}. \end{aligned}$$

STEP 3. In the described uncertain MDP, we now outline why a policy that is feasible according to Constraint (15) must be deterministic for states $\{s_{(1,0)}, \dots, s_{(N,0)}\}$. First we show that $\mathbb{1}^{(v_n,1)} + \mathbb{1}^{(v_n,2)}$ is 1 if the policy for state $s_{(n,0)}$ is deterministic and 0 otherwise. For any v_n in this MDP,

$$\begin{aligned} \mathbb{1}^{(v_n,1)} &= \mathbb{1} \left\{ \mathbb{E}_x \left(\sum_{t=0}^{\infty} \alpha^t \tilde{r}_t(x_t) \mid q, \pi, R^{(v_n,1)} \right) \geq 0 \right\} \\ &= \mathbb{1} \left\{ 0 + \sum_{i=0}^N q(i) \alpha \left(\pi(i,1) R_{(i,1)}^{(v_n,1)} + \pi(i,2) R_{(i,2)}^{(v_n,1)} \right) + \sum_{t=2}^{\infty} 0 \geq 0 \right\} \\ &= \mathbb{1} \left\{ \frac{\alpha}{N+1} \sum_{i=1}^N -\mathbb{1}\{i = n \wedge 1 = 1\} \pi(i,1) - \mathbb{1}\{i = n \wedge 1 = 2\} \pi(i,2) \geq 0 \right\} \\ &= \mathbb{1} \left\{ \frac{-\alpha \pi(n,1)}{N+1} \geq 0 \right\} \\ &= \mathbb{1} \{ \pi(n,1) = 0 \} = \mathbb{1} \{ \pi(n,2) = 1 \}, \end{aligned}$$

where we started by expanding the expectation term, then used the definition of $R^{(v_n,1)}$, and finally the fact that $\alpha > 0$, $\pi(n,1) > 0$, and that $\pi(n,1) + \pi(n,2) = 1$. By symmetry, it is also clear that $\mathbb{1}^{(v_n,2)} = \mathbb{1} \{ \pi(n,1) = 1 \}$.

It remains to show that, because $\mathbb{1}^{(v_n,1)} + \mathbb{1}^{(v_n,2)} \leq 1$, Constraint (2) can only be met with equality, which occurs if and only if the policy is deterministic for states $\{s_{(1,0)}, \dots, s_{(N,0)}\}$.

$$\begin{aligned} \sum_{n=1}^N \mathbb{P}(E^{(v_n,1)}) (\mathbb{1}^{(v_n,1)} + \mathbb{1}^{(v_n,2)}) + \sum_{m=1}^M \mathbb{P}(E^{(w_m)}) \mathbb{1}^{(w_m)} &\geq 0.75 \\ \Leftrightarrow \frac{0.25}{N} \sum_{n=1}^N \mathbb{1} \{ \pi(n, \cdot) \text{ is deterministic} \} + \frac{0.5}{M} \sum_{m=1}^M \mathbb{1}^{(w_m)} &\geq 0.75 \\ \Leftrightarrow (\pi(n, \cdot) \text{ is deterministic}) \wedge (\mathbb{1}^{(w_m)} = 1), \forall n \in \{1, \dots, N\}, \forall m \in \{1, \dots, M\}. \end{aligned}$$

STEP 4. With the described uncertain MDP, given a policy that is feasible according to Constraint (15), the assignment $v_i = \mathbb{1}\{\pi(i, 1) = 1\}$ for the variables in V satisfies all the clauses in W . We already showed in Step 3, that for a policy π to be feasible according to Constraint (15), it must be deterministic for states $\{s_{(1,0)}, \dots, s_{(N,0)}\}$ and satisfy $(\mathbb{1}^{(w_m)} = 1)$ for all $m \in \{1, \dots, M\}$. Now given a clause $w \in W$, for example $w_1 = v_1 \vee v_2 \vee \neg v_3$, then by construction of event $E^{(w_1)}$,

$$\begin{aligned} (\mathbb{1}^{(w_m)} = 1) &\leftrightarrow \mathbb{E}_x \left(\sum_{t=0}^{\infty} \alpha^t \tilde{r}_t(x_t) | q, \pi, R^{(w_1)} \right) \geq 0 \\ &\leftrightarrow 0 + \sum_{i=0}^N q(i) \alpha \left(\pi(i, 1) R_{(i,1)}^{(w_1)} + \pi(i, 2) R_{(i,2)}^{(w_1)} \right) + \sum_{t=2}^{\infty} 0 \geq 0 \\ &\leftrightarrow \frac{\alpha}{N+1} (-1 + \pi(1, 1) \cdot 1 + \pi(2, 1) \cdot 1 + \pi(3, 2) \cdot 1) \geq 0 \\ &\leftrightarrow \pi(1, 1) + \pi(2, 1) + \pi(3, 2) \geq 1 \\ &\rightarrow (\pi(1, 1) = 1) \vee (\pi(2, 1) = 1) \vee (\pi(3, 1) = 0) \\ &\rightarrow v_1 \vee v_2 \vee \neg v_3, \end{aligned}$$

given that π is deterministic. This can be shown for any clause $w \in W$ and it allows us to conclude that, given that the optimal value of Problem 3 is greater or equal to 0, the 3SAT problem is satisfiable and we can construct a satisfying assignment from the optimal point of the optimization problem.

STEP 5. It is also easy to show that if the optimal value for Problem 3 is smaller than 0, then there is no satisfying assignment for the 3SAT problem. This will be demonstrated by showing that given any satisfying assignment for the variables in V , there exists a policy π that is feasible according to Constraint (15). Using the satisfying assignment for the variables in V , we test the feasibility of policy

$$\begin{aligned} \pi(0, 1) &= 1 \\ \pi(i, 1) &= \mathbb{1}\{v_i\} \quad \forall i \in \{1, \dots, N\} \\ \pi(i, 2) &= \mathbb{1}\{\neg v_i\} \quad \forall i \in \{1, \dots, N\}, \end{aligned}$$

which is obviously a valid deterministic policy. But also, since, for example, clause $w_1 = (v_1 \vee v_2 \vee \neg v_3)$ is satisfied by the variable assignment, then $\pi(1, 1) + \pi(2, 1) + \pi(3, 2) \geq 1$ is necessarily satisfied. Thus, $\mathbb{1}^{(w_m)} = 1$ is satisfied for all $w_m \in W$. From the statements presented in steps 3 and 4, we get that Constraint (15) is satisfied by π .

$$\mathbb{P}_{\tilde{r}} \left(\mathbb{E}_x \left(\sum_{t=0}^{\infty} \alpha^t \tilde{r}_t(x_t) | q, \pi \right) \geq 0 \right) = \frac{0.25}{N} \sum_{n=1}^N \mathbb{1}\{\pi(n, \cdot) \text{ is deterministic}\} + \frac{0.5}{M} \sum_{m=1}^M \mathbb{1}^{(w_m)} = 0.75$$

STEP 6. The uncertain MDP that is used to solve the 3SAT problem can be constructed in polynomial time. First, the MDP presented in Figure 6 has $|S| = 3(N + 1) + 1$ states and $|A| = 2$ actions, the transition matrix has therefore $2|S|^2$ entries which are either 0 or 1. Then, each of the $2N + M$ events can be described by its probability, $0.25/N$ or $0.5/M$, and its associated $|S| \times |A|$ reward matrix, the entries of which are either -1 , 0 , or 1 in our construction. Overall, the problem can obviously be constructed in polynomial time and polynomial space. \square

We want to note the fact that the proof did not require the assumption of stationarity in the uncertainty for \tilde{r} , or the stationarity of the policy π . In fact, the proof is valid for both types of uncertainty and strategies (*i.e.*, stationary or non-stationary).

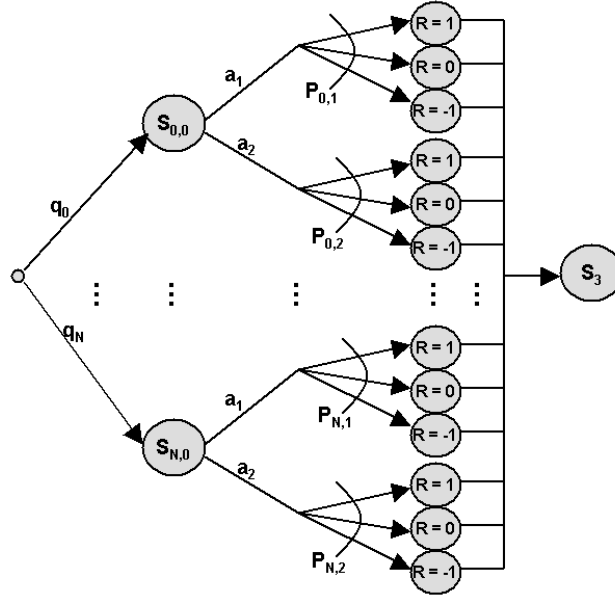


Figure 7 Structure for the MDP with uncertainty in the transitions (\tilde{P}) used in the proof of Corollary 1

Appendix C: Proof of Corollary 1

The proof of Corollary 1 follows similar lines as what was presented in Appendix B for Theorem 3. Given an instance of the NP-complete 3SAT Problem, one can construct in polynomial time the MDP with discrete transition uncertainty presented in Figure 7. Based on this instance of the 3SAT Problem, the same events, $E^{(v_n,1)}$, $E^{(v_n,2)}$ and $E^{(w_m)}$, can be created as described in Appendix B. However, in this proof, drawing $E^{(v_n,1)}$, $E^{(v_n,2)}$ or, $E^{(w_m)}$ will lead to using transition parameters $P^{(v_n,1)}$, $P^{(v_n,2)}$, or $P^{(w_m)}$ respectively in the uncertain MDP. These parameters are defined as follows:

$$P_{(i,j)}^{(v_n,1)}(k) = \begin{cases} 1 & \text{if } r(k) = R_{(i,j)}^{(v_n,1)} \\ 0 & \text{otherwise,} \end{cases} \quad P_{(i,j)}^{(v_n,2)}(k) = \begin{cases} 1 & \text{if } r(k) = R_{(i,j)}^{(v_n,2)} \\ 0 & \text{otherwise,} \end{cases}$$

$$P_{(i,j)}^{(w_m)}(k) = \begin{cases} 1 & \text{if } r(k) = R_{(i,j)}^{(w_m)} \\ 0 & \text{otherwise,} \end{cases}$$

where $R_{(i,j)}^{(v_n,1)}$, $R_{(i,j)}^{(v_n,2)}$ and $R_{(i,j)}^{(w_m)}$ refer to the definitions in Appendix B. Clearly, the resulting MDP from drawing each of these events are equivalent to the MDP instances that was originally associated with that event in Appendix B. Therefore, in the constructed MDP with transition uncertainty \tilde{P} , one can use similar arguments to show that the feasibility of

$$\mathbb{P}_{\tilde{P}} \left(\mathbb{E}_x \left(\sum_{t=0}^{\infty} \alpha^t r_t(x_t) \mid x_0 \propto q, \pi \right) \geq 0 \right) \geq 0.75, \quad (16)$$

is equivalent to the determining the satisfiability of the instance of the 3SAT Problem. \square

Appendix D: Proof of Lemma 2

Take the assignment $y_1 = \mu - \frac{\sigma}{\sqrt{\epsilon}}$, Markov's inequality can be used to show that y_1 is always feasible according to Constraint 11b:

$$P(\tilde{z} \geq y_1) = P(\tilde{z} \geq \mu - \frac{\sigma}{\sqrt{\epsilon}})$$

$$\begin{aligned}
&= P(|\tilde{z} - \mu| \leq \frac{\sigma}{\sqrt{\epsilon}}) + P(\tilde{z} - \mu \geq \frac{\sigma}{\sqrt{\epsilon}}) \\
&\geq 1 - P(|\tilde{z} - \mu| \geq \frac{\sigma}{\sqrt{\epsilon}}) \\
&\geq 1 - \epsilon.
\end{aligned}$$

On the other hand, for any $\delta \in (0, 1 - \epsilon)$, $y_2 = \mu + \frac{\sigma}{\sqrt{(1-\epsilon-\delta)}}$ is on the contrary always assured to be unfeasible:

$$\begin{aligned}
P(\tilde{z} \geq y_2) &= P(\tilde{z} \geq \mu + \frac{\sigma}{\sqrt{(1-\epsilon-\delta)}}) \\
&= 1 - P(|\tilde{z} - \mu| \leq \frac{\sigma}{\sqrt{(1-\epsilon-\delta)}}) - P(\tilde{z} - \mu \leq -\frac{\sigma}{\sqrt{(1-\epsilon-\delta)}}) \\
&\leq P(|\tilde{z} - \mu| \geq \frac{\sigma}{\sqrt{(1-\epsilon-\delta)}}) \\
&\leq 1 - \epsilon - \delta.
\end{aligned}$$

Therefore, $\mu - \frac{\sigma}{\sqrt{\epsilon}} \leq y^* \leq \mu + \frac{\sigma}{\sqrt{1-\epsilon}}$. \square

Notes

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

²Implementation 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/>.

³Implementation details: Matlab’s optimization toolbox was used to solve this non-linear optimization problem.

⁴The posterior updates are $\mu'_{(i,a)} = \sigma'_{(i,a)}(\mu_{(i,a)}/\sigma_{(i,a)} + \hat{r}(i,a)/\sigma_\nu)$ and $\sigma'_{(i,a)} = (\sigma_{(i,a)}^{-1} + \sigma_\nu^{-1})^{-1}$. Note that $\sigma'_{(i,a)}$ is independent of the observed $\hat{r}(i,a)$.

⁵Being an online method, MBIE only provides a rule, given a state, for choosing the action with highest exploration-exploitation potential. To adapt this method to our framework, we first draw a state randomly and then select the action with MBIE.

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.
- 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.

- Hazan, Elad, Adam Kalai, Satyen Kale, Amit Agarwal. 2006. Logarithmic regret algorithms for online convex optimization. *COLT*. 499–513.
- Howard, R., J. Matheson. 1972. Risk-sensitive Markov decision processes. *Management Science* **18**(7) 356–369.
- Howard, R. A. 1966. Information value theory. *IEEE Trans. on Systems Science and Cybernetics* **SSC-2**(1) 22–26.
- 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.
- Mannor, S., D. Simester, P. Sun, J. N. Tsitsiklis. 2006. Bias and variance in value function estimation. *Management Science*, In press .
- Martin, J. 1967. *Bayesian Decision Problems and Markov Chains*. Wiley.
- 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.
- Strehl, A. L., M. L. Littman. 2005. A theoretical analysis of model-based interval estimation. *Proc. ICML*. 857–864.
- van der Schaft, A. J. 1999. *L₂-gain and Passivity techniques in Non-linear Control*. Springer-Verlag.
- Wilks, S. S. 1962. *Mathematical Statistics*. John Wiley & Sons, Inc.