Percentile Optimization in Uncertain Markov Decision Processes with Application to Efficient Exploration

Erick Delage EDELAGE@STANFORD.EDU
Department of Electrical Engineering, Stanford University, Stanford, California, USA

Shie Mannor SHIE.MANNOR@MCGILL.CA
Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada

Abstract

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, learned from experience, or designed by hand, it is not surprising that the actual performance of a chosen strategy often significantly differs from the designer's initial expectations due to unavoidable model uncertainty. In this paper, we present a percentile criterion that captures the trade-off between optimistic and pessimistic points of view on MDP with parameter uncertainty. We describe tractable methods that take parameter uncertainty into account in the process of decision making. Finally, we propose a cost-effective exploration strategy when it is possible to invest (money, time or computation efforts) in actions that will reduce the uncertainty in the parameters.

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, learned from experience or designed by hand, 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., 2007). For this reason, criteria that address parameter uncertainty in general and specifically in MDPs are of interest (*e.g.*, Silver, 1963; Ben-Tal & Nemirovski, 1998).

Appearing in *Proceedings of the 24^{th} International Conference on Machine Learning*, Corvallis, OR, 2007.

To date, most efforts have focused on the study of robust MDPs (*e.g.*, Givan et al., 2000; Iyengar, 2005; Nilim & El Ghaoui,), a framework in which one makes the assumption that parameters can only lie in a bounded uncertainty set. Although this formulation for the MDP problem remains tractable under mild conditions, it suffers from relying on the union bound for bounding probabilistic events[1] and often generates overly conservative strategies.

In this paper we offer a more practical way of handling uncertainty in the parameters. Following some recent work by Mannor et al. (2007) that studied the effect of parameter uncertainty on the mean and variance of value function estimates, we consider the parameters as random variables and take a Bayesian point of view on the question of decision-making when faced with this extra layer of uncertainty in the MDP model. The Bayesian framework naturally leads to a performance measure we call the percentile criterion,[2] which is both conceptually natural and representative of the trade-off between optimistic and pessimistic strategies when facing parameter uncertainty. Unlike robust methods, our approach reasons directly about the effect of this uncertainty on the total cumulative reward itself. This in turn leads to the notion of a cost-effective exploration strategy when given the option to invest in the reduction of this uncertainty.

The percentile criterion (or chance constraint) that is widely studied for single-period optimization problems (*e.g.*, Charnes & Cooper, 1959; Prékopa, 1995; Calafiore & El Ghaoui, 2006) will be generalized in Section 2 to infinite-horizon MDPs. Although general percentile optimization problems are suspected to be "severely computationally intractable" (Nemirovski & Shapiro, 2006),

[1] As the size of the state space grows, one needs to consider larger uncertainty sets for each parameters to accommodate a probabilistic constraint.

[2] 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.

in Section 3 we demonstrate that the problem of reward uncertainty can reduce to a deterministic second order cone program (*c.f.*, Lobo et al., 1998) and that transition uncertainty can be addressed approximately. Section 4 presents a proposed cost-efficient strategy for the exploration-exploitation dilemma in the context of MDPs with non-negligible observation costs and compare its performance against popular exploration schemes.

2. Background

In the context of an MDP with parameter uncertainty, current methods either disregard parameter uncertainty, or prepare for the worse case. Our research focuses on a criterion that trades off between the two conflicting views.

2.1. The nominal MDP problem

We consider an infinite horizon Markov decision process described as follows: a finite state space S with $|S|$ states, a finite action space A with $|A|$ actions, a transition probability matrix $P \in \mathbb{R}^{|S|\times|A|\times|S|}$ with $P(s,a,s') = \mathbb{P}(s'|s,a)$, an initial distribution on states q, and a reward vector $r \in \mathbb{R}^{|S|}$. For reasons of tractability, we will limit our attention to the set of mixed stationary Markov policies, which is denoted by Υ. When considering an infinite horizon, an optimal discounted reward stationary policy π is a solution to the following optimization problem:

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

where $\alpha \in [0,1)$ is the discount factor.[3]

The nominal problem is known to be easily solvable using value iteration. However, it does not take into account any uncertainty in the choice of the parameters P and r. In practice, this uncertainty is unavoidable and using the most likely (or expected) parameters can actually lead to a significant bias in the performance of the chosen policy (see Mannor et al., 2007).

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 set (hopefully convex) and optimizes 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:

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

There are two types of uncertainty that are of interest. In the first type, termed fixed uncertainty, r and P are drawn once and remain fixed for all time steps. In the second type, termed repeated uncertainty, r and P are repeatedly drawn from their feasible set at each time step. In both cases, the optimal policy π^* for Problem (1) can be found efficiently (see Nilim & El Ghaoui,).

2.3. The percentile MDP problem

Consider a Bayesian setup where 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 $\tilde{r} \propto \mathcal{N}(\mu_{\tilde{r}}, \Sigma_{\tilde{r}})$, 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 **percentile** form:

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

$$\text{sub. to} \quad \mathbb{P}\left(\mathbb{E}(\textstyle\sum_{t=0}^{\infty} \alpha^t \tilde{r}_t(x_t)|x_0 \propto q,\pi) \geq y\right) \geq \eta \ , \quad (2b)$$

where the probability $\mathbb{P}$ 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 $\mathbb{E}(\cdot|x_0 \propto q,\pi)$ 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 percentile problem gives us an η 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 $\eta = 1$, Problem (2) and Problem (1) are equivalent; thus, $1-\eta$ is a measure of risk of the policy doing worse than y^*. In what follows, we will present the details from a Bayesian point of view in order to preserve the clarity of our derivations. However, frequentist extensions follow naturally as in Mannor et al., 2007. Section 3 will initially focus on how to find an optimal policy to Problem 2 with either reward or parameter uncertainty. Later, in Section 4, the percentile criterion will be used to guide exploration.

[3] Although our analysis will 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 $r(s,a,s')$. They can also be extended to the average reward criterion.

3. Decision making under parameter uncertainty

We first present solution methods for the percentile problem with fixed uncertainty.[4] Under the assumption of Gaussian rewards, solving the percentile MDP is not harder than solving the nominal MDP. We will then present a second order approximation for the problem of transition uncertainty with Dirichlet priors. Because of space constraint, we refer the reader to a full version of this paper for proofs and extensions of the presented solution methods to other distributions.

3.1. The case of reward uncertainty

The Gaussian assumption on reward uncertainty, $\tilde{r} \propto \mathcal{N}(\mu_{\tilde{r}}, \Sigma_{\tilde{r}})$, is a standard assumption in many applications as it allows the modeling of correlation between the reward obtained in different states. In what follows, we will show that finding an optimal stationary policy for the problem of maximizing the η-percentile of the total expected discounted reward (*i.e.*, Problem 2) under fixed Gaussian uncertainty in the reward can be explicitly expressed as a second order cone program (*c.f.*, Lobo et al., 1998). The first step is to convert the Constraint (2b) to a form where the expectation operator is expanded.[5]

$$\mathbb{P}_{\tilde{r}}(\sum_{t=0}^{\infty} q^{\mathsf{T}}(\alpha\Pi P)^t \tilde{r}_t \geq y) \geq \eta \,. \tag{3}$$

Using the assumption of fixed uncertainty, the following form is equivalent to Constraint (3):

$$\mathbb{P}_{\tilde{r}}(\upsilon^{\mathsf{T}}\tilde{r} \geq y) \geq \eta \tag{4a}$$

$$q^{\mathsf{T}} \sum_{t=0}^{\infty} (\alpha\Pi P)^t = \upsilon^{\mathsf{T}} \,. \tag{4b}$$

Lemma 3.1 : *(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*

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

is the same as those satisfying

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

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 $\mathcal{N}(0,1)$.

[4] Although this work focuses on fixed uncertainty, similar methods can be derived for the problem of repeated uncertainty.

[5] Here, $\Pi \in \mathbb{R}^{|S|\times|S|\times|A|}$ such that $\Pi(s_1, s_2, a) = \pi(s_1, a)\mathbb{1}\{s_1 = s_2\}$ and the matrix multiplication ΠP is carried along $\mathbb{R}^{|S|\times(|S|\,|A|)} \times \mathbb{R}^{(|S|\,|A|)\times|S|}$.

Using Lemma 3.1, Constraint (4a) can be converted into the equivalent deterministic convex constraint given that $\eta \geq 0.5$:

$$\upsilon^{\mathsf{T}}\mu_{\tilde{r}} - \Phi^{-1}(\eta)\| \begin{bmatrix} \upsilon^{\mathsf{T}}\Sigma_{\tilde{r}}^{\frac{1}{2}} \end{bmatrix} \|_2 \geq y \,.$$

Lemma 3.2 : *Using the change of variables $\rho = \upsilon^{\mathsf{T}}\Pi$,[6] Constraint* (4b) *is equivalent to:*

$$\upsilon^{\mathsf{T}} = q^{\mathsf{T}} + \alpha \textstyle\sum_a \rho_a^{\mathsf{T}} P_a$$
$$\upsilon^{\mathsf{T}} = \textstyle\sum_{a\in A} \rho_a^{\mathsf{T}} \;,\; \rho_a^{\mathsf{T}} \geq 0 \,, \quad \forall\; a \in A \,,$$

where ρ_a is the a-th column of ρ, and from feasible point (υ, ρ), an equivalent pair (υ, Π) feasible according to Constraint (4b) *can be retrieved using:*

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

The following theorem is proven using the constraint replacement technique presented in Lemma 3.1 and Lemma 3.2.

Theorem 3.3 : *For any $\eta \in [0.5, 1)$, the discounted reward percentile Problem 2 with fixed Gaussian reward uncertainty is equivalent to the convex second order cone program*

$$\max_{\rho\in\mathbb{R}^{|S|\times|A|}} \quad \textstyle\sum_a \rho_a^{\mathsf{T}}\mu_{\tilde{r}} - \Phi^{-1}(\eta)\| \sum_a \rho_a^{\mathsf{T}}\Sigma_{\tilde{r}}^{\frac{1}{2}} \|_2 \tag{6a}$$

$$\text{sub. to} \quad \textstyle\sum_a \rho_a^{\mathsf{T}} = q^{\mathsf{T}} + \sum_a \alpha\rho_a^{\mathsf{T}} P_a \tag{6b}$$

$$\rho_a^{\mathsf{T}} \geq 0 \,, \quad \forall\; a \in A \,, \tag{6c}$$

where given an optimal assignment ρ^*, an optimal policy π^* can be retrieved using:

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

Solving a second order cone program (SOCP) is often considered to be not much more computationally demanding than solving a linear program of comparable size (*i.e.*, it is feasible to solve problems of 10^3-10^4 variables).[7] This is an appealing feature for the percentile problem which is actually preserved under different reductions of the Gaussian assumption. However, one can show that the NP-complete 3SAT problem can be reduced to solving Problem 2 with a discrete distribution on the rewards. Hence,

Theorem 3.4 : *Solving the percentile MDP Problem 2 with **general uncertainty** in the reward parameters is NP-hard*

[6] By $\rho = \upsilon^{\mathsf{T}}\Pi$, we refer to $\rho \in \mathbb{R}^{|S|\times|A|}$ such that $\rho(s', a) = (\upsilon^{\mathsf{T}}\Pi)(s', a) = \sum_s \upsilon(s)\Pi(s, s', a)$.

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

3.2. The case of uncertainty in transition parameters

We now focus on the problem of transition uncertainty. This type of uncertainty is naturally 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 contains inherent uncertainty. Unfortunately, as was the case for reward uncertainty, one can show that the percentile problem is computationally hard in general.

Corollary 3.5 : *Solving percentile MDP Problem 2 for general uncertainty in the transition parameters is NP-hard.*

Because we cannot expect to solve this problem with general transition uncertainty, our analysis makes the Dirichlet assumption and proposes a solution method that generates near optimal solutions given a sufficient number of samples drawn from $\tilde{P}$.

Unlike in the case of reward uncertainty, where 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 non-trivial. More specifically, as presented in (Mannor et al. 2007), the expected return can be expressed as

$$\mathbb{E}(\sum_{t=0}^{\infty}\alpha^t r(x_t)|x_0 \propto q,\pi) = \mathbb{E}(q^{\mathsf{T}}\sum_{k=0}^{\infty}\alpha^k (X^{\pi}\Pi\Delta\tilde{P})^k X^{\pi} r) \,,$$

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 $\tilde{P}$ can only generate stochastic matrices, thus ensuring that $I - \alpha\Pi\mathbb{E}(\tilde{P})$ is nonsingular.[8] The expression $\mathbb{E}(\sum_{t=0}^{\infty}\alpha^t r(x_t)|x_0 \propto q,\pi)$ therefore depends on all the moments of the uncertainty in $\tilde{P}$. Because we expect the higher order moments of $\tilde{P}$ to decay quickly with the number of samples drawn from $\tilde{P}$, it is reasonable to focus on second order approximation

$$\mathbb{E}(\sum_{t=0}^{\infty}\alpha^t r(x_t)|x_0 \propto q,\pi) \approx q^{\mathsf{T}}X^{\pi}r + \alpha^2 q^{\mathsf{T}}X^{\pi}\Pi Q X^{\pi} r \,,$$

where $Q \in \mathbb{R}^{|S|\times|A|\times|S|}$, such that

$$Q_{(i,a,j)} = \left(\mathbb{E}(\Delta\tilde{P}X^{\pi}\Pi\Delta\tilde{P})\right)_{(i,a,j)} = \pi_{(i,a)}\Sigma^{(i,a)}_{(j,\cdot)}X^{\pi}_{(\cdot,i)} \,.$$

This is under the assumption that the rows of $\tilde{P}$ are independent and using $\Sigma^{(i,a)}$ to represent the covariance between the terms of the transition vector from state i with action a.

[8] Refer to footnote 5

Let $\mathbb{F}(\pi)$ be the second order approximation of the expected return under transition uncertainty, such that

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

In order to show that minimizing $\mathbb{F}(\pi)$ leads to a near-optimal percentile policy, we make the assumption that $\tilde{P}$ behaves according to a Dirichlet distribution. This allows us to bound the approximation error in terms of number of observed transitions. One can then show, using Markov's inequality, that the following theorem holds.

Theorem 3.6 : *Given state transition samples* $\{(s_1,a_1,s'_1),...,(s_M,a_M,s'_M)\}$ *and suppose that* $M^{a^*}_{i^*} = \min_{i,a} M^{(i,a)}$*, and* $\eta \in [0.5,1)$*, policy*

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

is $o(1/\sqrt{(1-\eta)M^{a^*}_{i^*}})$ *optimal according to the percentile MDP Problem 2 with known rewards, where the probability* $\mathbb{P}$ *is the probability of drawing* $\tilde{P}$ *from the posterior Dirichlet distribution given that* $M^{(i,a)}$ *transitions were observed from each state* i *and action* a*.*

3.3. A machine replacement problem with Dirichlet uncertainty in the transition parameters

We have chosen the machine replacement problem as an application for our methods. Let us assume that we are interested in the repair cost that is incurred by a factory that holds a large number of machines, given that each of these machines are modeled with the same underlying MDP for which the transition 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 η, this policy will have a low maintenance cost on average. This is exactly what the percentile criterion quantifies.

Our experiment uses 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 a 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 with a cost of 2, and $R2$ a more involved one with a cost of 10. A cost of 20 penalizes reaching an age of 8. In each of these states, one has access to three repair services for the machine. We assume a Dirichlet model for all transitions. In the case of each of the three repair options, we use slightly perturbed versions of a reference Dirichlet model that is presented in Figure 1. In this figure, the expected transition parameters are presented given that M transitions are observed from each state and action.

We apply three solution methods to this decision problem. First, the nominal problem is formulated using the expected

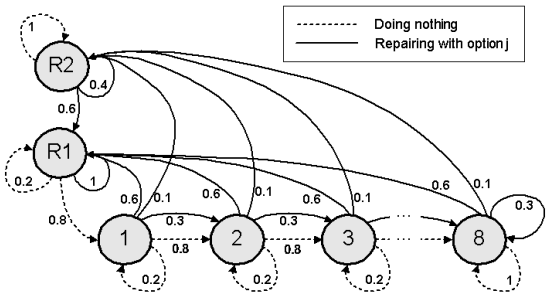

Figure 1. Instance of a machine replacement problem with Dirichlet uncertainty in the transition parameters.

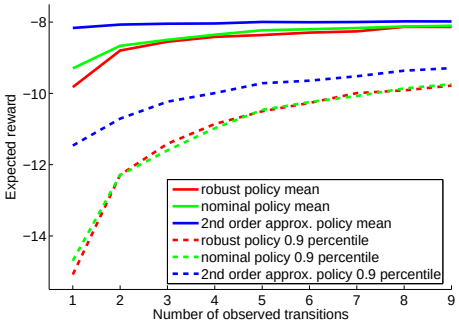

Figure 2. Performance comparisons between the optimal policies according to the nominal, robust and percentile criteria on 10000 runs of the machine replacement problem as the uncertainty is reduced.

transition probabilities. Then, we apply the robust method presented in Section 2.2, for which we choose to use as uncertainty set a box in $\mathbb{R}^{|S|\times|A|\times|S|}$ that contains $\tilde{P}$ with 90% confidence.[9] Finally, we use the "2nd order approximation" performance measure presented in Section 3.2 to find an optimal policy for this machine replacement problem.[10]

Figure 2 shows the mean and 90th percentile performances of the different methods on this problem as uncertainty in the parameters decreases (or M increases). It is interesting to see that the policy obtained by the 2nd order approximation method outperforms the policy obtained using the robust method and nominal method for a range of uncertainty levels (low to high). This is mainly due to the fact the 2nd order approximation method returns a policy that uses, in states 8 and $R1$, a mixed strategy over the repair options in order to reduce the transition variance and, indirectly, the overall expected cost.

[9]Implementation details: using 10000 samples drawn from $\tilde{P}$ and a given γ ratio, for each parameter $P_{(i,a,j)}$ we choose $A_{(i,a,j)}$ and $B_{(i,a,j)}$ so that they include a ratio of γ of the random samples. A search over γ is done to find the minimal γ that leads 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 $\tilde{P}$.

[10]Implementation details: Matlab's optimization toolbox was used to solve this constrained non-convex optimization of $\mathbb{F}(\pi)$.

4. Efficient exploration using percentile optimization

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 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. Methods such as R-max and model based interval estimation (see Strehl & Littman, 2005), lead with high-probability to near-optimal policies in polynomial time. We are interested in a slightly different framework. We assume that, before committing to an exploitation strategy (such as a repair policy for the problem described in Section 3.3), 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

$$\mathcal{V}(i,a) = \mathbb{E}\left(\max_{\pi'} \mathcal{G}(\pi', \tilde{r}', \tilde{P}')\right) - \max_{\pi} \mathcal{G}(\pi, \tilde{r}, \tilde{P}) \ , \quad (7)$$

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 then the observation cost. In what follows, we apply this simple learning strategy on the percentile problem with Gaussian priors on rewards, or Dirichlet priors on transitions.

4.1. Efficient learning of Gaussian rewards

We start by studying the case where we know the transition parameters of the MDP exactly but where we have uncer-

tainty about the rewards. We assume that we have the option of buying noisy measurements of the rewards $\hat{r}(i,a) = r(i,a) + \nu(i,a)$, where $\nu(i,a) \propto \mathcal{N}(0,\sigma_\nu)$. Using 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 $\tilde{r}(i,a) \propto \mathcal{N}(\mu_{(i,a)}, \sigma^2_{(i,a)})$, we can evaluate the posterior distribution $\tilde{r}'(i,a) \propto \mathcal{N}(\mu'_{(i,a)}, \sigma'^2_{(i,a)})$.[11] The value of information $\mathcal{V}(i,a)$, with $\mathcal{G}(\pi,\tilde{r})$ set as the optimal value of percentile Problem 6, 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^*:

$$\begin{aligned}\mathcal{V}(i,a) &= \mathbb{E}(\mathcal{G}(\pi^*, \tilde{r}'_{\hat{r}(i,a)})) - \mathcal{G}(\pi^*,\tilde{r}) \\ &= \mathbb{E}\left(\sum_a {\rho_a^*}^\mathsf{T} \mu_{\tilde{r}'}\right) - \Phi^{-1}(\eta)\| \sum_a {\rho_a^*}^\mathsf{T} \Sigma^{\frac{1}{2}}_{\tilde{r}'} \|_2 - \mathcal{G}(\pi^*,\tilde{r}) \\ &= \Phi^{-1}(\eta)\left(\| \sum_a {\rho_a^*}^\mathsf{T} \Sigma^{\frac{1}{2}}_{\tilde{r}} \|_2 - \| \sum_a {\rho_a^*}^\mathsf{T} \Sigma^{\frac{1}{2}}_{\tilde{r}'} \|_2\right) ,\end{aligned}$$

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.

The following experiments compare percentile based sampling to random sampling and the model based interval estimation (MBIE) strategy[12] 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 $\tilde{r}(i,a) \propto \mathcal{N}(\mu_{\hat{r}}(i,a),1)$ 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 3 presents the average 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 4 shows the average total percentile cost (final percentile return added to cost of extracted samples) of our learning strategy given different observation prices. 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.

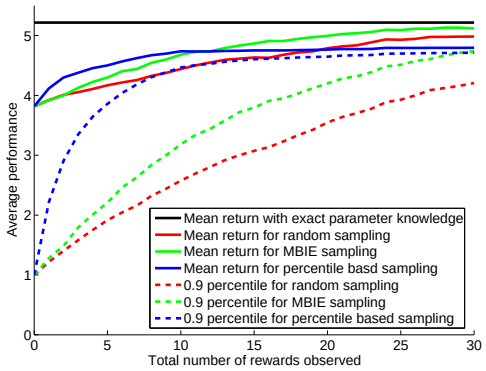

Figure 3. Average percentile and mean performances of sampling strategies on a set of 1000 random MDPs with reward uncertainty (free observations).

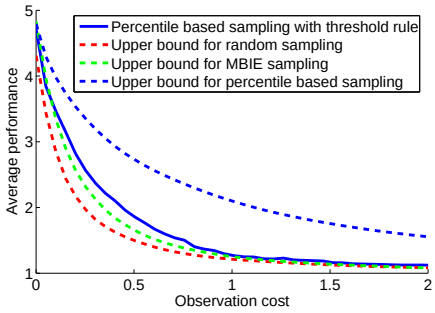

Figure 4. Average total percentile return on the MDPs of Figure 3 for a range of observation costs.

4.2. Efficient learning of Dirichlet transitions

In the case where we have transition uncertainty, we model our uncertainty using a Dirichlet prior and now have the option of buying state transition observations. The same problem arises in this framework: when should one stop paying for these observations and start exploiting the system as one

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

[12] Being 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.

knows it? Given M transition observations from state i using action a, one can update the Dirichlet prior as suggested in the Bayesian framework (see Gelman et al., 2003). Using $\mathcal{G}(\pi, \tilde{P}) = \mathbb{F}(\pi)$, the 2nd order approximation to the expected return presented in Section 3.2, the value of information $\mathcal{V}(i, a)$ can therefore be estimated with Monte Carlo methods. Here, computing the lower bound for $\mathcal{V}(i, a)$ as in Section 4.1 largely reduces the computational complexity of the Monte Carlo method by sparing us from performing the optimization inside the expectation term of Equation (7).

Our experiments compare the percentile learning rule to random sampling and model based interval estimation strategy on a set of 1000 randomly generated MDPs with transition uncertainty. Each model has 10 states, 4 actions, a discount factor of 0.8. For a given model, the rewards are generated from $\mathcal{N}(0, 1)$ for each state, and the initial uncertainty in a transition from (i, a) is generated by selecting 3 possible next states uniformly, drawing the Dirichlet parameters uniformly in the [0,1] interval and normalizing them to sum to 1. As in Section 4.1, Figure 5 presents the average percentile and mean performances over the set of uncertain MDPs given a number of observations chosen by the different strategies (no observation cost). Again, the percentile rule outperforms on average random sampling and MBIE in the choice of observations to make. Figure 6 shows the average total percentile cost of our learning strategy given different observation prices. Unlike the case of reward uncertainty, the stopping criterion does not outperform the lower bounds on other methods but we expect it to perform well against any reasonable stopping criterion based on random or MBIE sampling.

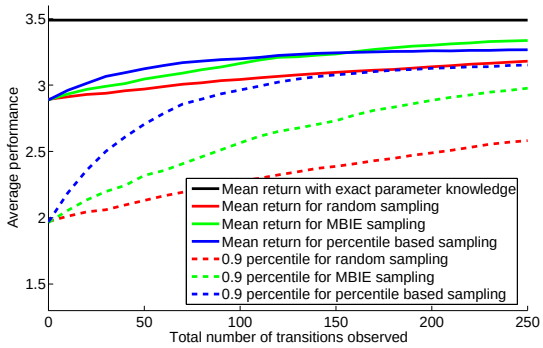

Figure 5. Average percentile and mean performances on a set of 1000 random MDPs with transition uncertainty (free observations).

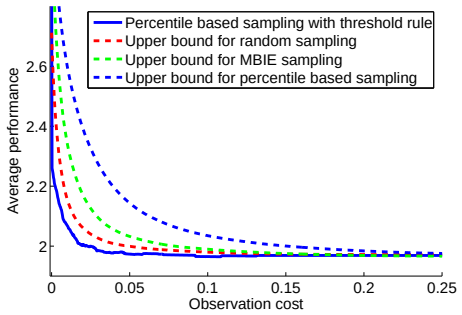

Figure 6. Average total percentile return on the MDPs of Figure 5 for a range of observation costs.

5. Discussion

In the context of high cost observations, the results of Section 4 demonstrate that random sampling and MBIE are less efficient exploration methods comparing to the proposed value of information exploration. This is mainly due to the fact that these methods disregard the cost of observations, but focus entirely on reaching an ϵ-optimal policy in the long term. Methods such as the E^3 algorithm (Kearns and Singh 1998) and the R-max algorithm (Brafman and Tennenholtz 2003) suffer similarly. When observations incur a non-negligible cost, the exploration-exploitation dilemma takes the shape of a problem better expressed through value of information. Ideally, one needs to reason about sequences of observations that will have a high expected impact on percentile return while preserving a low observation cost. Unfortunately, because this problem is intractable, we settle for a strategy that acts greedily with respect to a single decision. Such a strategy is therefore subject to aborting exploration when no single observation can lead to an immediate useful reduction of uncertainty although a series of them might. This phenomenon can be observed in Figure 3 where the percentile strategy does not lead in general to the optimal policy of the underlying MDP as more samples are used. However, in applications where one can only afford a small number of observations (compared to the size of the state space), Figure 4 shows that the percentile strategy is the best option.

The application of value of information to the exploration-exploitation dilemma is not new (see Dearden et al., 1999). However, previous work only applied this concept to the MDP in its nominal form without considering the value of risk-sensitive policies. Also, these methods have been considered to be impractical since they are confronted to the problem of evaluating $\mathbb{E}\left(\max_{\pi'} \mathcal{G}(\pi', \tilde{r}', \tilde{P}')\right)$ with $\mathcal{G}(\pi', \tilde{r}', \tilde{P}')$ being the optimal value of the nominal problem for each pair (i, a). This can only be done using Monte Carlo methods and the computation requirements grow quickly with the dimension of the state space, as one needs to solve an MDP for each Monte Carlo sample of each (i, a) pair. By studying the percentile problem, we obtain a form for $\mathcal{V}(i, a)$ which can be approximated efficiently using the lower bounds presented in Section 4.1

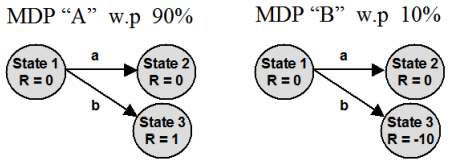

Figure 7. Uncertain MDP where 90th percentile based sampling is risk tolerant and chooses action b.

and 4.2.[13]

It is important to note that the success of our exploration strategy relies on the formulation of an adequate prior over the parameters and of a percentile threshold that truly reflects risk tolerance. Consider the uncertain MDP presented in Figure 7. If prior knowledge indicates that the system MDP "A" is with 90% probability, a 90th percentile based sampling chooses to exploit using action b without sampling any state. This might seem sub-optimal since by sampling the reward in state 3 it is possible to completely determine the system and then choose the policy that avoids the negative reward. Percentile based sampling disregards the risk related to this negative event since, based on the prior distribution, the risk is tolerated by the target percentile. In this example, one might feel more comfortable using a 99th percentile.

Finally, we believe that percentile based exploration strategy should naturally extend to model based online learning. We also expect that many important problems that have been addressed using standard MDP models and "naive" exploration methods, should be revisited and better resolved using the proposed risk-sensitive percentile criterion.

Acknowledgments

We would like to thank Constantine Caramanis, Xu Huan and David Varodayan for helpful discussions. We also thank an anonymous reviewer for suggesting the example presented in the discussion. E. Delage was supported by the Fonds Québécois de la recherche sur la nature et les technologies. S. Mannor was partially supported by the Natural Sciences and Engineering Research Council of Canada and by the Canada Research Chairs Program.

References

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

Brafman, R., & Tennenholtz, M. (2003). R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. of Machine Learning Research.*, *3*, 213–231.

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

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

Dearden, R., Friedman, N., & Andre, D. (1999). Model-based Bayesian exploration. *Proc. of Uncertainty in AI* (pp. 150–159).

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

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian data analysis, second edition*. Chapman & Hall/CRC.

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

Howard, R. (1966). Information value theory. *IEEE Trans. on Systems Science and Cybernetics*, *SSC-2*, 22–26.

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

Kearns, M., & Singh, S. (1998). Near-optimal reinforcement learning in polynomial time. *Proc. ICML* (pp. 260–268).

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

Mannor, S., Simester, D., Sun, P., & Tsitsiklis, J. (2007). Bias and variance in value function estimation. *Management Science*, *53*, 308–322.

Nemirovski, A., & Shapiro, A. (2006). Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, *17*, 969–996.

Nilim, A., & El Ghaoui, L. Robust Markov decision processes with uncertain transition matrices. *Operations Research*, *53*, 780–798.

Prékopa, A. (1995). *Stochastic programming*. Kluwer Academic Publishers.

Putterman, M. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. Wiley.

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

Strehl, A., & Littman, M. (2005). A theoretical analysis of model-based interval estimation. *Proc. ICML* (pp. 857–864).

[13] The lower bound approximation cannot be effectively applied to value of information with the nominal problem. In the case of reward uncertainty, $V(i,a) \approx \mathbb{E}(\mathcal{G}(\pi^*, \tilde{r}', \tilde{P}')) - \mathcal{G}(\pi^*, \tilde{r}', \tilde{P}') = \mathbb{E}(\sum_a {\rho_a^*}^\mathsf{T} \mu_{\tilde{r}'}) - \sum_a {\rho_a^*}^\mathsf{T} \mu_{\tilde{r}} = 0$, for all (i,a).