

Approximate Probabilistic Constraints and Risk-Sensitive Optimization Criteria in Markov Decision Processes

Dmitri A. Dolgov and Edmund H. Durfee

Department of Electrical Engineering and Computer Science

University of Michigan

Ann Arbor, MI 48109

{ddolgov,durfee}@umich.edu

Abstract

The majority of the work in the area of Markov decision processes has focused on expected values of rewards in the objective function and expected costs in the constraints. Although several methods have been proposed to model risk-sensitive utility functions and constraints, they are only applicable to certain classes of utility functions and allow limited expressiveness in the constraints. We propose a construction that extends the standard linear programming formulation of MDPs by augmenting it with additional optimization variables, which allows us to compute the higher order moments of the total costs (and/or reward). This greatly increases the expressive power of the model, and supports reasoning about the probability distributions of the total costs (reward). Consequently, this allows us to formulate more interesting constraints and to model a wide range of utility functions. In particular, in this work we show how to formulate the constraint that bounds the probability of the total incurred costs falling within a given range. Constraints of that type arise, for example, when one needs to bound the probability of overutilizing a consumable resource. Our construction, which greatly increases the expressive power of our model, unfortunately comes at the cost of significantly increasing the size and the complexity of the optimization program. On the other hand, it allows one to choose how many higher order moments of the costs (and/or reward) are modeled as a means of balancing accuracy against computational effort.

1 Introduction

Markov processes are widely used to model stochastic environments, due to their expressiveness and analytical tractability. In particular, unconstrained (e.g. [3, 4, 8, 15]) as well as constrained (e.g. [1, 2]) Markov decision processes have gained significant popularity in the AI and OR communities as tools for devising optimal policies under uncertainty. The vast majority of the work in the area has focused on optimization criteria and constraints that are based on the expected values of the rewards and costs. However, such risk-neutral approaches are not always applicable and expressive enough,¹ thus precipitating the need for extending the MDP framework to model risk-sensitive utility functions and constraints.

Several approaches [9, 12, 13] to modeling risk-sensitive utility functions have been proposed that work by transforming risk-sensitive problems into equivalent risk-neutral problems, which can then be solved by dynamic programming. However, this transformation only works for a certain class of utility functions. Namely, this has been done for exponential utility functions that are characteristic of agents that have “constant local risk aversion” [14] or obey the “delta property” [9], which says that a decision maker’s risk sensitivity is independent of his current wealth. This approximation has a number of very nice analytical properties, but is generally considered somewhat unrealistic [9]. Our work attempts to address this issue via approximate modeling of a more general class of utility functions.

As with utility functions, there has been a significant amount of work on risk-sensitive constraints in MDPs. These methods typically work by constraining or including in the objective function the variance of the costs [6, 7, 10, 18, 19] or reasoning about sample-path costs in the case of per unit-time problem formulations [2, 16, 17]. However, in some cases, reasoning about the variance only is also not expressive enough (see [5] for a more detailed discussion).

We propose a method that allows explicit reasoning about the probability distributions of the total reward in the objective function and the distribution of costs in constraints, thus allowing us to represent a wide class of interesting optimization criteria and constraints. In this work, we describe a method for handling probabilistic constraints, but the approach is also directly applicable to risk-sensitive objective functions. We focus on transient (or episodic) Markov processes [11] and

¹As pointed out, for example, by Ross and Chen in the telecommunication domain [16].

base our approach on the standard occupancy-measure linear programming formulation of constrained Markov decision processes (CMDPs). We augment the classical program with additional optimization variables, which allows us to compute the higher order moments of the total incurred costs for stationary Markov policies. This enables us to reason about the probability distributions of the total costs, and consequently, to express more interesting constraints such as bounding the probability that the total costs fall within a given range (or exceed a threshold).

It is important to note that, in general, arbitrary utility functions and arbitrary constraints do not obey the Markov property, which means that stationary Markov (history-independent) policies are not guaranteed to be optimal under such utility functions and constraints. However, given the practical difficulty of implementing non stationary history-dependent policies, in this work we limit the search to the class of stationary Markov policies, i.e. we are interested in finding the best policy that satisfies the given constraints and maximizes the given utility function, among the class of stationary history-independent policies.

2 Preliminaries

We formulate our optimization problem as a stationary, discrete-time, fully-observable constrained Markov decision process. In this section, we review some well-known facts from the theory of standard [3, 15] and constrained [1] fully-observable Markov decision processes.

A standard constrained Markov decision process (CMDP) can be defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \boldsymbol{\alpha}, \mathbf{r}, \mathbf{c} \rangle$, where \mathcal{S} is a finite set of states, \mathcal{A} is a finite set of actions, $\mathbf{P} = [P_{ij}^a] : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ defines the transition function (P_{ij}^a is the probability that the agent will go to state j if it executes action a in state i), $\boldsymbol{\alpha} = [\alpha_i] : \mathcal{S} \rightarrow [0, 1]$ is the initial probability distribution over the state space, $\mathbf{r} = [r_{ia}] : \mathcal{S} \times \mathcal{A} \rightarrow \mathfrak{R}$ defines the reward function (agent receives a reward of r_{ia} for executing action a in state i), and $\mathbf{c} = [c_i] : \mathcal{S} \rightarrow \mathfrak{R}$ are the costs.²

A solution to a CMDP is a policy that prescribes a procedure for selecting an action that typically maximizes some measure of performance (based on the rewards \mathbf{r}), while satisfying constraints (based on the costs \mathbf{c}). A stationary Markov policy π can be described as a mapping of states to probability distributions over actions: $\pi = [\pi_{ia}] : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$. We address the problem of finding optimal stationary Markov policies, under probabilistic constraints (defined in section 3).

For a Markov system, the initial probability distribution $\boldsymbol{\alpha}$, the transition probabilities, and the policy together completely determine the evolution of the system in a stochastic sense:

$$\boldsymbol{\rho}(t+1) = \tilde{\mathbf{P}}\boldsymbol{\rho}(t), \quad \boldsymbol{\rho}(0) = \boldsymbol{\alpha}, \quad (1)$$

where $\boldsymbol{\rho}(t) = [\rho_i(t)]$ is the probability distribution of the system at time t , and $\tilde{\mathbf{P}} = \tilde{\mathbf{P}}(\pi) = [\tilde{P}_{ij}^a]$ is the probability transition matrix induced by the policy ($\tilde{P}_{ij}^a = \sum_a \pi_{ia} P_{ij}^a$).

In this work we focus our attention on discrete-time *transient* (or episodic) problems, where there is no predefined number of time steps that the agent spends in the system, but it cannot stay there forever. Given a finite state space, this assumption implies that there have to exist some state-action pairs $\{i, a\}$ for which $\sum_j P_{ij}^a < 1$.

For most of the analysis in this paper we use the expected total reward as the policy evaluation criterion: $V(\pi, \boldsymbol{\alpha}) = \sum_{t=0}^{\infty} \sum_i \rho_i(t) \sum_a \pi_{ia} r_{ia}$, which, for a transient system with bounded rewards, converges to a finite value.

A standard CMDP where constraints are imposed on the expected total costs, and the expected total reward is being maximized can be formulated as the following linear program [1, 15]:

$$\max \sum_i \sum_a x_{ia} r_{ia} \quad \left| \quad \begin{array}{l} \sum_a x_{ja} - \sum_i \sum_a x_{ia} P_{ij}^a = \alpha_j \\ \sum_i c_i \sum_a x_{ia} \leq C_0, \quad x_{ia} \geq 0, \end{array} \right. \quad (2)$$

where C_0 is the upper bound on the expected total incurred cost, and the optimization variables x_{ia} are called the *occupancy measure* of a policy and be interpreted as the expected total number of times action a is executed in state i .

3 Problem Description

Given a standard constrained MDP model $\langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \boldsymbol{\alpha}, \mathbf{r}, \mathbf{c} \rangle$, we would like to find a policy that maximizes the total expected reward, while satisfying the constraints on the probability of that the cost exceed a given upper bound, i.e. $P[C \geq C_0] \leq p_0$,

²The costs are said to be incurred for visiting states rather than executing actions, but all results can be trivially extended to the latter case. Also note that most interesting problems involve several costs, and while we make the simplification that there is only one cost incurred for visiting a state, this is done purely for notational brevity. All results presented in this work are directly applicable (and most useful) for problems with several costs.

where C is the total cost. In other words, we need to solve the following optimization problem:

$$\max \sum_i \sum_a r_{ia} x_{ia} \quad \left| \begin{array}{l} \sum_a x_{ja} - \sum_i \sum_a x_{ia} P_{ij}^a = \alpha_j \\ P[C \geq C_0] \leq p_0, \end{array} \right. \quad (3)$$

where the optimization variables x_{ia} have the standard interpretation of the expected total number of times action a is executed in state i . If $P[C \geq C_0]$ could be expressed as a simple function of $\mathbf{x} = [x_{ia}]$, the problem would be solved, as one could simply plug the expression into the above program. Unfortunately, things are not as wonderful, and the above dependency is significantly more complex.

A simple linear approximation to the above program (when costs are non-negative) can be obtained by using the Markov inequality:

$$P(C \geq C_0) \leq \frac{E[C]}{C_0} = \frac{1}{C_0} \sum_i c_i \sum_a x_{ia}, \quad (4)$$

which allows one to express $P(C \geq C_0)$ as a linear function of the occupancy measure \mathbf{x} . Our investigation [5] of this approximation showed, unsurprisingly, that this linear approximation is computationally cheap but usually leads to suboptimal policies, because the Markov inequality provides a very rough upper bound on the probability that the total cost exceed a given limit C_0 . The purpose of this work is to improve this approximation.

To this end, we are going to come up with a system of coordinates \mathbf{y} , such that the constraint $P[C \geq C_0] \leq p_0$ can be expressed as a simple function of \mathbf{y} , so that the expression can be plugged into the optimization program (eq. 3). However, one has to note that there are only $|\mathcal{S}||\mathcal{A}|$ free parameters in the system, so not all optimization variables \mathbf{y} are going to be independent and additional constraints might have to be imposed.

As mentioned earlier, the method presented in this paper works for more general problems than (eq. 3), which we use as an interesting example to illustrate the approach. Section 7 comments on other problems that this method applies to.

4 Calculating the Probability of Exceeding Cost Bounds

To find the probability of exceeding the cost bounds, it would be very helpful to know the probability density function (pdf) $f_C(C)$.³ Then, the probability of exceeding the cost bounds could be expressed simply as $P[C \geq C_0] = \int_{C_0}^{\infty} f_C(C) dC$.

Unfortunately, $f_C(C)$ is not easily available. However, it is a well-known fact that under some conditions the moments of a random variable completely specify its distribution.⁴ The k^{th} moment of a random variable x is defined as the expected value of x^k : $E_x^k = \int_{-\infty}^{\infty} f_x(x) x^k dx$. One way to compute the pdf $f_x(x)$, given the moments E_x^k is via an inverse Legendre transform.⁵ Indeed, the Legendre polynomials $\mathcal{P}_l(x) = \frac{1}{2^l l!} \frac{d^l}{dx^l} (x^2 - 1)^l$ form a complete orthogonal set on the interval $[-1, 1]$: $\int_{-1}^1 \mathcal{P}_l \mathcal{P}_m = \frac{2}{2l+1} \delta_{lm}$. Therefore, a function on that interval $[-1, 1]$ can be approximated as a weighted sum of Legendre polynomials: $f(x) = \sum_{l=0}^{\infty} b_l \mathcal{P}_l(x)$, where $\mathcal{P}_l(x)$ is the l^{th} Legendre polynomial, and b_l is a constant coefficient, obtained by multiplying the polynomials by $f(x)$, integrating over $[-1, 1]$, and using the orthogonality condition:

$$b_l = \frac{2l+1}{2} \int_{-1}^1 f(x) \mathcal{P}_l(x) dx = \sum_k a_{kl} E_x^k \quad (5)$$

Realizing that $\int f(x) \mathcal{P}_l(x) dx$ is just a linear combination of several moments E_x^k , we get:

$$f(x) = \sum_{l=0}^{\infty} \sum_k a_{kl} E_x^k \mathcal{P}_l(x), \quad (6)$$

where in the second summation, the index k runs over all powers of x present in \mathcal{P}_l . Therefore, for an $x \in [-1, 1]$, we can express the probability that x is greater than some x_0 as a linear function of the moments:

$$P[x \geq x_0] = \int_{x_0}^1 f_x(x) dx = \int_{x_0}^1 \sum_{l=0}^{\infty} \sum_k a_{kl} E_x^k \mathcal{P}_l(x) dx = \sum_k \psi_k(x_0) E_x^k, \quad (7)$$

³Note that, in general, for an MDP with finite state and action spaces, the total costs have a discrete distribution. However, we make no assumptions about the continuity of the pdf $f_C(C)$, and our analysis carries through for both continuous and discrete density functions; in the latter case, $f_C(C)$ can be represented as a sum of Dirac delta functions: $f_C(C) = \sum_k p_k \delta(x - p_k)$.

⁴This is true when the power series of the moments that specifies the characteristic function converges, which holds in our case due to the transient nature of the Markov process and the fact that costs are finite.

⁵A more common and natural way involves inverting the characteristic function of x via a Fourier transform, but the method does not work for this problem.

where $\psi_k(x_0) = \sum_l a_{kl} \int_{x_0}^1 P_l(x) dx$, in which the index l runs over all polynomials that include the k^{th} power of x .

Therefore, if we normalize C to be in the interval $[-1, 1]$ (section 6 discusses in more detail the necessary transformation for a transient system), we could use the above method to express $P[C \geq C_0]$ as a linear function of the moments E_C^k . Now, if we could come up with a system of coordinates \mathbf{y} , such that the moments E_C^k could be expressed via \mathbf{y} , we might be able to formulate a manageable approximation to (eq. 3). However, it is important to note that unless we use an infinite number of moments, the resulting program will be an approximation to the original one.

5 Computing the Moments

As mentioned in the previous section, the properties of the pdf of the total cost are not immediately obvious, as the total cost is a sum of a random number of dependent random variables. We do, however, know how the system evolves with time, i.e. given the initial probability distribution, a policy, and the corresponding transition probabilities over states, we know the probability that the system is in state i at time t – it is simply $(\tilde{\mathbf{P}}^t \boldsymbol{\alpha})_i$, where $\tilde{\mathbf{P}} = \tilde{\mathbf{P}}(\pi) = [\tilde{P}_{ij}]$ is the probability transition matrix induced by the policy ($\tilde{P}_{ij} = \sum_a \pi_{ia} P_{ij}^a$). In other words, we know the probability distribution for the random variables $n_i(t) = \{0, 1\}$, where $n_i(t) = 1$ if state i is visited at time t , and 0 otherwise.

Let us also define for every state a random variable $N_i = \sum_{t=0}^{\infty} n_i(t)$ that specifies the total number of times state i is visited. Then, the moments E_C^k of C can be expressed as linear functions of the cross-moments $E_{i_1 \dots i_k} = \langle N_{i_1} N_{i_2} \dots N_{i_k} \rangle$ (the expected value of the product) as follows:

$$\begin{aligned} E_C^1 &= \left\langle \sum_i c_i N_i \right\rangle = \sum_i c_i \langle N_i \rangle = \sum_i c_i E_i \\ E_C^2 &= \left\langle \left(\sum_i c_i N_i \right)^2 \right\rangle = \sum_i \sum_j c_i c_j \langle N_i N_j \rangle = \sum_i \sum_j c_i c_j E_{ij} \\ E_C^k &= \sum_{i_1} \sum_{i_2} \dots \sum_{i_k} c_{i_1} c_{i_2} \dots c_{i_k} E_{i_1 i_2 \dots i_k} \end{aligned} \quad (8)$$

Let us now compute the first moments $E_i = \langle \sum_{t=0}^{\infty} n_i(t) \rangle = \sum_{t=0}^{\infty} \langle n_i(t) \rangle$. Recalling that $n_i(t) = \{0, 1\}$, and, therefore, its mean equals the probability that $n_i(t)$ is 1:⁶

$$E_i = \sum_{t=0}^{\infty} P[n_i(t)] = \sum_{t=0}^{\infty} (\tilde{\mathbf{P}}^t \boldsymbol{\alpha})_i = ((\mathbf{I} - \tilde{\mathbf{P}})^{-1} \boldsymbol{\alpha})_i, \quad (9)$$

where \mathbf{I} is the identity matrix, and $\sum_{t=0}^{\infty} \tilde{\mathbf{P}}^t = (\mathbf{I} - \tilde{\mathbf{P}})^{-1}$ holds, because $\lim_{t \rightarrow \infty} \tilde{\mathbf{P}}^t = 0$ for our transient system. Multiplying by $(\mathbf{I} - \tilde{\mathbf{P}})$, we get:

$$E_i - \sum_j \tilde{P}_{ji} E_j = \alpha_i \quad (10)$$

Note that the above is exactly the ‘‘conservation of probability’’ constraint in (eq. 3). Indeed, since $\tilde{P}_{ij} = \sum_a P_{ij}^a \pi_{ia}$ and $x_{ia} = E_i \pi_{ia}$, the two are identical. Let us now compute the second moments in a similar fashion:

$$\begin{aligned} E_{ij} = \langle N_i N_j \rangle &= \left\langle \left(\sum_{t_1=0}^{\infty} n_i(t_1) \right) \left(\sum_{t_2=0}^{\infty} n_j(t_2) \right) \right\rangle = \sum_{t_1=0}^{\infty} \sum_{t_2=0}^{\infty} \langle n_i(t_1) n_j(t_2) \rangle \\ &= \sum_{t_1=0}^{\infty} \sum_{t_2=t_1}^{\infty} \langle n_i(t_1) n_j(t_2) \rangle + \sum_{t_2=0}^{\infty} \sum_{t_1=t_2}^{\infty} \langle n_i(t_1) n_j(t_2) \rangle - \sum_{t=0}^{\infty} \langle n_i(t) n_j(t) \rangle \end{aligned} \quad (11)$$

Once again recalling that $n_i(t)$ are binary variables, and since the system can only be in one state at a particular time, the mean of their product is:

$$\langle n_i(t_1) n_j(t_2) \rangle = \begin{cases} P[n_i(t_1), n_j(t_2)], & \text{if } t_1 \neq t_2 \\ \delta_{ij} P[n_i(t_1)], & \text{if } t_1 = t_2, \end{cases} \quad (12)$$

⁶Hereafter we use the notation $P[x]$ for binary variables as a shorthand for $P[x = 1]$, and $P[x, y]$ for $P[(x = 1) \wedge (y = 1)]$

where $P[n_i(t_1), n_j(t_2)]$ is the probability that state i is visited at time t_1 and state j is visited at time t_2 . Also, since the system is Markovian, for $t_1 \leq t_2$, we have:

$$P[n_i(t_1), n_j(t_2)] = P[n_i(t_2)|n_j(t_1)]P[n_j(t_1)] = (\tilde{P}^{t_2-t_1})_{ij}P[n_i(t_1)] \quad (13)$$

Substituting, we obtain:

$$\begin{aligned} E_{ij} &= \sum_{t_1=0}^{\infty} \sum_{t_2=t_1}^{\infty} (\tilde{P}^{t_2-t_1})_{ij}P[n_i(t_1)] + \sum_{t_2=0}^{\infty} \sum_{t_1=t_2}^{\infty} (\tilde{P}^{t_1-t_2})_{ij}P[n_j(t_2)] - \sum_{t=0}^{\infty} \delta_{ij}P[n_i(t_1)] \\ &= \sum_{t_1=0}^{\infty} P[n_i(t_1)] \sum_{\Delta t=0}^{\infty} (\tilde{P}^{\Delta t})_{ij} + \sum_{t_2=0}^{\infty} P[n_j(t_2)] \sum_{\Delta t=0}^{\infty} (\tilde{P}^{\Delta t})_{ji} - \delta_{ij} \sum_{t=0}^{\infty} P[n_i(t_1)] \\ &= (\mathbf{I} - \tilde{\mathbf{P}})_{ij}^{-1} E_i + (\mathbf{I} - \tilde{\mathbf{P}})_{ji}^{-1} E_j - \delta_{ij} E_i \end{aligned} \quad (14)$$

Unfortunately, as can be seen from the above expression, the second moments cannot be tied to the first moments via a linear function. Therefore, we cannot use the moments as the optimization variables directly. Instead, we are going to work with the following asymmetric terms where the order of indexes of M corresponds to an temporal ordering of the terms in the sums:

$$\begin{aligned} M_i &= \sum_{t=0}^{\infty} \langle n_i(t) \rangle = \sum_j \alpha_j (\mathbf{I} - \tilde{\mathbf{P}})_{ij}^{-1} \\ M_{ij} &= \sum_{t_1=0}^{\infty} \sum_{t_2=t_1}^{\infty} \langle n_i(t_1) n_j(t_2) \rangle = \sum_{t_1=0}^{\infty} \sum_{t_2=t_1}^{\infty} P[n_j(t_2)|n_i(t_1)]P[n_i(t_1)] = M_i (\mathbf{I} - \tilde{\mathbf{P}})_{ij}^{-1} \\ M_{i_1 i_2 \dots i_{k-1} i_k} &= M_{i_1 \dots i_{k-1}} (\mathbf{I} - \tilde{\mathbf{P}})_{i_{k-1} i_k}^{-1} \end{aligned} \quad (15)$$

We will refer to the above terms as the *asymmetric moments* (although they do not correspond to moments of any real variables). Clearly, all of the k^{th} order asymmetric moments can be expressed as a linear function of the asymmetric moments of order $k-1$ by moving the $(\mathbf{I} - \tilde{\mathbf{P}})$ term to the left-hand side. For example, for the second moments this step can be easily done by rewriting (eq. 15) in matrix form:

$$\mathbf{M}'' = \mathbf{D}(M_i)(\mathbf{I} - \tilde{\mathbf{P}})^{-1}, \quad (16)$$

where $\mathbf{M}'' = [M_{ij}]$ is the matrix of second order asymmetric moments, and $\mathbf{D}(M_i)$ is a diagonal matrix, with values of the first order moment M_i on the diagonal. Multiplying by $(\mathbf{I} - \tilde{\mathbf{P}})$ on the right, we get $\mathbf{M}'' - \mathbf{M}'' \tilde{\mathbf{P}} = \mathbf{D}(M_i)$. Similarly, for the other moments we get:

$$\begin{aligned} M_i - \sum_j M_j \tilde{P}_{ji} &= \alpha_i \\ M_{ij} - \sum_k M_{ik} \tilde{P}_{kj} &= \delta_{ij} M_i \\ M_{i_1 i_2 \dots i_{k-1} i_k} - \sum_j M_{i_1 i_2 \dots i_{k-1}} \tilde{P}_{i_{k-1} j} &= \delta_{i_1 i_k} M_{i_1 i_2 \dots i_{k-1}} \end{aligned} \quad (17)$$

Furthermore, it can be seen that the true moments of order k can be expressed as linear functions of the asymmetric moments of orders 1 through k :

$$\begin{aligned} E_i &= M_i, \quad E_{ij} = M_{ij} + M_{ji} - \delta_{ij} M_i \\ E_{ijk} &= M_{ijk} + M_{ikj} + M_{jik} + M_{jki} + M_{kij} + M_{kji} - \delta_{ij} M_{ik} - \delta_{ij} M_{ki} - \delta_{ik} M_{ij} - \delta_{ik} M_{ji} - \\ &\quad \delta_{jk} M_{ji} - \delta_{jk} M_{ij} + \delta_{ijk} M_i \end{aligned} \quad (18)$$

Indeed, for the first moments, there is only one state index and therefore the first asymmetric moment equals the true moment. For the second moments, both M_{ij} and M_{ji} include the term $(t_1 = t_2)$, and thus we have to subtract $\sum_{t_1} \sum_{t_2=t_1} \langle n_i(t_1) n_j(t_2) \rangle = \delta_{ij} M_i$. The expressions for other moments are obtained in a similar manner.

The last step that remains in formulating the optimization program is to substitute the transition probabilities $\tilde{P}_{ij} = \sum_a P_{ij}\pi_{ia}$ and to define the actual optimization variables. This is where we hit a problem that breaks the linearity of the program. Recall that for the standard CMDP, the optimization variables are defined as $x_{ia} = E_i\pi_{ia} = M_i\pi_{ia}$ and have the interpretation of the expected total number of times action a is executed in state i . As mentioned earlier, this allows one to express the first-order constraint from (eq. 17) as a linear function of x_{ia} . Indeed, since $\sum_a \pi_{ia} = 1$, the first moments are simply $M_i = \sum_a x_{ia}$, and recalling that $x_{ia} = E_i\pi_{ia} = M_i\pi_{ia}$, we have for the first order constraint:

$$M_i - \sum_j M_j \tilde{P}_{ji} = \sum_a x_{ia} - \sum_j M_j \sum_a P_{ji}\pi_{ja} = \sum_a x_{ia} - \sum_j \sum_a P_{ji}x_{ja} \quad (19)$$

Unfortunately, the same trick does not work for the higher order constraints. If we were to define similar variables for the higher-order moments (ex. $x_{ija} = M_{ij}\pi_{ja}$), we could, of course, rewrite (eq. 17) as linear functions of these variables. However, by doing this, we would introduce too many free parameters into the program. To retain the original desired interpretation of the variables, we would also have to add constraints to ensure that the policy π implied by the higher-order variables is the same as the one computed from the first-order x_{ia} . Clearly, these new constraints would be quadratic:

$$\frac{x_{ja_1}}{x_{ja_2}} = \frac{x_{ija_1}}{x_{ija_2}} = \dots = \frac{\pi_{ja_1}}{\pi_{ja_2}} \quad (20)$$

Hence, it appears that there is no easy way to avoid the quadratic expressions ($M_{ij}\pi_{ja}$) in the constraints on the moments:

$$\begin{aligned} \sum_a x_{ia} - \sum_j \sum_a x_{ia} P_{ji}^a &= \alpha_i \\ M_{ij} - \sum_k M_{ik} \sum_a P_{kj}^a \pi_{ka} &= \delta_{ij} M_i \\ M_{ijk} - \sum_l M_{ijl} \sum_a P_{lk}^a \pi_{la} &= \delta_{ik} M_{ij} \\ M_{i_1 i_2 \dots i_{k-1} i_k} - \sum_j M_{i_1 i_2 \dots i_{k-1}} \sum_a P_{i_{k-1} j}^a \pi_{i_{k-1} a} &= \delta_{i_1 i_k} M_{i_1 i_2 \dots i_{k-1}} \\ \pi_{ia} x_{i0} &= \pi_{i0} x_{ia} \end{aligned} \quad (21)$$

We are therefore left with an optimization program in x_{ia} and the asymmetric moments ($M_{ij}, M_{ji}, M_{ijk}, \dots$) that has: a linear objective function $\sum_i \sum_a x_{ia} r_{ia}$, a constraint on the probability of the total cost exceeding a given threshold, which is linear in the moments E (eq. 7), which are linear in M (eq. 18), and a system of quadratic constraints that synchronizes the moments (eq. 21).

6 An Example

As an example of the use of the method presented in the previous sections, let us consider a toy problem, for which we can analytically compute the distribution of the total cost, and formulate a constrained optimization program for it using the first three moments of the total cost. In this section we present a more careful derivation of the optimization program, paying more attention to some steps that were just briefly described in section 4. We also present a preliminary empirical analysis that shows how closely our model approximates the true cumulative probability distribution function of the total cost. The purpose of the latter is to serve as a rough indication of the accuracy of our approach, which we cannot yet directly report on, as at the time of writing we do not have the optimization implemented yet.

Consider the problem depicted in figure 1(a). In this problem, there are two states, one of which ($i = 2$) is a sink state. If the agent starts in state 1 ($\alpha = [1, 0]$), the total received reward is the same as the total incurred cost, and both equal the total number of visits to state 1. The obvious optimal policy for the unconstrained problem is to always execute action 1 in state 1 ($\pi = [1, 0, 1, 0]$). If we set the upper bound on the cost as $C_0 = 10$, the unconstrained optimal policy has about a 30% chance of exceeding that bound. As we decrease the acceptable probability of exceeding the cost bound, the policies should become more conservative, i.e. they should prescribe higher probabilities of executing action 2 in state 1.

In order to apply the Legendre approximation from section 4, one has to ensure that the total cost C is in the range $[-1, 1]$. Clearly, this is not generally the case for our problem. Therefore, to satisfy this condition, we apply a transformation

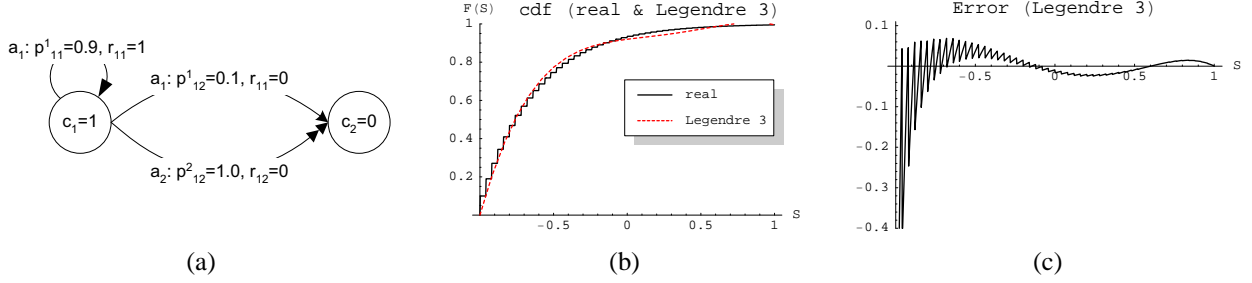


Figure 1: (a) – simple problem with two states and two actions; (b) – the actual cdf of the total cost and the cdf computed from a third-degree Lagrange approximation of the pdf of the cost; (c) – relative error of the approximation in (b).

$S = 2C/C_{max} - 1$ with $C_{max} = 50$ (a reasonable approximation, as $P[C \geq 50] = 0.004$).⁷ Figure 1(b) shows the resulting cumulative distribution function $F_S(S)$ for the unconstrained optimal policy and the cdf computed from a third-degree approximation of the pdf $f(S)$. Figure 1(c) shows the relative error of the third-degree approximation and serves to show that we can expect to get a reasonably good approximation of the cdf using just the first three moments. Note that the pdf for this problem is discrete (thus, harder to approximate with continuous Legendre polynomials), but we can still get a reasonably good approximation of the cdf, which is what we really care about, as our goal is to estimate $P(S > S_0) = 1 - F_S(S_0)$. Let us now compute a third-order approximation to $f_S(S)$ for an arbitrary policy by computing the coefficients b_l (eq. 5):

$$b_0 = \frac{1}{2}E_S^0, \quad b_1 = \frac{3}{2}E_S^1, \quad b_2 = -\frac{5}{4}E_S^0 + \frac{15}{4}E_S^2, \quad b_3 = -\frac{21}{4}E_S^1 + \frac{35}{4}E_S^3. \quad (22)$$

Notice that here we have to use the moments of $S \in [-1, 1]$. However, the constraints in (eq. 21) operate on moments of C , and since our optimization variables are going to be the moments of $C \in [0, C_{max}]$, we have to be able to express the former via the latter. This can be easily done by solving the following linear system of equations for E_S :

$$\begin{aligned} E_C^0 &= \int_0^{C_{max}} f_C(C) dC = \frac{1}{2}C_{max} \int_{-1}^1 f_S(S) dS = \frac{1}{2}C_{max} E_S^0, \\ E_C^1 &= \frac{1}{4}C_{max}^2 (E_S^1 + E_S^0), \quad E_C^2 = \frac{1}{8}C_{max}^3 (E_S^2 + 2E_S^1 + E_S^0), \quad E_C^3 = \frac{1}{16}C_{max}^4 (E_S^3 + 3E_S^2 + 3E_S^1 + E_S^0) \end{aligned} \quad (23)$$

Now, the probability of exceeding the cost bounds is simply $P[C \geq C_0] = \int_{C_0}^{C_{max}} f_C(C) dC = C_{max}/2 \int_{S_0}^1 f_S(S) dS$, where $S_0 = 2C_0/C_{max} - 1$.

7 Conclusions

In this paper we have introduced a method for approximately reasoning about the probability distributions of rewards and costs in Markov decision processes. The main three sources of approximation error in our method are: 1) the use of Legendre polynomials to approximate the true pdf, 2) the use of a finite number of moments (the more moments are used, the better the approximation), and 3) truncation of the costs at some upper bound C_{max} (the lower the mass of the remaining cost $\int_{C_{max}}^{\infty} C f_C(C)$, the better the approximation).

We demonstrated the approach on a specific problem that bounds the probability that the total cost exceeds a given upper bound. However, it is easy to see that the method allows one to model a wide range of risk-sensitive objective functions and constraints. Indeed, one can just as easily approximate the distribution of the total (normalized) reward and express the expected value of the utility function as (similarly to (eq. 7)):

$$E[u(R)] = \int_{-1}^1 u(R) f_R(R) dR = \int_{-1}^1 \sum_{l=0}^{\infty} \sum_k a_{kl} E_R^k u(R) \mathcal{P}_l(R) dR = \sum_k \phi_k E_R^k, \quad (24)$$

where $\phi_k = \sum_l a_{kl} \int_{-1}^1 u(R) \mathcal{P}_l(R) dR$, $u(R)$ is the utility of getting reward R , and $f_R(R)$ is the distribution of the total reward. Then, the approximation methods of sections 4 and 5 are directly applicable. Of course, the above relies on the fact

⁷For a transient system with bounded rewards $\lim_{C_0 \rightarrow \infty} P[C > C_0] = 0$. Thus one can always compute or estimate a reasonable C_{max} for which $P[C > C_{max}]$ is arbitrary small.

that there either exists a natural upper bound on the largest possible utility value of a policy, or that there exists an upper bound R_{max} such that the weighted tail of the utility distribution $\int_{R_{max}}^1 f_R(R)u(R)dR$ is sufficiently small.

Even though our construction yields more complex optimization programs than the standard constrained MDP approach, it is more expressive than the standard risk-neutral CMDP techniques, because our formulation allows one to reason about the probability distributions instead of the expected values of the total cost and rewards. Our ongoing efforts in extending this work include several directions such as looking at ways of efficiently encoding and implementing the optimization program, more careful investigation of the complexity and convergence properties of the model (as more moments are used), exploring heuristics for choosing an appropriate number of moments, and a formal analysis of properties of the problem and the corresponding solutions.

Acknowledgments This work was supported by DARPA and the Air Force Research Laboratory under contract F30602-00-C-0017 as a subcontractor through Honeywell Laboratories and also by Honeywell Laboratories through the “Industrial Partners of CSE” program at the University of Michigan. The authors thank the anonymous reviewers for their comments.

References

- [1] E. Altman. *Constrained Markov Decision Processes*. Chapman and HALL/CRC, 1999.
- [2] E. Altman and A. Shwartz. Adaptive control of constrained Markov chains: Criteria and policies. *Annals of Operations Research, special issue on Markov Decision Processes*, 28:101–134, 1991.
- [3] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.
- [4] T. Dean and M. Wellman. *Planning and Control*. Morgan Kaufmann, 1991.
- [5] D. A. Dolgov and E. H. Durfee. Approximating optimal policies for agents with limited execution resources. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 1107–1112, 2003.
- [6] J. A. Filar and L. C. M. Kallenberg. Variance-penalized markov decision processes. *Math. of OR*, 14:147–161, 1989.
- [7] J. A. Filar and H. M. Lee. Gain/variability tradeoffs in undiscounted markov decision processes. In *Proceedings of 24th Conference on Decision and Control IEEE*, pages 1106–1112, 1985.
- [8] R. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, 1960.
- [9] R. Howard and J. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972.
- [10] Y. Huang and L. Kallenberg. On finding optimal policies for Markov decision chains. *Math. of OR*, 19:434–448, 1994.
- [11] L. Kallenberg. *Linear Programming and Finite Markovian Control Problems*. Math. Centrum, Amsterdam, 1983.
- [12] S. Koenig and R. G. Simmons. Risk-sensitive planning with probabilistic decision graphs. In J. Doyle, E. Sandewall, and P. Torasso, editors, *KR’94: Principles of Knowledge Representation and Reasoning*, pages 363–373. Morgan Kaufmann, San Francisco, California, 1994.
- [13] S. Marcus, E. Fernandez-Gaucherand, D. Hernandez-Hernandez, S. Colaruppi, and P. Fard. Risk-sensitive markov decision processes, 1997.
- [14] J. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1-2):122–136, 1964.
- [15] M. L. Puterman. *Markov Decision Processes*. John Wiley & Sons, New York, 1994.
- [16] K. Ross and B. Chen. Optimal scheduling of interactive and non-interactive traffic in telecommunication systems. *IEEE Transactions on Auto Control*, 33:261–267, 1988.
- [17] K. Ross and R. Varadarajan. Multichain Markov decision processes with a sample path constraint: A decomposition approach. *Math. of Operations Research*, 16:195–207, 1991.
- [18] M. Sobel. Maximal mean/standard deviation ratio in undiscounted mdp. *OR Letters*, 4:157–188, 1985.
- [19] D. J. White. A mathematical programming approach to a problem in variance penalised markov decision processes. *OR Spectrum*, 15:225–230, 1994.