Probabilistic Goal Markov Decision Processes*

Huan Xu
Department of Mechanical Engineering
National University of Singapore, Singapore
mpexuh@nus.edu.sg

Shie Mannor
Department of Electrical Engineering
Technion, Israel
shie@ee.technion.ac.il

Abstract

The Markov decision process model is a powerful tool in planing tasks and sequential decision making problems. The randomness of state transitions and rewards implies that the performance of a policy is often stochastic. In contrast to the standard approach that studies the expected performance, we consider the policy that maximizes the probability of achieving a pre-determined target performance, a criterion we term *probabilistic goal Markov decision processes*. We show that this problem is NP-hard, but can be solved using a pseudo-polynomial algorithm. We further consider a variant dubbed "chance-constraint Markov decision problems," that treats the probability of achieving target performance as a constraint instead of the maximizing objective. This variant is NP-hard, but can be solved in pseudo-polynomial time.

1 Introduction

The Markov Decision Process (MDP) model is a powerful tool in planning tasks and sequential decision making problems [Puterman, 1994; Bertsekas, 1995]. In MDPs, the system dynamics is captured by transition between a finite number of states. In each decision stage, a decision maker picks an action from a finite action set, then the system evolves to a new state accordingly, and the decision maker obtains a reward. Both the state transition and the immediate reward are often random in nature, and hence the cumulative reward X may be inherently stochastic. Classical approach deals with the maximization of the *expected value* of X, which implicitly assumes that the decision maker is risk-neutral.

In this paper we propose and study an alternative optimization criterion: fix a target level $V \in \mathbb{R}$, the decision goal is to find a policy π that maximizes the probability of achieving the target, i.e., to maximize $Pr(X_\pi \geq V)$. The criterion of maximizing the probability of achieving a goal, which we call *probabilistic goal*, is an intuitively appealing objective, and justified by axiomatic decision theory [Castagnoli and LiCalzi, 1996]. Moreover, empirical research has also concluded that in daily decision making, people tends to regard *risk* primarily as failure to achieving a predetermined goal [Lanzillotti, 1958; Simon, 1959; Mao, 1970; Payne *et al.*, 1980; Payne *et al.*, 1981].

Different variants of the probabilistic goal formulation such as chance constrained programs[1], have been extensively studied in *single-period optimization* [Miller and Wagner, 1965; Prékopa, 1970]. However, little has been done in the context of sequential decision problem including MDPs. The standard approaches in risk-averse MDPs include maximization of expected utility function [Bertsekas, 1995], and optimization of a coherent risk measure [Riedel, 2004; Le Tallec, 2007]. Both approaches lead to formulations that can not be solved in polynomial time, except for special cases including exponential utility function [Chung and Sobel, 1987], piecewise linear utility function with a single break down point [Liu and Koenig, 2005], and risk measures that can be reduced to robust MDPs satisfying the so-called "rectangular condition" [**?**; Iyengar, 2005].

Two notable exceptions that explicitly investigate the probabilistic goal or chance constrained formulation in the context of MDPs are [Filar *et al.*, 1995] and [Delage and Mannor, 2010]. The first reference considers the average reward case of MDPs, and reduces the probability of achieving a target to the probability of entering irreducible chains. The analysis is thus very specific and can not be extended to finite horizon case or discounted reward infinite horizon case. The second reference investigated, instead of the *internal randomness*, the *parametric uncertainty* of the cumulative reward. That is, the parameters of the MDP, namely the transition probability and the reward, are not precisely known. While the decision maker is risk-neutral to internal randomness – performance fluctuation due to the stochasticity of the state transition, immediate reward and possible randomness of the action, he/she is risk averse to the performance deviation due to the parameter uncertainty. Conceptually, in this setup one can think of of having many identical machines all with the same uncertain parameter, and the goal is to maximize the probability that the *average reward* (w.r.t. all machines) achieves a certain target. Because of this apparent difference in modeling,

*H. Xu is supported by an NUS startup grant R-265-000-384-133. S. Mannor is supported by the Israel Science Foundation under contract 890015.

[1] Instead of maximizing the probability of achieving a target, chance constrained program requires such probability to be no less than a fixed threshold.

the techniques used in [Delage and Mannor, 2010] do not extend to the probabilistic goal of the internal randomness, the setup that we consider.

To the best of our knowledge, the probabilistic goal or chance constrained formulation has not been investigated for finite-horizon or infinite-horizon discounted reward MDPs. Our contributions include the following:

1. We compare different policy sets: we show that for a probabilistic goal MDP, an optimal policy may depend on accumulated reward, and randomization does not improve the performance.
2. We show that the probabilistic goal MDP is NP-hard. Thus, it is of little hope that such problem can be solved in polynomial time in general.
3. We propose a pseudo-polynomial algorithm based on state-augmentation, that solves the probabilistic goal MDP.
4. We investigate chance constrained MDPs and show it can be solved in pseudo polynomial time.

Before concluding this section, let us briefly discuss some practical examples that motivate this study. Consider the following scenario that many of us have experienced. Suppose one wants to drive to the airport to catch a plane which is soon to take off. He/she needs to decide which route to take, while the time he/she will spend on each link is random. This can be modeled as an MDP with a deterministic transition probability and random reward. It is clear that the expected time the decision maker will spend on route is less critical than whether he/she will catch the plane, a natural fit for probabilistic goal MDP. Other examples can be found in finance, where synthetic derivatives that are triggered by discrete events are becoming common, and hence minimizing or maximizing the probability of such event is relevant and important; and airplanes design, where one seek a reconfiguration policy that maximizes the chance of not having a critical failure.

2 Setup

In this section we present the problem setup and some necessary notations. A (finite) MDP is a sextuple $< T, \mathcal{S}, \mathcal{A}, \mathcal{R}, p, g >$ where

- T is the time horizon, assumed to be finite;
- $\mathcal{S}$ is a finite set of states, with $s_0 \in \mathcal{S}$ the initial state;
- $\mathcal{A}$ is a collection of finite action sets, one set for each state. For $s \in \mathcal{S}$, we denote its action set by $\mathcal{A}_s$;
- $\mathcal{R}$ is a finite subset of $\mathbb{R}$, and is the set of possible values of the immediate rewards. We let $K = \max_{r\in\mathcal{R}} |r|$;
- p is the transition probability. That is, $p_t(s'|s, a)$ is the probability that the state at time $t+1$ is s', given that the state at time t is s, and the action chosen at time t is a.
- g is a set of reward distributions. In particular, $g_t(r|s, a)$ is the probability that the immediate reward at time t is r, if the state is s and action is a.

As standard, we use symbol π to denote a policy of an MDP. A *history-dependent, deterministic* policy is a mapping from the history $H_t = (s_{0:t}, a_{0:t}, r_{0:t})$ of the process to an action $a \in \mathcal{A}_{s_t}$. We use Π^h to represent the set of history-dependent deterministic policies. The set of deterministic policies that only depend on the current state (and the time horizon), which is often called Markovian policies, is denoted by $\Pi^{t,s}$. As we show in the sequel, it is sometimes beneficial to incorporate the accumulated reward in the decision. The set of policies that depend on the time horizon, the current state, and the accumulated reward up-to-now, which we called "pseudo-Markovian" policy, is denoted by $\Pi^{t,s,x}$.

Besides deterministic policy, we also considered randomized policy. That is, assuming there is available a sequence of i.i.d. random variables $U_1, \cdots, U_T$, independent to everything else, and a history-dependent, randomized policy is a mapping from (H_t, U_t) to an action. We denote the set of such policies by $\Pi^{h,u}$. The set of Markovian and pseudo-Markovian randomized policies are defined similarly, and denoted respectively by $\Pi^{t,s,u}$ and $\Pi^{t,s,x,u}$.

Note that given the distribution of the reward parameter, the total reward of the MDP under a policy π is a well defined random variable, denoted by X_π. We are interested in the following problems. As standard in the study of computational complexity, the first problem we consider is a "yes/no" decision problem.

Problem 1 (Decision Problem). *Given $V \in \mathbb{R}$ and $\alpha \in (0, 1)$. Is there a $\pi \in \Pi^{h,u}$ such that*

$$Pr(X_\pi \geq V) \geq \alpha?$$

We call this problem $D(\Pi^{h,u})$. Similarly, we define $D(\Pi^{t,s,u})$, $D(\Pi^{t,s,x,u})$, and their deterministic counterparts.

A related problem of more practical interest is the optimization one.

Problem 2 (probabilistic goal MDP). *Given $V \in \mathbb{R}$, find $\pi \in \Pi^{h,u}$ that maximizes*

$$Pr(X_\pi \geq V).$$

Let us remark that while we focus in this paper the finite horizon case, most results easily extend to infinite horizon discounted reward case, due to the fact that the contribution of the tail of the time horizon can be made arbitrarily small by increasing the time horizon.

3 Comparison of policy sets

In this section we compare different policy sets. We say a policy set Π is "inferior" to another Π', if $D(\Pi)$ being true implies that $D(\Pi')$ is true, and there exists an instance where the reverse does not hold. We say two policy sets Π and Π' are "equivalent" if $D(\Pi)$ is true if and only if $D(\Pi')$ is true.

We first show that randomization does not help: Π^h is equivalent to $\Pi^{h,u}$, $\Pi^{t,s,x}$ is equivalent to $\Pi^{t,s,x,u}$, and similarly $\Pi^{t,s}$ is equivalent to $\Pi^{t,s,u}$. This essentially means that for probabilistic goal MDP, it suffices to focus on Π^h, $\Pi^{t,s,x}$ and $\Pi^{t,s}$. Note that it suffices to show the following theorem, since the reverse direction holds trivially.

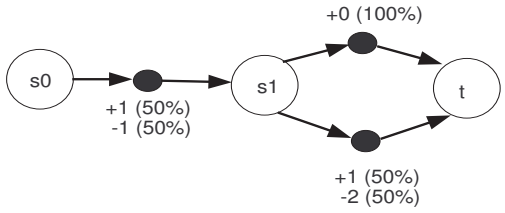

Figure 1: Illustration of Example 1

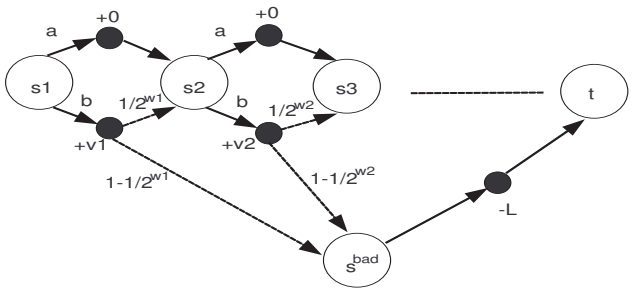

Figure 2: Example of NP-hard probabilistic goal MDP

Theorem 1. *Given an MDP, $\alpha \in [0,1]$, and V, if there exists $\pi_u \in \Pi^{h,u}$ (respectively $\Pi^{t,s,x,u}$ and $\Pi^{t,s,u}$) such that*

$$Pr(X_{\pi_u} \geq V) \geq \alpha,$$

then there exists $\pi \in \Pi^h$ (respectively $\Pi^{t,s,x}$ and $\Pi^{t,s}$) such that

$$Pr(X_{\pi} \geq V) \geq \alpha.$$

Proof. Since $T < \infty$, the set Π^h is finite, i.e., there is a finite number of deterministic history dependent policies. For succinctness of presentation, we write $\pi_u \sim \pi$, where $\pi_u \in \Pi^{h,u}$ and $\pi \in \Pi^h$, to denote the event (on probability space of U) that $\pi_u(H_t, U_{0:t}) = \pi(H_t)$ for all t and H_t. Fix a randomized policy $\pi \in \Pi^{h,u}$, due to theorem of total probability we have

$$\begin{aligned} Pr(X_{\pi_u} \geq V) &= \sum_{\pi \in \Pi^h} Pr(X_{\pi_u} \geq V | \pi_u \sim \pi) Pr(\pi_u \sim \pi) \\ &= \sum_{\pi \in \Pi^h} Pr(X_{\pi} \geq V) Pr(\pi_u \sim \pi). \end{aligned}$$

Note that $\sum_{\pi \in \Pi^h} Pr(\pi_u \sim \pi) = 1$, we have that

$$\max_{\pi \in \Pi^h} Pr(X_{\pi} \geq V) \geq Pr(X_{\pi_u} \geq V),$$

which establishes the theorem for the history-dependent case. The other two cases are essentially same and hence omitted.

It is straightforward to generalize the finite horizon case to the discounted infinite horizon case, due to the fact that the contribution of the tail of the time horizon can be made arbitrarily small. (The proof is by contradiction: if there is a difference in performance in the limit, it must appear in finite time as well.) □

We next show that in general, $\Pi^{t,s}$ is inferior to $\Pi^{t,s,x}$. This essentially means that including the information of accumulated reward can improve the performance of a policy for probabilistic goal MDP.

Example 1. Consider the MDP as in Figure 1: $S = \{s_0, s_1, t\}$, where t is the terminal state. The initial state s_0 has one action, which leads to state s_1, and the immediate reward is that with probability 0.5 it is $+1$, and otherwise -1. State s_1 has two actions a and b, both lead to state t. The immediate reward of a is always 0, and that of b is that with probability 0.5 it is $+1$, and otherwise -2. Observe that for either fixed action a or b, $Pr(X \geq 0) = 0.5$. In contrast, consider the policy π that at state s_1, takes action a if the accumulated reward is $+1$, otherwise takes action b. Observe that $Pr(X_\pi \geq 0) = 0.75$, which shows that $\Pi^{t,s}$ is inferior to $\Pi^{t,s,x}$.

Finally, we remark that adding extra information beyond the cumulative reward does not improve the performance for probabilistic goal MDP, as we will show in Section 5. The policies clasees $\Pi^{t,s,x}$ and Π^h are therefore equivalent.

4 Complexity

In this section we show that in general, solving the probabilistic goal MDP is a computationally hard problem.

Theorem 2. *The problem $D(\Pi)$ is NP-hard, for Π equals to Π^h, $\Pi^{h,u}$, $\Pi^{t,s}$, $\Pi^{t,s,u}$, $\Pi^{t,s,x}$ or $\Pi^{t,s,x,u}$.*

Proof. We start with proving the result for $\Pi^{t,s}$, by showing that we can reduce a well-known NP complete problem, KnapSack Problem (KSP), into a probabilistic goal MDP, and hence establish its NP-hardness. Recall that a KSP is the following problem. There are n items, 1 through n, each item i has a value v_i and a weight w_i. We can assume they are all non-negative integers. The KSP problem is to decide given two positive number W and V, whether there exists a $I \subseteq [1:n]$ such that

$$\sum_{i \in I} w_i \leq W; \quad \sum_{i \in I} v_i \geq V.$$

Given n, w_i, v_i, W and V, we construct the following MDP. Let $T = n + 2$. There are $n + 2$ states: $\mathcal{S} = \{s_1, \cdots, s_n, s^{bad}, t\}$: s_1 is the initial state and t is the terminal state. Each state s_i has two actions: if action a is taken, then the immediate reward is 0 and the next state will be s_{i+1}; if action b is taken, then the immediate reward is v_i; furthermore, with probability $1/2^{w_i}$ the next state will be s_{i+1} (respectively terminal state t if $i = n$) and otherwise the next state will be s^{bad}. The state s^{bad} has only one action which incurs an immediate reward $-L \triangleq -2\sum_{i=1}^{n} v_i$, and the next state will be t. See Figure 2.

Now consider the decision problem $D(\Pi^{t,s})$ with $\alpha = 1/2^W$. That is, to answer whether there exists $\pi \in \Pi^{t,s}$ such that

$$Pr(X_\pi \geq V) \geq \alpha.$$

We now show that the answer to $D(\Pi^{t,s})$ is positive if and only if the answer to KSP is positive.

Suppose the answer to $D(\Pi^{t,s})$ is positive, i.e., there exists a policy $\pi \in \Pi^{t,s}$ such that

$$Pr(X_\pi \geq V) \geq \alpha.$$

Define set I' as all $i \in [1:n]$ such that π takes b in s_i. Observe this implies that

$$\sum_{i \in I'} v_i \geq V.$$

Furthermore, due to the extreme large negative reward in s^{bad},

$$\begin{aligned} Pr(X_\pi \geq V) &\leq Pr(s^{bad} \text{ is never reached.}) \\ &= \Pi_{i \in I'} \frac{1}{2^{w_i}} = \frac{1}{2^{\sum_{i \in I'} w_i}}, \end{aligned}$$

which implies that

$$\frac{1}{2^{\sum_{i \in I'} w_i}} \geq 1/2^W \quad \Rightarrow \quad \sum_{i \in I'} w_i \leq W.$$

Thus, the answer to KSP is also positive.

Now suppose the answer to KSP is positive, i.e., there exists $I \subset [1:n]$ such that

$$\sum_{i \in I} w_i \leq W; \quad \sum_{i \in I} v_i \geq V.$$

Consider the following policy π' of the MDP: take action b for all s_i where $i \in I$, and a otherwise. We have

$$\begin{aligned} &Pr(s^{bad} \text{ is never reached.}) \\ &= \prod_{i \in I} \frac{1}{2^{w_i}} = \frac{1}{2^{\sum_{i \in I} w_i}} \geq 1/2^W = \alpha. \end{aligned}$$

Furthermore, when s^{bad} is never reached, the cumulative reward is $\sum_{i \in I} v_i \geq V$. Thus, we have that

$$Pr(X_{\pi'} \geq V) \geq \alpha,$$

i.e., the answer to $D(\Pi^{t,s})$ is also positive.

Thus, determining the answer to KSP is reduced to answering $D(\Pi^{t,s})$, the probabilistic goal MDP. Since the former is NP-complete, we conclude that probabilistic goal MDP is NP-hard.

Notice that in this example, $\Pi^{t,s} = \Pi^{t,s,x} = \Pi^h$. Thus, NP-hardness for the decision problems with respect to these policy sets are established as well. Furthermore, Theorem 1 implies that $D(\Pi^{h,u})$, $D(\Pi^{t,s,u})$, $D(\Pi^{t,s,x,u})$ are also NP-hard. □

5 Pseudo-polynomial algorithms

In the previous section we showed that it is of little hope that the probabilistic goal MDP can be solved in polynomial time. In this section we develop a pseudo-polynomial algorithm to handle this question. Recall that the running time of a pseudo-polynomial algorithm is polynomial in the number of parameters and the size of the parameter, as opposed to polynomial algorithms whose running time is polynomial in the number of parameters and *polylogarithmic* in the size of the parameters.

5.1 Integer Reward

The main technique of our pseudo-polynomial algorithm is state augmentation. That is, we construct a new MDP in which the state space is the Cartesian product of the original state-space and the set of possible accumulated rewards. To better understand the algorithm, we first consider in this subsection a special case where the immediate reward is integer-valued, i.e., $\mathcal{R} \subset \mathbb{Z}$. We show that the probabilistic goal MDP can be solved in time polynomial to $|\mathcal{S}|$, $|\mathcal{A}|$ and K.

Theorem 3. *Suppose that $\mathcal{R} \subset \mathbb{Z}$. In computational time polynomial in T, $|\mathcal{S}|$, $|\mathcal{A}|$ and K, one can solve $D(\Pi^{h,u})$, i.e., determine the correctness of the following claim:*

$$\exists \pi \in \Pi^{h,u} : \ Pr(X_\pi \geq V) \geq \alpha.$$

Similarly, in computational time polynomial to T, $|\mathcal{S}|$, $|A|$ and K, one can solve the probabilistic goal MDP, i.e., find $\pi \in \Pi^{h,u}$ that maximizes $Pr(X_\pi \geq V)$. Moreover, there exists an optimal solution to probabilistic goal MDP that belongs to $\Pi^{t,s,x}$.

Proof. It suffices to show that an optimal solution to the probabilistic goal MDP

$$\pi^* = \arg\max_{\pi \in \Pi^h} Pr(X_\pi \geq V),$$

together with the optimal value $Pr(X_{\pi^*} \geq V)$, can be obtained by solving an MDP with $2TK|\mathcal{S}|$ states.

We construct a new MDP as follows: each state is a pair (s, C) where $s \in S$ and C is an integer in $[-TK, TK]$, and the initial state is $(s_0, 0)$. The action set for (s, C) is $\mathcal{A}_s$, for all C. In time t, the transition between states is defined as follows:

$$Pr\big((s', C')\big|a, (s, C)\big) = p_t(s'|a, s) \times g_t(C' - C|s, a).$$

That is, a transition to (s', C') happens if in the original MDP, a transition to state s' happens and the immediate reward is $C' - C$. Notice that since the immediate reward of the original MDP is bounded in $[-K, K]$ and there are only T stages, the accumulated reward of the original MDP is bounded in $[-TK, TK]$. The immediate-reward of the new MDP is zero except in the last stage, in which for states (s, C) with $C \geq V$, a reward $+1$ is incurred.

It is easy to see that the solution to probabilistic goal MDP is equivalent to a policy that maximizes the expected reward for the new MDP. Note that there exists a deterministic, Markovian policy to the latter, then the probabilistic goal MDP has an optimal solution in $\Pi^{t,s,x}$. □

Theorem 3 leads to the following algorithm for solving probabilistic goal MDPs. Note that, not surprisingly, the proposed algorithm indeed parallels the standard algorithm (also pseudo-polynomial) solving KSP.

5.2 General Reward

In this section we relax the assumption that the reward is integer valued. We discretize the real-valued reward parameters to obtain a new MDP that approximates the original one. The new MDP is equivalent to an integer valued MDP, and hence can be solved in pseudo-polynomial time thanks to results

- Input: MDP $< T, \mathcal{S}, \mathcal{A}, \mathcal{R}, p, g >$ with $\mathcal{R} \subset \mathbb{Z}$, $V \in \mathbb{R}$.
- Output: $\pi \in \Pi^{t,s,w}$ that maximizes $Pr(X_\pi \geq V)$.
- Algorithm:
  1. Construct MDP with augmented state-space.
  2. Find a Markovian $\hat{\pi}^*$ that maximizes the expected reward of the new MDP.
  3. Construct the policy of the original MDP as follows: the action taken at stage t, state s, with an accumulated reward C equals to $\hat{\pi}_t(s, C)$. Output the policy.

in Section 5.1. To show that such approximation is legitimate, we bound the performance gap that diminishes as the discretization becomes finer.

Theorem 4. *There exists an algorithm, whose running time is polynomial in T, $|\mathcal{S}|$, $|\mathcal{A}|$, K and $1/\epsilon$, that finds a policy $\hat{\pi}$, such that*

$$Pr(X_\pi \geq V + \epsilon) \leq Pr(X_{\hat{\pi}} \geq V); \quad \forall \pi \in \Pi^{h,u}.$$

Proof. Given an MDP $\langle T, \mathcal{S}, \mathcal{A}, \mathcal{R}, p, g \rangle$, with $K = \max_{r \in \mathcal{R}} |r|$, and the grid size $\delta > 0$, we can construct a new MDP $< T, \mathcal{S}, \mathcal{A}, \hat{\mathcal{R}}, p, \hat{g} >$, where $\hat{\mathcal{R}} = \{i\delta | i \in \mathbb{Z};\ -K/\delta \leq i \leq K/\delta\}$, and

$$\hat{g}_t(r|s,a) = \begin{cases} \sum_{r': r \leq r' < r+\delta} g_t(r'|s,a), & \text{if } r \in \hat{\mathcal{R}}; \\ 0, & \text{otherwise.} \end{cases}$$

Observe that, by scaling the parameters, the new MDP is equivalent to an integer reward MDP whose reward parameters are bounded by K/δ. Hence the probabilistic goal MDP for the new MDP can be solved in time polynomial to T, $|S|$, $|A|$, K and $1/\delta$. Furthermore, as the next lemma shows, we can bound the error introduced by discretization.

Lemma 1. *Let π be any policy, and $\hat{\pi}$ be the solution to the new probabilistic goal MDP. We have that*

$$Pr(X_\pi \geq V + T\delta) \leq Pr(X_{\hat{\pi}} \geq V).$$

Proof. Fix t, s, a, by definition of $\hat{g}_t$, we have that for any c

$$\sum_{r \geq c} \hat{g}_t(r|s,a) \leq \sum_{r \geq c} g_t(r|s,a) \leq \sum_{r \geq c-\delta} \hat{g}_t(r|s,a).$$

Since the realization of reward parameters are independent, and there are T stages, this implies that

$$Pr(X_\pi \geq V + T\delta) \leq Pr(\hat{X}_\pi \geq V) \leq Pr(X_\pi \geq V),$$

where $\hat{X}$ denotes the (random) total reward for the new MDP. Since $\hat{\pi}$ maximizes $Pr(\hat{X}_\pi \geq V)$, the lemma follows. □

The theorem follows by setting $\delta = \epsilon/T$. □

6 Chance Constrainted MDPs

Thus far we investigated the case where we want to find a policy that *maximizes* the probability of achieving a predetermined target V. Another reasonable formulation is to treat this probability as a constraint: the decision maker may be interested in maximizing the expected reward while in the mean time ensures that the probability of achieving the target V is *larger than a threshold*. This leads to the Chance Constrained MDP (CC-MDP) formulation.

Problem 3 (Chance Constrained MDP). *Given $V \in \mathbb{R}$ and $\alpha \in (0, 1)$, find a policy $\pi \in \Pi^{h,u}$ that solves*

$$\begin{array}{ll} \textit{Maximize:} & \mathbb{E}(X_\pi) \\ \textit{Subject to:} & Pr(X_\pi \geq V) \geq \alpha. \end{array}$$

Notice that checking the the feasibility of CC-MDP is equivalent to (NP-hard) $D(\pi^{h,u})$, which implies the NP-hardness of CC-MDP. Nevertheless, similar as the probabilistic goal MDP, there exists a pseudo-polynomial algorithm that solves CC-MDP. However, in contrast to the probabilistic goal MDP, the optimal policy may be randomized. For succinctness we focus in this section the integer-reward case. Note that the general reward case can be approximated to arbitrary precision using discretization.

Theorem 5. *Given an MDP such that $\mathcal{R} \in \mathbb{Z}$, $V \in \mathbb{R}$, $\alpha \in [0, 1]$, the Chance Constrained MDP can be solved in a time polynomial in $T, |S|, |A|, K$. Furthermore, there exists an optimal solution π belongs to $\Pi^{t,s,x,u}$.*

Proof. Similar to the probabilistic goal MDP case, we construct a new MDP with augmented state space (s, C). The transition probability is defined in a same way as in the probabilistic goal MDP. However, we define the immediate reward as a pair (r^1, r^2), both of which take non-zero values only at the last stage. In the last stage, if the state is (s, C), the first reward component is C and the second one is $\mathbf{1}_{(C \geq V)}$. Thus, observe that for a given policy π of this new MDP, the expected value of the first and the second component of the accumulated reward equals to the expected reward and probability of meeting the target of the original MDP, respectively. Thus, the chance constrained MDP is equivalent to find a policy of the second MDP that solves the following problem.

$$\begin{array}{ll} \text{Maximize:} & \mathbb{E}(X^1_\pi) \\ \text{Subject to:} & \mathbb{E}(X^2_\pi) \geq \alpha, \end{array} \tag{1}$$

where X^1_π and X^2_π are the first and second component of the cumulative reward, respectively. Observe that Equation (1) is the well-studied constrained MDP formulation, which can be solved in polynomial time by converting it into a linear program [Altman, 1999]. It is also known that there exists a Markovian *randomized* optimal solution. Therefore, the chance constrained MDP can be solved in pseudo-polynomial time, and one optimal solution belongs to $\Pi^{t,s,x,u}$. □

We note that for CC-MDP , $\Pi^{t,s,x}$ is inferior to $\Pi^{t,s,x,u}$, i.e., randomization may be necessary for optimal policies. This property is inherited from constrained MDPs and is due to the fact that we want to satisfy the probabilistic constraint while maximizing the expected reward.

7 Numerical Results

In this section we briefly report some simulation results that illustrate the performance of the proposed approach. We consider the following machine replacement example which is

similar in spirit to [Delage and Mannor, 2010]. There are four possible states for a machine. The larger the state number, the worse the machine is. One can either replace the machine, which incurs a random cost, and the next state is $s0$, or "do nothing," with no cost, and the next state will be one state larger; see Figure 3. Fix $T = 20$. We compute the policy that minimizes the expected cost as in the standard MDP, and the policies that maximizes the probability that the cost is no larger than $5, 6, 7, 8, 9$, respectively. For each policy, 1000 simulations are performed and the relative error is under 1%. We plot the performance of each policy in Figure 4: for a given x, the y-value is the frequency that the cumulative cost of a policy is smaller or equal to x. The simulation result clearly shows that each policy maximizes the chance of reaching its respective target.

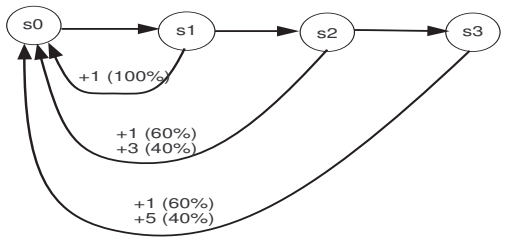

Figure 3: Simulation: Machine replacement dynamics

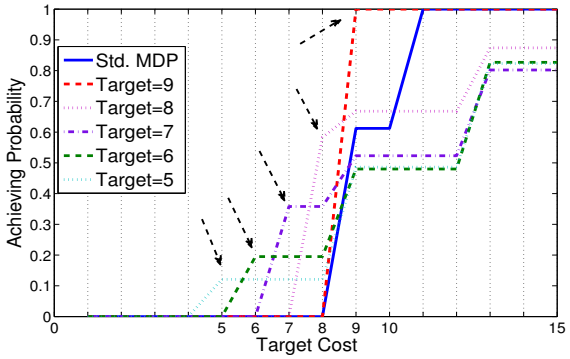

Figure 4: Simulation Result

8 Conclusion and Discussions

In this paper we proposed and investigated the probabilistic goal MDP model where the decision maker maximizes the probability that the cumulative reward of a policy is above a fixed target. We showed that while this formulation is NP-hard, it can be solved in pseudo-polynomial time using state augmentation. We further discussed the chance constrained MDP formulation and showed it to be solvable in pseudo-polynomial time.

The classical objective for MDP, considers only the expected cumulative reward, and hence may not capture the risk preference of the decision maker. The main thrust of this work is to address this shortcoming, with a decision criterion that is both intuitively appealing and also justified from a decision theory perspective, without sacrificing too much of the computational efficiency of the classical approach.

From an algorithmic perspective, the proposed algorithm for solving probabilistic MDPs may appear to be "easy" and not scale well, partly due to the fact that the problem is NP-hard. However, we believe that by embedding the probabilistic MDP into a large scale MDP, it opens doors to use more scalable methods, in the spirit of approximate dynamic programming, that have been developed to handle large-scale MDPs, to approximately solve the probabilistic MDP. This is an interesting and important issue for future research.

References

[Altman, 1999] E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

[Bertsekas, 1995] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 1995.

[Castagnoli and LiCalzi, 1996] E. Castagnoli and M. LiCalzi. Expected utility without utility. *Theory and Decisions*, 41:281–301, 1996.

[Chung and Sobel, 1987] K. Chung and M. Sobel. Discounted MDP's: distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.

[Delage and Mannor, 2010] E. Delage and S. Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.

[Filar *et al.*, 1995] J. A. Filar, D. Krass, and K. W. Ross. Percentile performance criteria for limiting average Markov decision processes. *IEEE Transactions on Automatic Control*, 40(1):2–10, 1995.

[Iyengar, 2005] G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[Lanzillotti, 1958] R. F. Lanzillotti. Pricing objectives in large companies. *American Economic Review*, 48:921–940, 1958.

[Le Tallec, 2007] Y. Le Tallec. *Robust, Risk-Sensitive, and Data-driven Control of Markov Decision Processes*. PhD thesis, MIT, 2007.

[Liu and Koenig, 2005] Y. Liu and S. Koenig. Risk-sensitive planning with one-switch utility functions: Value iteration. In *Proceedings of the 20th AAAI Conference on Artificial Intelligence*, pages 993–999, 2005.

[Mao, 1970] J. Mao. Survey of capital budgeting: Theory and practice. *Journal of Finance*, 25:349–360, 1970.

[Miller and Wagner, 1965] L. B. Miller and H. Wagner. Chance-constrained programming with joint constraints,. *Operations Research*, 13:930–945, 1965.

[Nilim and El Ghaoui, 2005] A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, September 2005.

[Payne *et al.*, 1980] J. W. Payne, D. J. Laughhunn, and R. Crum. Tanslation of gambles and aspiration level effects in risky choice behaviour. *Management Science*, 26:1039–1060, 1980.

[Payne *et al.*, 1981] J. W. Payne, D. J. Laughhunn, and R. Crum. Further tests of aspiration level effects in risky choice behaviour. *Management Science*, 26:953–958, 1981.

[Prékopa, 1970] Prékopa. On probabilistic constrained programming. In *In Proceedings of the Princeton Symposium on Mathematical Programming*, pages 113–138, 1970.

[Puterman, 1994] M. L. Puterman. *Markov Decision Processes*. John Wiley & Sons, New York, 1994.

[Riedel, 2004] F. Riedel. Dynamic coherent risk measure. *Stochastic Processes and Applications*, 112:185–200, 2004.

[Simon, 1959] H. A. Simon. Theories of decision-making in economics and behavioral science. *American Economic Review*, 49(3):253–283, 1959.