Stationary Deterministic Policies for Constrained MDPs with Multiple Rewards, Costs, and Discount Factors

Dmitri Dolgov and Edmund Durfee
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109
{ddolgov, durfee}@umich.edu

Abstract

We consider the problem of policy optimization for a resource-limited agent with multiple time-dependent objectives, represented as an MDP with multiple discount factors in the objective function and constraints. We show that limiting search to stationary deterministic policies, coupled with a novel problem reduction to mixed integer programming, yields an algorithm for finding such policies that is computationally feasible, where no such algorithm has heretofore been identified. In the simpler case where the constrained MDP has a single discount factor, our technique provides a new way for finding an optimal deterministic policy, where previous methods could only find randomized policies. We analyze the properties of our approach and describe implementation results.

1 Introduction

Markov decision processes [Bellman, 1957] provide a simple and elegant framework for constructing optimal policies for agents in stochastic environments. The classical MDP formulations usually maximize a measure of the aggregate reward received by the agent. For instance, in widely-used discounted MDPs, the objective is to maximize the expected value of a sum of exponentially discounted scalar rewards received by the agent. Such MDPs have a number of very nice properties: they are subject to the principle of local optimality, according to which the optimal action for a state is independent of the choice of actions for other states, and the optimal policies for such MDPs are stationary, deterministic, and do not depend on the initial state of the system. These properties translate into very efficient dynamic-programming algorithms for constructing optimal policies for such MDPs (e.g., [Puterman, 1994]), and policies that are easy to implement in standard agent architectures.

However, there are numerous domains where the classical MDP model proves inadequate, because it can be very difficult to fold all the relevant feedback from the environment (i.e., rewards the agent receives and costs it incurs) into a single scalar reward function. In particular, the agent's actions, in addition to producing rewards, might also incur costs that might be measured very differently from the rewards, making it hard or impossible to express both on the same scale. For example, a natural problem for a delivery agent is to maximize aggregate reward for making deliveries, subject to constraints on the total time spent en route. Problems naturally modeled as constrained MDPs also often arise in other domains: for example, in telecommunication applications (e.g., [Lazar, 1983]), where it is desirable to maximize throughput subject to delay constraints.

Another situation where the classical MDP model is not expressive enough is where an agent receives multiple reward streams and incurs multiple costs, each with a different discount factor. For example, the delivery agent could face a rush-hour situation where the rewards for making deliveries decrease as a function of time (same delivery action produces lower reward if it is executed at a later time), while the traffic conditions improve with time (same delivery action can be executed faster at a later time). If the rewards decrease and traffic conditions improve on different time scales, the problem can be naturally modeled with two discount factors, allowing the agent to evaluate the tradeoffs between early and late delivery. Problems with multiple discount factors also frequently arise in other domains: for example, an agent can be involved in several financial ventures with different risk levels and time scales, where a model with multiple discount factors would allow the decision maker to quantitatively weigh the tradeoffs between shorter- and longer-term investments. Feinberg and Shwartz [1999] describe more examples and provide further justification for constrained models with several discount factors.

The price we have to pay for extending the classical model by introducing constraints and several discount factors is that stationary deterministic policies are no longer guaranteed to be optimal [Feinberg and Shwartz, 1994; 1995]. Searching for an optimal policy in the larger class of non-stationary randomized policies can dramatically increase problem complexity; in fact, the complexity of finding optimal policies for this broad class of constrained MDPs with multiple costs, re-

This material is based upon work supported by Honeywell International, and by the DARPA/IPTO COORDINATORs program and the Air Force Research Laboratory under Contract No. FA8750–05–C–0030. The views and conclusions contained in this document are those of the authors, and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

wards, and discount factors is not known, and no solution algorithms exist (aside from some very special cases [Feinberg and Shwartz, 1996]). Furthermore, even if they could be found, these non-stationary randomized policies might not be reliably executable by basic agent architectures. For example, [Paruchuri *et al.*, 2004] described how executing randomized policies in multiagent systems can be problematic.

In this paper, therefore, we focus on finding optimal stationary deterministic policies for MDPs with multiple rewards, costs, and discount factors. This problem has been studied before and has been proven to be NP-complete by Feinberg [2000], who formulated it as a non-linear and non-convex mathematical program. Unfortunately, aside from intractable techniques of general non-convex optimization, these problems have heretofore not been practically solvable.

Our contribution in this paper is to present an approach to solving this problem that reduces it to a mixed-integer linear program – a formulation that, while still NP-complete, has available a wide variety of very efficient solution algorithms and tools that make it practical to often find optimal stationary deterministic policies. As we will show, moreover, our approach can also be fruitfully employed for the subclass of MDPs that have multiple costs, but only a single reward function and discount factor. For these problems, linear programming can, in polynomial time, find optimal stationary randomized policies [Kallenberg, 1983; Heyman and Sobel, 1984], but the problem of finding optimal stationary deterministic policies is NP-complete [Feinberg, 2000], with no implementable solution algorithms existing previously (aside from the general non-linear optimization techniques). We show that our integer-programming-based approach finds optimal stationary deterministic policies, which can then be compared empirically to optimal randomized policies.

In the remainder of this paper, we first (in Section 2) establish a baseline by briefly reviewing techniques for solving unconstrained MDPs. In Section 3, we move on to constrained MDPs, and present our approach to solving constrained MDPs with a single reward, multiple costs, and one discount factor for the rewards and costs. We next expand this to the case with multiple rewards and costs, and several discount factors, in Section 4. Section 5 provides some empirical evaluations and observations, and Section 6 discusses our results and some thoughts about applying the same techniques to other flavors of constrained MDPs.

2 Background: Unconstrained MDPs

An unconstrained, stationary, discrete-time, fully-observable MDP can be defined as a 4-tuple $\langle \mathcal{S}, \mathcal{A}, p, r \rangle$, where $\mathcal{S} = \{i\}$ is a finite set of states the agent can be in; $\mathcal{A} = \{a\}$ is a finite set of actions the agent can execute; $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the stochastic ($\sum_j p_{iaj} = 1$) transition function (p_{iaj} is the probability the agent goes to state j if it executes action a in state i); $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the bounded reward function (the agent gets a reward of r_{ia} for executing action a in state i).

A solution to an MDP is a policy (a procedure for selecting an action in every state) that maximizes some measure of aggregate reward. In this paper we will focus on MDPs with the total expected discounted reward optimization criterion, but our results can be extended to other optimization criteria (as discussed in Section 6). A policy is said to be *Markovian* (or *history-independent*) if the choice of action does not depend on the history of states and actions encountered in the past, but only on the current state and time. If, in addition to that, the policy does not depend on time, it is called *stationary* (by definition, a stationary policy is always Markovian). A *deterministic* policy always prescribes the execution of the same action in a state, while a *randomized* policy chooses actions according to a probability distribution.

A stationary randomized policy π can be described as a mapping of states to probability distributions over actions: $\pi : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$, where π_{ia} defines the probability that the agent will execute action a when it encounters state i. A deterministic policy can be viewed as a degenerate case of a randomized policy for which there is only one action for each state that has a nonzero probability of being executed.

A policy π and the initial conditions $\alpha : \mathcal{S} \mapsto [0, 1]$ that specify the probability distribution over the state space at time 0 (the agent starts in state i with probability α_i) together determine the evolution of the system and the total expected discounted reward the agent will receive:

$$U_\gamma(\pi, \alpha) = \sum_{t=0}^{\infty} \sum_{i,a} \gamma^t \varphi_i(t) \pi_{ia} r_{ia}, \tag{1}$$

where $\varphi_i(t)$ refers to the probability of being in state i at time t, and $\gamma \in [0, 1)$ is the discount factor.

It is well-known (e.g., [Puterman, 1994]) that, for an unconstrained MDP with the total expected discounted reward optimization criterion, there always exists an optimal policy π^* that is stationary, deterministic, and *uniformly-optimal*, where the latter term means that the policy is optimal for all initial probability distributions over the starting state (i.e., $U_\gamma(\pi^*, \alpha) \geq U_\gamma(\pi, \alpha) \; \forall \pi, \alpha$).

There are several standard ways of solving such MDPs (e.g., [Puterman, 1994]); some use dynamic programming (value or policy iteration), others, which are much better suited for constrained problems, reduce MDPs to linear programs (LPs). A discounted MDP can be formulated as the following LP [D'Epenoux, 1963; Kallenberg, 1983] (this maximization LP is the dual to the more-commonly used minimization LP in the value function coordinates):

$$\max \sum_{i,a} r_{ia} x_{ia} \left| \begin{array}{l} \displaystyle\sum_a x_{ja} - \gamma \sum_{i,a} x_{ia} p_{iaj} = \alpha_j, \\ x_{ia} \geq 0. \end{array} \right. \tag{2}$$

The set of optimization variables x_{ia} is often called the *occupation measure* of a policy, where x_{ia} can be interpreted as the total expected discounted number of times action a is executed in state i. Then, $\sum_a x_{ia}$ gives the total expected discounted *flow* through state i, and the constraints in the above LP can be interpreted as the conservation of flow through each of the states. An optimal policy can be computed from a solution to the above LP as:

$$\pi_{ia} = x_{ia} / \sum_a x_{ia}. \tag{3}$$

Although this appears to lead to randomized policies, in the absence of external constraints, and if we use strictly positive initial conditions ($\alpha_i > 0$), a basic feasible solution to this LP always maps to a deterministic policy that is uniformly-optimal [Puterman, 1994; Kallenberg, 1983]. This LP (2) for the unconstrained MDP serves as the basis for solving constrained MDPs that we discuss next.

3 Constrained MDPs

Suppose that the agent, besides getting rewards for executing actions, also incurs costs: $c^k : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R},\ k \in [1, K]$, where c^k_{ia} is the cost of type k incurred for executing action a in state i (e.g., actions might take time and consume energy, in which case we would say that there are two types of costs). Then, a natural problem to pose is to maximize the expected discounted reward subject to some upper bounds on the total expected discounted costs. Let us label the total expected discounted cost of type $k \in [1, K]$ as:

$$C^k_\gamma(\pi, \alpha) = \sum_{t=0}^{\infty} \sum_{i,a} \gamma^t \varphi_i(t) \pi_{ia} c^k_{ia}. \tag{4}$$

Then, we can abstractly write the optimization problem with cost constraints as

$$\max_\pi U_\gamma(\pi, \alpha) \,\big|\, C^k_\gamma(\pi, \alpha) \le \widehat{c}^k, \tag{5}$$

where $\widehat{c}^k$ is the upper bound on the cost of type k. If this problem is feasible, then there always exists an optimal stationary policy, and it can be computed as a solution to the following LP [Kallenberg, 1983; Heyman and Sobel, 1984]:

$$\max \sum_{i,a} r_{ia} x_{ia} \left| \begin{array}{l} \sum_a x_{ja} - \gamma \sum_{i,a} x_{ia} p_{iaj} = \alpha_j, \\ \sum_{i,a} c^k_{ia} x_{ia} \le \widehat{c}^k, \\ x_{ia} \ge 0. \end{array} \right. \tag{6}$$

Therefore, constrained MDPs of this type can be solved in polynomial time, i.e., adding constraints with the same discount factor does not increase the complexity of the MDP. However, due to the addition of constraints, the problem (5), in general, will not have uniformly-optimal policies. Furthermore, the LP (6) will yield randomized policies, which (as argued in Section 1) are often more difficult to implement than deterministic ones.

Thus, it can be desirable to compute optimal solutions to (5) from the class of stationary deterministic policies. This, however, is a much harder problem: Feinberg [2000] studied this problem, showed that it is NP-complete (using a reduction similar to [Filar and Krass, 1994]), and reduced it to a mathematical program by augmenting (6) with the following constraint, ensuring that only one x_{ia} per state is nonzero:

$$|x_{ia} - x_{ia'}| = x_{ia} + x_{ia'} \tag{7}$$

However, the resulting program (6,7) is neither linear nor convex, and thus presents significant computational challenges.

We show how (5) can be reduced to a mixed integer linear program (MILP) that is equivalent to (6,7). This is beneficial because MILPs are well-studied (e.g., [Wolsey, 1998]), and there exist efficient implemented algorithms for solving them. Our reduction uses techniques similar to the ones employed in [Dolgov and Durfee, 2004b], where an MDP with limited non-consumable resources is reduced to a MILP. The following proposition provides the basis for our reduction.

Proposition 1 *Consider an MDP $\langle \mathcal{S}, \mathcal{A}, p, r, \alpha \rangle$, a policy π, its corresponding occupation measure x (given α), a constant $X \ge x_{ia}\ \forall i \in \mathcal{S}, a \in \mathcal{A}$, and a set of binary variables $\Delta_{ia} = \{0, 1\},\ \forall i \in \mathcal{S}, a \in \mathcal{A}$.*

If x and Δ satisfy the following conditions

$$\sum_a \Delta_{ia} \le 1, \quad \forall i \in \mathcal{S}, \tag{8}$$

$$x_{ia}/X \le \Delta_{ia}, \quad \forall i \in \mathcal{S}, a \in \mathcal{A} \tag{9}$$

then, for all states i that, under π and α, have a nonzero probability of being visited ($\sum_a x_{ia} > 0$), π is deterministic, and the following holds:

$$\Delta_{ia} = 1 \Leftrightarrow x_{ia} > 0 \tag{10}$$

Proof: Consider a state i^* that, under policy π and initial distribution α, has a nonzero probability of being visited, i.e., $\sum_a x_{i^*a} > 0$. Then, since the occupation measure is non-negative, there must be at least one action in this state that has a non-zero occupation measure:

$$\exists a^* \in \mathcal{A} \quad \text{s.t.} \quad x_{i^*a^*} > 0.$$

Then, (9) forces $\Delta_{i^*a^*} = 1$, which, due to (8), forces zero values for all other Δ's for state i^*:

$$\Delta_{i^*a} = 0 \quad \forall a \ne a^*.$$

Given (9), this, in turn, means that the occupation measure for all other actions has to be zero:

$$x_{i^*a} = 0 \quad \forall a \ne a^*,$$

which, per (3), translates into the fact that the policy π is deterministic and $\Delta_{ia} = 1 \Leftrightarrow x_{ia} > 0$. ■

Proposition 1 immediately leads to the following MILP whose solution yields optimal stationary deterministic policies for (5):

$$\max \sum_i \sum_a x_{ia} r_{ia} \left| \begin{array}{l} \sum_a x_{ja} - \gamma \sum_{i,a} x_{ia} p_{iaj} = \alpha_j, \\ \sum_{i,a} c^k_{ia} x_{ia} \le \widehat{c}^k, \\ \sum_a \Delta_{ia} \le 1, \\ x_{ia}/X \le \Delta_{ia}, \\ x_{ia} \ge 0, \quad \Delta_{ia} \in \{0, 1\}, \end{array} \right. \tag{11}$$

where X can be computed in polynomial time by, for example, solving the LP (2) with the objective function replaced by $\max \sum_{i,a} x_{ia}$ and setting X to its maximum value.

The above reduction to an MILP is most valuable for domains where it is difficult to implement a randomized stationary policy because of an agent's architectural limitations. It is also of interest for domains where such limitations are not present, as it can be used for evaluating the quality vs. implementation-difficulty tradeoffs between randomized and deterministic policies during the agent-design phase.

4 Constrained MDPs with Multiple Discounts

We now turn our attention to the more general case of MDPs with multiple streams of rewards and costs, each with its own discount factor $\gamma_n,\ n \in [1, N]$. The total expected reward is a weighted sum of the N discounted reward streams:

$$U(\pi, \alpha) = \sum_n \beta_n U_{\gamma_n}(\pi, \alpha) = \sum_n \beta_n \sum_{t=0}^{\infty} \sum_{i,a} \gamma_n^t \varphi_i(t) \pi_{ia} r_{ia}^n$$

where β_n is the weight of the n^{th} reward stream that is defined by the reward function $r^n : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$. Similarly, each of the K total expected costs is a weighted sum of N cost streams:

$$C^k(\pi, \alpha) = \sum_n \beta_{kn} C_{\gamma_n}^k(\pi, \alpha) = \sum_{n,i,a} \beta_{kn} \sum_{t=0}^{\infty} \gamma_n^t \varphi_i(t) \pi_{ia} c_{ia}^{kn}$$

where β_{kn} is the weight of the n^{th} discounted stream of cost of type k, defined by the cost function $c^{kn} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$.

Notice that in this MDP with multiple discount factors, we have N reward functions and KN cost functions (unlike the constrained MDP from the previous section that had 1 and K reward and cost functions, respectively).

Our goal is to maximize the total weighted discounted reward, subject to constraints on weighted discounted costs:

$$\max_{\pi} U(\pi, \alpha) \,\big|\, C^k(\pi, \alpha) \le \widehat{c}^k, \quad \forall k \in [1, K]. \tag{12}$$

Feinberg and Shwartz [1994; 1995] developed a general theory of constrained MDPs with multiple discount factors and demonstrated that, in general, optimal policies are neither deterministic nor stationary. However, except for some special cases [Feinberg and Shwartz, 1996], there are no implementable algorithms for finding optimal policies for such problems. Because of this, and given the complexity of implementing non-stationary randomized policies (even if we could find them), it is worthwhile to consider the problem of constructing optimal stationary deterministic policies for such MDPs. Feinberg [2000] showed that finding optimal policies that belong to the class of stationary (randomized or deterministic) policies is an NP-complete task. He also formulated the problem of finding optimal stationary policies as the following mathematical program (again, based on (eq. 6)):

$$\max \sum_n \beta_n \sum_{i,a} r_{ia}^n x_{ia}^n \left| \begin{array}{l} \displaystyle\sum_a x_{ja}^n - \gamma_n \sum_{i,a} x_{ia}^n p_{iaj} = \alpha_j, \\ \displaystyle\sum_n \beta_{kn} \sum_{i,a} c_{ia}^{kn} x_{ia}^n \le \widehat{c}^k, \\ \displaystyle x_{ia}^n / \sum_a x_{ia}^n = x_{ia}^{n+1} / \sum_a x_{ia}^{n+1}, \\ x_{ia}^n \ge 0, \end{array} \right. \tag{13}$$

This program has an occupation measure x^n for each discount factor $\gamma_n,\ n \in [1, N]$ and expresses the total reward and total costs as weighted linear functions of these occupation measures. The first set of constraints contains the conservation of flow constraints for each of the N occupation measures, and the third set of constraints ensures that all occupation measures map to the same policy (recall (3)).

As in the previous section, we can limit the search to deterministic policies by imposing the following additional constraint on the occupation measures in (13) [Feinberg, 2000]:

$$\Big| \sum_n \left(x_{ia}^n - x_{ia'}^n \right) \Big| = \sum_n \left(x_{ia}^n + x_{ia'}^n \right) \tag{14}$$

Because of the synchronization of the different occupation measures and the constraint that forces deterministic policies, this program (13,14) is non-linear and non-convex, and is thus very difficult to solve.

For finding optimal stationary deterministic policies, we present a reduction of the program (12) to a linear integer program that is equivalent to (13,14). Just like in the previous section, this reduction to an MILP allows us to exploit a wide array of efficient solution techniques and tools. Our reduction is based on the following proposition.

Proposition 2 *Consider an MDP $\langle \mathcal{S}, \mathcal{A}, p, r, \alpha \rangle$ with several discount factors $\gamma_n,\ n \in [1, N]$, a set of policies $\pi^n,\ n \in [1, N]$ with their corresponding occupation measures x^n (policy π^n and discount factor γ_n define x^n), a constant $X \ge x_{ia}^n\ \forall n \in [1, N], i \in \mathcal{S}, a \in \mathcal{A}$, and a set of binary variables $\Delta_{ia} = \{0, 1\}$.*

If x^n and Δ satisfy the following conditions

$$\sum_a \Delta_{ia} \le 1, \quad \forall i \in \mathcal{S}, \tag{15}$$

$$x_{ia}^n / X \le \Delta_{ia}, \quad \forall n \in [1, N], i \in \mathcal{S}, a \in \mathcal{A}, \tag{16}$$

then, the sets of reachable states $\mathcal{I}^n = \{i : \sum_a x_{ia}^n > 0\}$ defined by all occupation measures are the same, i.e., $\mathcal{I}^n = \mathcal{I}^{n'},\ \forall n, n' \in [1, N]$. Furthermore, all π^n are deterministic on $\mathcal{I}^n$, and $\pi_{ia}^n = \pi_{ia}^{n'}\ \forall n, n' \in [1, N]$.

Proof: Consider an initial state i^* (i.e., $\alpha_{i^*} > 0$). Following the argument of Proposition 1, the policy for that state is deterministic:

$$\exists a^* : \ \ x_{i^*a^*}^n > 0, \Delta_{i^*a^*} = 1;\ x_{i^*a}^n = 0, \Delta_{i^*a} = 0\ \forall a \ne a^*$$

This implies that all N occupation measures x^n must prescribe the execution of the same deterministic action a^* for state i^*, because all x_{ia}^n are tied to the same Δ_{ia} via (16).

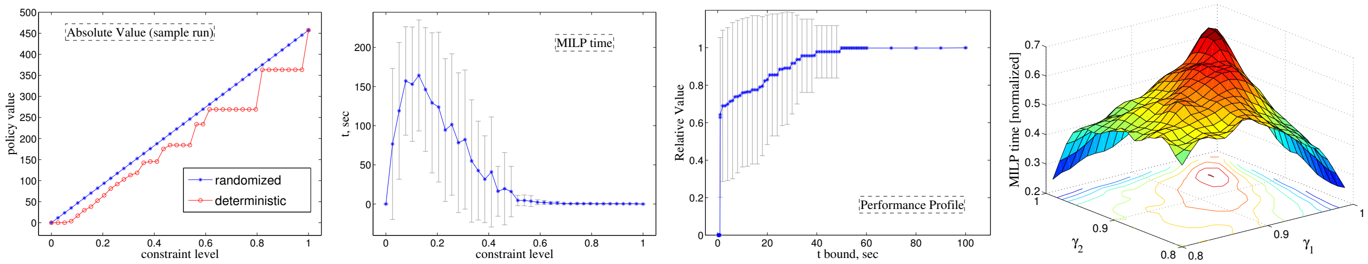
Therefore, all occupation measures x^n correspond to the same deterministic policy on the initial states $\mathcal{I}_0 = \{i : \alpha_i > 0\}$. We can then expand this statement by induction to all reachable states. Indeed, clearly the set of states $\mathcal{I}_1$ that are reachable from $\mathcal{I}_0$ in one step will be the same for all x^n. Then, by the same argument as above, all x^n map to the same deterministic policy on $\mathcal{I}_1$, and so forth. ■

It immediately follows from Proposition 2 that the problem of finding optimal stationary deterministic policies for an MDP with weighted discounted rewards and constraints (12) can be formulated as the following MILP:

$$\max \sum_n \beta_n \sum_{i,a} r_{ia}^n x_{ia}^n \left| \begin{array}{l} \displaystyle\sum_a x_{ja}^n - \gamma_n \sum_{i,a} x_{ia}^n p_{iaj} = \alpha_j, \\ \displaystyle\sum_n \beta_{kn} \sum_{i,a} c_{ia}^{kn} x_{ia}^n \le \widehat{c}^k, \\ x_{ia}^n / X \le \Delta_{ia}, \\ \displaystyle\sum_a \Delta_{ia} \le 1, \\ x_{ia}^n \ge 0, \Delta_{ia} \in \{0, 1\}, \end{array} \right. \tag{17}$$

where $X \ge \max x_{ia}^n$ is a constant, as in Proposition 2.

Although this MILP produces policies that are only optimal in the class of stationary deterministic ones, at present



(a) (b) (c) (d)

Figure 1: Value of deterministic and randomized policies (a); solution time (b) and profile (c); MDP with two discounts (d).

time there are (to the best of our knowledge) no practical algorithms for finding optimal solutions from any larger policy class for constrained MDPs with multiple discount factors.

5 Experimental Observations

We have implemented the MILP algorithm for finding optimal stationary deterministic policies for constrained MDPs and empirically evaluated it on a class of test problems. In the following discussion, we focus on the constrained MDPs from Section 3, because these problems are better studied, and the existing algorithms for finding optimal randomized policies can serve as benchmarks, whereas there are no alternative algorithms for finding policies that are optimal for general constrained MDPs with multiple discount factors.

In our empirical analysis, we tried to answer the following questions: 1) how well do deterministic policies perform, compared to optimal randomized ones, and 2) what is the average-case complexity of the resulting MILPs. The answers to these questions are obviously domain-dependent, so the following discussion should not be viewed as a comprehensive characterization of the behavior of our algorithms on constrained MDPs. However, we believe that our experiments provide some interesting observations about such problems.

We experimented with a large set of randomly-generated problems and with a more meaningful manually-constructed domain, which we randomly perturbed in several ways. The big picture resulting from the experiments on the randomly-generated domains was very similar to the one from the manually-constructed example, providing a certain measure of comfort about the stability and validity of our observations. We report the results for the manually-constructed domain.

For our test domain, we used a simplistic model of an autonomous delivery agent, as mentioned in the introduction (based on the multiagent example from [Dolgov and Durfee, 2004b]). In the domain, an agent is operating in a grid world with delivery sites placed randomly throughout the grid. The agent moves around the grid (incurring small negative rewards for every move) and receives positive rewards for making deliveries. The agent's movement is nondeterministic, and the agent has some probability of getting stuck in randomly-placed dangerous locations. The agent also incurs a scalar cost (e.g., time) per move, and the objective is to maximize the total expected discounted reward subject to an upper bound on the total expected discounted cost.

The results of our experiments are summarized in Figure 1. Figure 1a shows the values of randomized and deterministic policies as functions of the constraint level ($[0, 1]$), where 0 means that only policies that incur zero cost are feasible (strictest possible constraint), whereas 1 means that the upper bound on cost equals the cost of the optimal unconstrained policy (agent is not constrained at all). The first observation, as illustrated in Figure 1a, is that the value of stationary deterministic policies for constrained problems is reasonably close to optimal. We can also observe that the value of deterministic policies changes in a very discrete manner (i.e., it jumps up at certain constraint levels), whereas the value of randomized policies changes continuously. This is, of course, only natural, given that the space of randomized policies is continuous, and randomized policies can gradually increase the probability of taking "better" actions as cost constraints are relaxed. On the other hand, the space of deterministic policies is discrete, and their quality jumps when the constraints are relaxed to permit the agent to switch to a better action. While the number and the size of these jumps in the value function depends on the dynamics of the MDP, the high-level picture was the same in all of our experiments.

Figure 1b shows the running time of the MILP solver as a function of the constraint level (here and in Figure 1c the plots contain values averaged over 100 runs, with the error bars showing the standard deviation). The data indicates that our MILPs (11) have an easy-hard-easy complexity profile, although without a sharp phase transition from hard to easy, i.e., the problems very quickly become hard, and then gradually get easier as cost constraints are relaxed.

This complexity profile gives rise to the question regarding the source of the difficulty for solving MILPs in the "hard" region: is it difficult to find good feasible solutions, or is it time-consuming to prove their optimality? Figure 1c suggests that the latter is the case, which can be considered as the more fortunate outcome, since algorithms with such performance profiles can be successfully used in an anytime manner. The figure contains a plot of the quality of the best solution found as a function of the time bound imposed on the MILP solver[1] for problems in the hardest constraint region (constraint level value of 0.13). As the graph shows, very good policies are usually produced rather quickly.

Let us conclude with a somewhat intriguing observation about the MILP solution time for constrained MDPs with multiple discount factors (Section 4). We generated and solved a large number of random MDPs with two discount

[1] CPLEX 8.1 on a P4 performed the role of the MILP solver.

factors and plotted (after cubic smoothing) the average solution time (shown in Figure 1d). An interesting observation about this plot is that the problem instances where the two discount factors are equal (or close) appear to be the hardest (notice the contours in the γ_1-γ_2 plane). This is counterintuitive, because such MDPs are equivalent to standard MDPs with one discount factor. A possible explanation might be that when discount factors are far apart, one of the reward functions dominates the other and the problem becomes simpler, while when the discount factors are close, the tradeoffs become more complicated (with the equivalence to a standard MDP hidden in the MILP translation). However, this is speculation and the phenomenon deserves a more careful analysis.

6 Discussion and Conclusions

We have presented algorithms for finding optimal deterministic policies for two classes of constrained MDPs, and in both cases we were maximizing a measure of the total expected discounted reward subject to constraints on the total expected discounted costs. However, our technique of finding optimal stationary deterministic policies via mixed integer programming also applies to other classes of MDPs.

In particular, the same methodology applies to MDPs with average per-time rewards and constraints (e.g., [Puterman, 1994]). Similarly to the constrained total-reward discounted MDP model described in Section 3, the MDP with average rewards and constraints can also be formulated as an LP (similar to (6)) that yields optimal stationary randomized policies. The problem of finding optimal stationary deterministic policies for such MDPs is also known to be NP-complete [Filar and Krass, 1994]. Our MILP reduction of Section 3 carries through with almost no changes and can thus be used to find optimal stationary deterministic policies for such MDPs.

The techniques of Section 4 also apply more generally. For instance, consider an MDP with general utility functions in the optimization criteria and constraints on the *probability* that the total cost exceeds some upper bound. This class of MDPs was discussed in [Dolgov and Durfee, 2004a], and the problem of finding approximately-optimal stationary randomized policies was reduced to a non-convex quadratic program in the space of the higher-order moments of the state visits. The non-convex quadratic constraints resulted from the requirement that the moments of different orders had to correspond to the same policy, and were almost identical to the quadratic constraints in (14) that synchronized the occupation measures for different discount factors. In fact, the two are so similar that our MILP reduction from Section 4 can be used to approximate optimal stationary deterministic policies for MDPs in [Dolgov and Durfee, 2004a].

To summarize, we have presented a general integer programming method for finding optimal stationary deterministic policies in constrained MDPs. We have demonstrated the method on two classes of MDPs: (i) a constrained discounted MDP and (ii) a constrained MDP with multiple discount factors. For (i), our methodology is of most value for domains where randomized policies (which work better and are easier to compute) are undesirable or difficult to implement because of an agent's architectural limitations. However, even in the absence of such limitations, the approach is useful in situations where it is desirable to compare the quality of randomized and deterministic policies, such as when an agent is being designed for a particular task and it is necessary to weigh the cost of implementing a more complex policy-execution mechanism against the gain in expected performance. For problem (ii), to the best of our knowledge, no feasible algorithms have been reported for finding optimal solutions in any interesting policy class, and thus our MILP approach for finding optimal stationary deterministic policies provides the first practical approach to dealing with constrained MDPs with multiple discount factors.

References

[Bellman, 1957] Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[D'Epenoux, 1963] D'Epenoux. A probabilistic production and inventory problem. *Management Science*, 10:98–108, 1963.

[Dolgov and Durfee, 2004a] Dmitri A. Dolgov and Edmund H. Durfee. Approximate probabilistic constraints and risk-sensitive optimization criteria in Markov decision processes. In *Proc. of the 8th Int. Symposiums on AI and Math (AI&M 7-2004)*, 2004.

[Dolgov and Durfee, 2004b] Dmitri A. Dolgov and Edmund H. Durfee. Optimal resource allocation and policy formulation in loosely-coupled Markov decision processes. In *Int. Conf. on Automated Planning and Scheduling*, pages 315–324, 2004.

[Feinberg and Shwartz, 1994] E. A. Feinberg and A. Shwartz. Markov decision processes with weighted discounted criteria. *Math. of OR*, 19:152–168, 1994.

[Feinberg and Shwartz, 1995] E. A. Feinberg and A. Shwartz. Constrained Markov decision processes with weighted discounted criteria. *Math. of OR*, 20:302–320, 1995.

[Feinberg and Shwartz, 1996] E. Feinberg and A. Shwartz. Constrained discounted dynamic programming. *Math. of OR*, 21:922–945, 1996.

[Feinberg and Shwartz, 1999] E. Feinberg and A. Shwartz. Constrained dynamic programming with two discount factors: Applications and an algorithm. *IEEE Transactions on Automatic Control*, pages 628–630, 1999.

[Feinberg, 2000] Eugene A. Feinberg. Constrained discounted Markov decision processes and hamiltonian cycles. *Math. of Operations Research*, 25(1):130–140, 2000.

[Filar and Krass, 1994] J. A. Filar and D. Krass. Hamiltonian cycles and Markov chains. *Math. of OR*, 19:223–237, 1994.

[Heyman and Sobel, 1984] D. P. Heyman and M. J. Sobel. *Volume II: Stochastic Models in Operations Research*. McGraw-Hill, New York, 1984.

[Kallenberg, 1983] L.C.M. Kallenberg. *Linear Programming and Finite Markovian Control Problems*. Math. Centrum, Amsterdam, 1983.

[Lazar, 1983] A. Lazar. Optimal flow control of a class of queuing networks in equilibrium. *IEEE Transactions on Automatic Control*, 28(11):1001–1007, 1983.

[Paruchuri *et al.*, 2004] Praveen Paruchuri, Milind Tambe, Fernando Ordonez, and Sarit Kraus. Towards a formalization of teamwork with resource constraints. In *Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, pages 596–603, 2004.

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

[Wolsey, 1998] L.A. Wolsey. *Integer Programming*. John Wiley & Sons, 1998.