CONSTRAINED DYNAMIC PROGRAMMING WITH TWO DISCOUNT FACTORS: APPLICATIONS AND AN ALGORITHM

EUGENE A. FEINBERG†

W.A. Harriman School for Management and Policy
SUNY at Stony Brook
Stony Brook, NY 11794-3775.
Phone: 516 632-7189. Fax: 516 632 9412
Email: EFEINBER@FAC.HAR.SUNYSB.EDU

ADAM SHWARTZ*

Electrical Engineering
Technion—Israel Institute of Technology
Haifa 32000, Israel
Phone: (972) 4 8294743. Fax: (972) 4 8323041
Email: adam@ee.technion.ac.il

June 1997

Abstract

We consider a discrete time Markov Decision Process, where the objectives are linear combinations of standard discounted rewards, each with a different discount factor. We describe several applications that motivate the recent interest in these criteria. For the special case where a standard discounted cost is to be minimized, subject to a constraint on another standard discounted cost but with a different discount factor, we provide an implementable algorithm for computing an optimal policy.

---

† Please address all correspondence to this author.

* Research performed in part while the author was on Sabbatical leave, at Bell Laboratories–Lucent Technologies, and later at the Department of Management Science and Information Systems, Graduate School of Business, Rutgers University.

I. Introduction.

In recent years there is growing interest in criteria which combine long and short term considerations. These could be average and discounted costs, or several discounted costs with different discount factors. The areas of interest include Markov Decision Processes Feinberg [2], Krass [12], Filar, Krass and Sinha [9], Fernandez-Gaucherand, Ghosh and Marcus [7], Feinberg and Shwartz [4, 5], controlled diffusions Ghosh and Marcus [10], and stochastic games Filar and Vrieze [8], Altman Feinberg and Shwartz [1]. A comprehensive theory for discrete time Markov Decision Processes with multiple discount factors, including computation aspects in the finite case, was developed in [4]. Feinberg and Shwartz [5] investigate the existence and structure of optimal policies for the mixed-discount problem under cost constraints. Some computational procedures were suggested, to obtain ϵ-optimal policies. Krass [12] and Feinberg and Shwartz [5] describe several applications.

In this note we concentrate on two issues: the first is motivation for the constrained mixed discount problem. We provide additional examples which illustrate the applicability of this model. The second is the derivation of a computational procedure for obtaining optimal policies for maximizing a standard discounted cost, subject to a constraint on another standard discounted cost, but with a different discount factor.

Example 1. Consider the problem of managing a computer facility with many users. The objective is to provide acceptable service, while spending as little as possible. The state x encodes information about the number of users and about available computer resources that influence the performance of the system, such as RAM memory, computation power/speed, disk space, etc.

For the sake of illustration, consider the simple case in which the only available decision (control) is to add a GBytes of disk space. Suppose that the cost per GByte of disk storage at time $n = 0$ is c_d. The price of disk space has been decreasing at an exponential rate over the last ten years: denote the rate of decrease per unit time by β_1. Then the cost of purchasing a GBytes of disk space at time n is $\beta_1^n \cdot a \cdot c_d$.

Let $c_p(x)$ be a combined measure of performance, when the state (users, resources) is x. To model the fact that demands from computer performance as a whole increase at an exponential rate, we let the $\beta_p < 1$ be the ratio between required performances at consecutive decision epochs. Then maintaining "adequate performance" in the long run may be modeled by the requirement that, for some appropriate constant B_p,

$$\mathbb{E}_x^\pi \sum_{n=0}^{\infty} \beta_p^n c_p(x_n) \geq B_p .$$

Combining the different criteria we arrive at the following optimization problem:

$$\begin{aligned} \text{Minimize} \qquad & \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \beta_1^n c_1(x_n, a_n) \\ \text{subject to} \qquad & \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \beta_p^n c_p(x_n) \geq B_p \end{aligned}$$

where $c_1(x,a) = a \cdot c_d$. This is a mixed-discount Markov decision model with a constraint.

In Example 1, the different discount factors are a consequence of two mechanisms. The first is the explicit dependence of the cost on time (exponential decrease of the price of disk space), while the second concerns the use of discounting as a measure of future value. Next we describe a case where different discounts arise due to different time scales.

Example 2. Consider a discounted problem with discount factor β. Suppose that a component of the cost is a total discounted cost with immediate cost function c_1. However, suppose that this cost is accumulated only until some unobservable event occurs: below we give examples where this may be the case. We assume that the unobservable event occurs at a random, geometrically distributed time. We model the geometric random variable as a product of $\{0,1\}$ valued, i.i.d. random variables ζ_t with mean β_1, where $\zeta_t = 0$ if and only if the unobservable event occurred at time t. The total cost then takes the form

$$\mathbb{E}_x^\pi \left(\sum_{n=0}^{\infty} \beta^n c_1(x_n, a_n) \prod_{t=0}^{n-1} \zeta_t \right) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \beta^n c_1(x_n, a_n) \beta_1^n \ , \tag{1}$$

where the second expression is a result of the fact that, since $\{\zeta_t\}$ are not observable, then under any policy the random variables $\{x_n, a_n\}$ are independent of $\{\zeta_t\}$.

We can view $1 - \beta_1$ as our estimate of the rate of occurrence of the unobservable event. Let us denote the other immediate costs—those costs that do not depend on the unobservable event—by $c(x,a)$. Then if our intention is to minimize costs, subject to a constraint that the functional (1) stays below a given bound B, we obtain the optimization problem

$$\begin{aligned} \text{Minimize} \qquad & \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \beta^n c(x_n, a_n) \\ \text{subject to} \qquad & \mathbb{E}_x^\pi \sum_{n=0}^{\infty} (\beta \cdot \beta_1)^n c_1(x_n, a_n) \leq B \ , \end{aligned}$$

which is a mixed-discount Markov decision model with a constraint.

As a specific example, consider a research group trying to develop a new product (for example, a new medicine). The usual discounted cost serves to aggregate the immediate cost and expected future costs and revenues. The discount factor in this case is related to the interest rate (or, more precisely, to the opportunity cost). In addition, there is a possibility that a competitor develops the same product first. If the competitor submits a patent application first, we may not become aware of this until much later, when our patent application process is completed (rejected, in this case). Similarly, for example in medical research, the time between the development of a drug until it is approved for marketing is extremely long (due to the required clinical tests and FDA approval). So, again, we may not become aware of the success of a competitor until long after the completion of the competitor's research. In either case, the financial consequences can be expected to be quite significant: we may be unable to go into production, or may have to transfer a large part of the profit to the competitor.

So, we assume that we do not know if and when a competitor completes the development: we can only model his success as a geometric random variable. We can view $1 - \beta_1$ as our estimate of the rate of success of a competitor. Since the time from completion of development to actual marketing can be very long, we may take the horizon to be infinite. We assume that at time 0 the product development has not been completed by the competitors (more generally, it suffices that we know the probability that development was completed by time 0).

Introduce a new state, say x_0, so that when we succeed in our research (and submit a patent application), we set the state $x = x_0$. This state is visited at most once. Suppose that, if we submit our patent application before the competition does, then our profit (from the time of submission) is c_p. Then the choice $c_1(x, a) = c_p$ if and only if $x = x_0$, and $c_1(x, a) = 0$ otherwise, implies that our total cost criterion (1) measures the total expected discounted profit.

The mathematical program then solves the constrained problem of minimizing our costs, subject to expected profit being above a certain margin. Other choices of c_1 yield different measures, such as expected lead time in submitting for a patent, etc.

Note that if we know about completion of development by a competitor, then the equality in (1) is not valid: the policy π will then depend on the ζ_t, leading to statistical dependence of the state process.

It is also possible to model this application by enlarging the state space, and obtaining an MDP with partial information: however, both structural results and algorithms for the latter representation are even less developed than for our representation.

Feinberg and Shwartz [4, 5] develop the theory for Markov Decision Processes with mixed discounts. For the unconstrained optimization problem, they derive explicit algorithms to obtain the optimal policy. The complexity of these algorithms is similar to that of standard discounted problems. However, no algorithms are available for the problem with constraints. Feinberg [3] shows that the constrained problem, restricted to stationary policies, is NP-hard. The complexity of the constrained problem in the class of all policies is not known.

We now describe an iterative algorithm for the computation of optimal policies in the special case of a standard discounted cost, with a constraint on a another standard discounted cost functional with a different discount factor. In the interest of brevity, we introduce only the non-standard part of the notation. For details the reader is referred to [4, 5].

II. The Model.

Let $\mathbb{N}_0 = \{0, 1, ...\}$, $\mathbb{N} = \{1, 2, ...\}$, $\|\cdot\|$ be an Euclidean norm in $\mathbb{R}^2$, and $\langle u, v\rangle$ the scalar product of $u, v \in \mathbb{R}^2$. Consider a discrete-time controlled Markov chain with

(i) finite state space $\mathbf{X}$,

(ii) finite action space $\mathbf{A}$, with actions $\mathbf{A}(x)$ available at x,

(iii) transition probabilities $\{p(y \mid x, a)\}$,

(iv) immediate reward functions $c_i(x, a)$ and discounts $0 < \beta_i < 1$, $i = 1, 2$.

A strategy π is a function that assigns to each prehistory $h_n = x_0 a_0 x_1 ... x_n$ a probability distribution $\pi(\cdot \mid h_n)$ on $\mathbf{A}$ so that $\pi(\mathbf{A}(x_n) \mid h_n) = 1$. A strategy π is called (N, ∞)-*stationary* if for any $x \in \mathbf{X}$ there exists a stationary strategy ϕ such that

$$\pi_n(\phi(x_n) \mid h_n) = 1 \qquad \text{for} \qquad n = N+1, N+2, \ldots \tag{2}$$

for any prehistory h_n. Each strategy π and initial state x induce a probability measure $\mathbb{P}_x^\pi$ on the canonical space $H_\infty = (\mathbf{X} \times \mathbf{A})^\infty$, with corresponding expectation operator $\mathbb{E}_x^\pi$.

We are interested in solving the following constrained optimization problem: given a bound B,

$$\text{maximize} \quad V_1(x, \pi) = \mathbb{E}_x^\pi \sum_{n=0}^\infty \beta_1^n c_1(x_n, a_n), \tag{3}$$

$$\text{subject to} \quad V_2(x, \pi) = \mathbb{E}_x^\pi \sum_{n=0}^\infty \beta_2^n c_2(x_n, a_n) \geq B, \tag{4}$$

where $\beta_1 \neq \beta_2$. Denote the vector $(V_1(x, \pi), V_2(x, \pi))$ by $\overline{V}(x, \pi)$ and define the “performance space” $U(x) = \left\{\overline{V}(x, \pi) : \pi \in \Pi\right\}$. To develop an algorithm for (3)–(4), we use the following relation between Pareto optimality and maximization. Recall that $u = (u_1, u_2)$ in $\mathbb{R}^2$ is called Pareto optimal in a set U if $v \in U$ and $v_i \geq u_i$, $i = 1, 2$, imply $v = u$. Let $\mathbb{R}_0^2 = \{u \in \mathbb{R}^2 : u_1 \geq 0,\ u_2 \geq 0\}$.

Lemma 1. *Let $U \in \mathbb{R}^2$ be closed, convex, and bounded above. Consider the problem:*

$$\text{maximize } \{\langle b, v\rangle :\ v \in U\}. \tag{5}$$

(i) If $u \in U$ is Pareto optimal, then u is a solution of (5) for some $b \in \mathbb{R}_0^2$, with $b_1 + b_2 > 0$.

(ii) If u is a solution of (5) for some $b \in \mathbb{R}_0^2$ and $b_1 > 0,\ b_2 > 0$, then u is Pareto optimal.

Lemma 1 is well-known. In particular, Lemma 1(i) follows from Lemma 3.2(i,ii,iv,v) in [6], and Lemma 1(ii) follows from [11, page 75].

III. Constrained optimization: infinite horizon.

Recall that the initial condition is fixed at some x: in constrained problems, the optimal policy depends on the initial state. Assume the problem is feasible, that is

$$\max_{\pi} V_2(x, \pi) \geq B.$$

Checking this condition is a standard discounted problem and, unless this inequality holds, problem (3)–(4) has no solution. The problem of optimizing each of the V_k is a standard discounted problem. Let $A_k(x)$ be the set of conserving actions for the corresponding problem; these are the actions that achieve the maximum in the optimality equation. We then have from the general theory of discounted Markov processes that the maximal value is obtained for any strategy π such that $\pi(A_k(x_n)|h_n) = 1$ for any $h_n = x_0 a_0 x_1 ... x_n$, $n = 0, 1, \ldots$. As a result we can treat the two extreme cases rather easily. For $i, j = 1, 2,\ i \neq j$ define a strategy σ to be (i,j)-lexicographic-optimal at x if σ maximizes $V_i(x, \pi)$ at x, and σ provides the maximal value for $V_j(x, \pi)$ among all strategies whose action sets are $A_i(\cdot)$. The computation of an (i,j)-lexicographic-optimal policy σ proceeds as follows. First compute $A_i(x)$ for all x. Now solve the problem of maximizing $V_j(x, \pi)$ over policies π, where the action sets are restricted to $A_i(\cdot)$. This yields sets $A_{ij}(x)$ of conserving actions, for all x. Any policy σ whose actions are chosen from $A_{ij}(x)$ is an (i,j)-lexicographic-optimal policy.

We denote the corresponding value vector as $V^{ij}(x) \triangleq \{V_1^{ij}(x), V_2^{ij}(x)\}$.

Theorem 2. *If*

$$\max_{\pi} V_2(x, \pi) = B \tag{6}$$

then the $(2,1)$*-lexicographic-optimal strategy provides a solution to the constrained optimization problem.*

Proof. By the hypothesis, only strategies with actions in $A_2(x)$ are feasible. By definition and by properties of standard discounted problems, the $(2,1)$-lexicographic-optimal strategy provides the maximal value among all feasible strategies. ■

Theorem 3. *If the* $(1,2)$*-lexicographic-optimal strategy* σ *is such that*

$$V_2(x,\sigma) \geq B \tag{7}$$

then σ *is a solution to the constrained optimization problem.*

Proof. By definition. ■

We now consider the remaining cases. For $0 < b < 1$ define

$$W_b(x,\pi) = bV_1(x,\pi) + (1-b)V_2(x,\pi). \tag{8}$$

We shall rely on the following result concerning weighted discounted criteria. Let $i,j = 1,2$ be such that $\beta_i > \beta_j$.

Theorem 4. *There exists a function* $N = N(b)$ *such that, if* σ *is optimal for (8) then* σ *agrees with some* (i,j)*-lexicographic-optimal strategy at* x*, for all time* $t \geq N(b)$*.* $N(b)$ *is finite in* $(0,1)$*, decreasing in* b *when* $\beta_1 > \beta_2$ *and increasing when* $\beta_1 < \beta_2$*.*

Proof. The first claim is established in [4, Corollary 3.4 and Theorem 3.8]. Assume without loss of generality that $c_i(x,a) \geq 0$, so that $V_i(x,\pi) \geq 0$. Suppose $\beta_1 > \beta_2$ and recall the definitions [4, Equation (3.1) and preceding equation] of N and ϵ_1. Since $bV_1(x)$ is decreasing in b, $\epsilon_1(b)$ is increasing. Since $(1-b)V_2(x) - (1-b)V_2^-(x)$ is decreasing in b (where $V_2^-(x)$ is defined in [4]), it now follows that $N(b)$ is decreasing. The proof for the case $\beta_1 < \beta_2$ is identical. ■

By [5, Corollary 3.4], the sets $U(x)$ are convex for each x. Consequently, if both (6) and (7) do not hold, then necessarily the solution to the constrained optimization problem is obtained at a point V so that $V_2(x,\pi) = B$, and moreover, the slope of the normal to any tangent hyperplane to $U(x)$ at V is bounded away from 0 and ∞. But this implies that such a solution may be found by solving (8) with b bounded away from 0 and from 1. We then have the following structural result.

Theorem 5. *For some finite $N = 0, 1, \ldots$, there is an (N, ∞)-stationary strategy which solves the constrained optimization problem.*

Remark. Theorem 5 is a consequence of [5, Theorem 6.8]: however, the (relatively) simple structure of our problem allows for a direct and more intuitive proof.

Proof. If either (6) or (7) holds, Theorems 2 and 3 provide the existence of stationary optimal strategies. Now assume that both (6) and (7) do not hold. By assumption (using the convexity of $U(x)$), the normal to any tangent hyperplane to $U(x)$ at $V_2(x) = B$ has positive slope which is bounded away from 0 and ∞. Therefore such a point is Pareto optimal, and can be found by solving (8), with the appropriate (but yet unknown) value of b, and moreover $0 < b_0 \leq b \leq b_1 < 1$ for some b_0 and b_1. Theorem 4 now implies that there is a number N so that the solutions to (8) for any b with $b_0 \leq b \leq b_1$ can be chosen as stationary (i, j)-lexicographic-optimal from time N onward. Fix such a strategy σ. Then our problem reduces to computing the optimal solution under constraints of a finite (duration N) problem with terminal rewards $\beta_k^N V_k(y, \sigma)$ in state y. This problem can be solved via Algorithm 4.3 of [5]. ∎

Theorem 5 is an existence result. The proof of this theorem defines a number b such that $N = N(b)$, where b is described in Theorem 4. In order to obtain a practical algorithm we need to be able to compute $N = N(b)$. A natural way to this is to compute b first. Once $N = N(b)$ is computed, the computation of an optimal (N, ∞)-stationary policy is straightforward. This policy coincides with an optimal (i, j)-lexicographic-optimal policy after step N. Then we should solve an N-horizon problem with one constraint; see [5] where this computation is explained for any number of constraints.

The general theory of mixed-discount MDP's under constraints [5] suggests that the number of state-time pairs where randomization is necessary is limited by the number of constrains. In particular, in the present case where there is a single constraint, there is an optimal policy with the following property. There is at most one state and at most one point in time where a random action is necessary, and furthermore, at most two actions need to be chosen with positive probability at this time-state pair. Otherwise, the policy is nonrandomized. This also follows from the argument of the proof of Theorem 5. The lexicographic policy is not randomized, and the finite-horizon problem can be solved via a linear program defined in [5]. The presence of a single constraint implies a single randomization.

Recall the definition of V^{ij}, preceding Theorem 2.

Lemma 6. *If (7) does not hold, then a bound b_0 is obtained by considering $(b_0, 1 - b_0)$ the normal* *to the line* (V_1^{12}, B)–V^{21} *in the* V_1, V_2 *plane.*

Proof. Follows from convexity of $U(x)$, since by the assumptions the boundary of $U(x)$ must cross the line $V_2 = B$ at a point V with $V_1 < V_1^{12}$. ■

This bound suggests the following algorithm.

Algorithm 1.

1. Compute $b_0 > 0$ so that $(b_0, 1 - b_0)$ is normal to the line (V_1^{12}, B)–V^{21}.
2. Compute $N = N(b_0)$ using [4], Algorithm 3.7.
3. Find any (i, j)-lexicographic-optimal stationary strategy σ and compute $V_k(y, \sigma)$, $k = 1, 2$.
4. Use a finite horizon algorithm [5], Algorithm 4.3 to compute an optimal strategy up to time N.

Corollary 7. *If* $\beta_1 > \beta_2$ *then an optimal solution can be found using Algorithm 1.*

Proof. Since $N(b)$ is decreasing and b_0 is a lower bound on b, the value $N(b_0)$ is an upper bound on $N(b)$. The Corollary now follows from Theorem 5. ■

We do not have such a simple algorithm for computing the other bound b_1. Here is a search algorithm to obtain b_1. This algorithm terminates after a finite number of steps.

Algorithm 2.

1. Compute the value of b corresponding to the normal $(b, 1 - b)$ to the line V^{12}–V^{21}.
2. Compute an optimal policy σ for (8) using this value of b. If $V_2(x, \sigma) \leq B$ set $b_1 = b$ and stop. Otherwise
3. Replace b with $(b + 1)/2$ and repeat from 2.

That this algorithm terminates in a finite number of steps follows from the fact that, since (7) does not hold, the slope of the normal to any tangent hyperplane to $U(x)$ at $V_2(x) = B$ is bounded away from 0 and ∞.

To conclude, we have established the following finite algorithm for computing optimal strategies for the constrained problem.

Algorithm 3.

1. Check for the cases in Theorems 2 and 3.
2. If they do not hold, compute bounds for b using Lemma 6 and Algorithm 2.
3. Compute N using [4, Algorithm 3.7].
4. Find any (i, j)-lexicographic-optimal stationary strategy σ and compute $V_k(y, \sigma)$, $k = 1, 2$.
5. Use a finite horizon algorithm to compute an optimal strategy.

Remark. Consider the following two-dimensional constrained optimization problem.

$$\begin{aligned} \text{maximize} \quad & \sum_{k=1}^{2} a_k V_k(x,\pi) \\ \text{subject to} \quad & \sum_{k=1}^{2} c_k V_k(x,\pi) \geq B. \end{aligned}$$

Then Theorems 2–4 continue to hold, except that $N(b)$ is not monotone. The argument of Theorem 5 continues to hold as well. However, the computation of a bound on N (or, equivalently, on b) is more complicated.

Remark. Some of the ideas herein extend to the higher dimensional problem

$$\begin{aligned} \text{maximize} \quad & V_1(x,\pi) \\ \text{subject to} \quad & V_k(x,\pi) \geq B_k, \qquad k = 2, \ldots, K, \end{aligned}$$

where each V_k is a standard discounted criterion with discount factor β_k. For example, if

$$\max_{\pi} V_k(x,\pi) = B_k$$

for some $k \geq 2$, then we can restrict to actions which are conserving for V_k, and consider the problem with $K-2$ constraints. However, due to the higher dimentionality, the geometry is less transparent and there are more specific cases to deal with.

Acknowledgments Research of the first author was partially supported by NSF Grant DMI-9500746. Research of the second author was supported in part by the fund for promotion of research at the Technion, and by the Israel Science Foundation, administered by the Israel Academy of Sciences and Humanities. It was performed in part while the author was on Sabbatical Leave: the hospitality and support of the Mathematical Center, Bell Laboratories, Murray Hill and of the Department of Management Science and Information Systems, Graduate School of Business, Rutgers University is gratefully acknowledged.

References

[1] E. Altman, E. Feinberg and A. Shwartz (1996), "Weighted Discounted Stochastic Games with Perfect Information," CC Pub. #146, Electrical Engineering, Technion, Israel.

[2] E. Feinberg (1982), "Controlled Markov processes with arbitrary numerical criteria," *SIAM Theory Prob. Appl.* **27** pp. 486–503.

[3] E. Feinberg (1997), "Constrained discounted Markov decision processes and Hamiltonian cycles," preprint, SUNY at Stony Brook.

[4] E. Feinberg and A. Shwartz (1994), "Markov decision models with weighted discounted criteria," *Math. of Operations Research* **19** pp. 152–168.

[5] E. Feinberg and A. Shwartz (1995), "Constrained Markov decision models with weighted discounted rewards," *Math. of Operations Research* **20** pp. 302–320.

[6] E. Feinberg and A. Shwartz (1996), "Constrained discounted dynamic programming," *Math. of Operations Research* **21** pp. 922–945.

[7] E. Fernandez-Gaucherand, M.K. Ghosh and S.I. Marcus (1994), "Controlled Markov processes on the infinite planning horizon: weighted and overtaking cost criteria," *ZOR-Methods and Models in Operations Research* **39** pp. 131–155.

[8] J. Filar and O.J. Vrieze (1992), "Weighted reward criteria in competitive Markov decision processes," *ZOR—Methods and Models of Operations Research*, **36** pp. 343–358.

[9] J. Filar, D. Krass and S.S. Sinha (1992), "A weighted Markov decision process," *Operations Research* **40** pp. 1180-1187.

[10] M.K. Ghosh and S.I. Marcus (1991), "Infinite horizon controlled diffusion problems with some nonstandard criteria," *J. Mathematical Systems, Estimation and Control* **1** pp. 45–69.

[11] R.L. Keeney and H. Raiffa (1976), *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, New York.

[12] D. Krass, *Contributions to the Theory and Applications of Markov Decision Processes*, Ph. D. Thesis, The Johns Hopkins University, Baltimore, 1989.