Markov Decision Models with Weighted Discounted Criteria*

EUGENE A. FEINBERG

SUNY at Stony Brook

ADAM SHWARTZ

Technion—Israel Institute of Technology.

June 1991

Revised: *August 1992*

Abstract

We consider a discrete time Markov Decision Process with infinite horizon. The criterion to be maximized is the sum of a number of standard discounted rewards, each with a different discount factor. Situations in which such criteria arise include modeling investments, production, modeling projects of different durations and systems with multiple criteria, and some axiomatic formulations of multi-attribute preference theory. We show that for this criterion for some positive ϵ there need not exist an ϵ-optimal (randomized) stationary strategy, even when the state and action sets are finite. However, ϵ-optimal Markov (non-randomized) strategies and optimal Markov strategies exist under weak conditions. We exhibit ϵ-optimal Markov strategies which are stationary from some time onward. When both state and action spaces are finite, there exists an optimal Markov strategy with this property. We provide an explicit algorithm for the computation of such strategies and give a description of the set of optimal strategies.

AMS 1980 subject classification. Primary: 90C40.

IAOR 1973 subject classification. Main: Programming, Markov Decision.

OR/MS Index 1978 subject classification. Primary: 119 Dynamic Programming/Markov.

Key words. Dynamic programming, Markov, sum of discounted rewards with different discount factors.

---
* **To appear, MOR 1993**

The use of a discounted reward criterion in Markov decision models is consistent with the notion that what happens far in the future is unimportant. Discounted future cost can be given the economic interpretation of an opportunity cost. It can also arise through the subjective notion that immediate rewards are better than delayed rewards. Existing theory deals with the following three situations: the case of a fixed discount rate (Shapley 1953, Blackwell 1962, 1965, Denardo 1967, Bertsekas 1987, Dynkin and Yushkevich 1979, Heyman and Sobel 1984, Ross 1984, Whittle 1982), the case when the discount rate depends on current states and actions (Schäl 1975), and the case when the discount rate is a function of the history of the process (Hinderer 1970). Naturally, results in the last situation are very limited, and no effective computational procedures are available.

Discount factors depend on perceived investment opportunities. When there are several different investment opportunities then it is natural to consider a weighted discounted criterion. This criterion is the sum of several expected total discounted rewards with different discount factors. Such criteria arise in models of investments with different risk classes. Two cash flow streams with different risks would have different required rates of return and hence different discount rates. Nevertheless, the value of the portfolio consisting of both cash flow streams would be the sum of their individual values, a principle appropriately named value additivity; Brealey and Myers (1988). Thus, the value of the portfolio is the sum of the discounted values of each cash flow stream in the portfolio.

Weighted discounted criteria also arise naturally in production. Standard models of production costs involve a component which decreases exponentially, due to the effect of learning: the "learning curve" is a model of costs which decrease as a result of improved technology and know-how. The rate involved in learning is not related to the discount rate arising from the opportunity cost. Thus the natural model could involve one discount factor arising from the opportunity cost, and one additional (different) discount factor for each technological area which affects production. Since the ultimate objective involves total costs, a weighted discounted criterion arises.

Weighted discounted criteria also arise in the axiomatic formulation of multi-attribute preference theory. Sobel (1991) describes general preference axioms leading to discounted and weighted discounted criteria.

Other examples when weighted discounted criteria arise in economics include investments in different assets within a company, investments in different pension funds, management of state budgets, and investment opportunities in a country with a prevalent underground economy.

Another possible interpretation of total discounted rewards is the sum of total rewards in models with finite but random horizon; Ross (1984). If the model reflects the managing of long-term and short-term projects in parallel, weighted discounted criteria arise.

Weighted discounted criteria arise also in models with multiple criteria when each criterion is an expected total discounted reward and different criteria may have different discount factors. Linear-weighted averages (Keeney and Raiffa 1976) for such models are weighted discounted criteria. A related approach to this multiple criteria problem, via constrained optimization, is discussed in Feinberg and Shwartz (1992).

Various weighted criteria were considered by Feinberg (1982b) as an illustration of possible applications of methods developed in that paper. The case of a weighted sum of two criteria was considered by Filar and Vrieze (1989) in the context of stochastic games. They consider the finite model and obtain existence of ϵ-optimal Markov strategies for the sum of two criteria; one being the a total discounted cost, and the other either the discounted or the average reward per unit time. When the first criterion is a total discounted cost and the second one is average reward per unit time, the weighted Markov decision problem was considered by Krass, Filar, and Sinha (1990) for discrete time models and by Ghosh and Marcus (1991) for continuous time models. Models with average reward per unit time may be described through discounted models with discount factors close to 1; see Blackwell (1962), Veinott (1966), Denardo (1971). Thus, our model covers the management of projects with long and short durations. By considering different discount factors we can optimize not only interactions of two projects, long and short, but more generally model interactions of several projects with different durations.

We present a theory for the weighted discounted criteria, which is defined more precisely in (1.1)–(1.2) below. In Section I we describe the model and the problems under investigation, and show that even the simplest weighted problem may not possess the structural properties of the standard discounted problem. In particular, even for finite state and action spaces there may not exist a stationary (non-randomized) optimal strategy; in fact, there may not exist an ϵ-optimal randomized stationary strategy! Moreover, randomized stationary strategies perform strictly better then (nonrandomized) stationary strategies, and the best stationary (and randomized stationary) strategies may depend on the initial state. In these examples, the optimal strategy turns out to be stationary only *after some initial time*. In Sections II–III we show that this is generic.

In Section II we prove the existence of ϵ-optimal and optimal Markov strategies. We also establish the existence of ϵ-optimal strategies which are Markov and are stationary from some time N onward: we call strategies of this form (N, ∞)-stationary. In Section III we consider models with finite state and action sets. For this model we show that there exist *optimal* strategies of this form and provide an explicit algorithm to compute an optimal (N, ∞)-stationary strategy.

This algorithm is of the same level of complexity as the computation of optimal strategies for the standard discounted reward problem. We also give a description of the set of optimal strategies for models with finite state and action sets.

I. Definitions and examples.

Consider a discrete-time controlled Markov chain with

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

(ii) measurable action space $\mathbf{A}$ endowed with a σ-field $\mathcal{A}$ containing all one-point subsets of $\mathbf{A}$,

(iii) sets of actions $\mathbf{A}(x)$ available at $x \in \mathbf{X}$. These sets are assumed to be elements of $\mathcal{A}$,

(iv) transition probabilities $\{p(y \mid x, a)\}$. For each $x, y \in \mathbf{X}$ the function $p(y \mid x, a)$ is nonnegative and measurable in a, and $\sum_{y \in \mathbf{X}} p(y \mid x, a) = 1$ for each $x \in \mathbf{X}$ and $a \in \mathbf{A}(x)$.

Let $H_n = \mathbf{X} \times (\mathbf{A} \times \mathbf{X})^n$ be the space of histories up to the time $n = 0, 1, \ldots, \infty$. Let $H = \bigcup_{0 \le n < \infty} H_n$ be the space of all finite histories. The spaces H_n and H are endowed with σ-fields generated by $2^{\mathbf{X}}$ and $\mathcal{A}$. A strategy π is a function that assigns to each history $h_n = x_0 a_0 x_1 \ldots x_{n-1} a_{n-1} x_n \in H_n$, $n = 0, 1, \ldots$, a probability measure $\pi(\cdot \mid h_n)$ on $(\mathbf{A}, \mathcal{A})$ satisfying the following conditions:

(a) $\pi(\mathbf{A}(x_n) \mid h_n) = 1$,

(b) for any $B \in \mathcal{A}$ the function $\pi(B \mid \cdot)$ is measurable on H.

A *Markov* strategy ϕ is a sequence of mappings $\phi_n : \mathbf{X} \to \mathbf{A}$, $n = 0, 1, \ldots$, such that $\phi_n(x) \in \mathbf{A}(x)$ for any $x \in \mathbf{X}$. We say that a Markov strategy ϕ is (N, ∞)-stationary, where $N = 0, 1, \ldots$, if $\phi_n(x) = \phi_N(x)$ for any $n = N+1, N+2, \ldots$ and for any $x \in \mathbf{X}$. A $(0, \infty)$-stationary strategy is called *stationary*. A stationary strategy is determined by a function $\phi : \mathbf{X} \to \mathbf{A}$ such that $\phi(x) \in \mathbf{A}(x)$, $x \in \mathbf{X}$. We will also consider *randomized stationary* and *randomized Markov* strategies. A randomized stationary strategy ϕ is defined by conditional distributions $\phi(\cdot \mid x)$, $x \in \mathbf{X}$, over $(\mathbf{A}, \mathcal{A})$ such that $\phi(\mathbf{A}(x) \mid x) = 1$ for any $x \in \mathbf{X}$. A randomized Markov strategy ϕ is defined by conditional distributions $\phi_n(\cdot \mid x)$, $n = 0, 1, \ldots$, $x \in \mathbf{X}$, over $(\mathbf{A}, \mathcal{A})$ such that $\phi_n(\mathbf{A}(x) \mid x) = 1$ for all n and x.

Using standard notation and construction, each strategy π and initial state x induce a probability measure $\mathbb{P}_x^\pi$ on H_∞. We denote the corresponding expectation operator by $\mathbb{E}_x^\pi$. In contrast with traditional models, we also have

(v) a collection of one-step rewards $\{r_k(x, a),\ k = 1, 2, \ldots, K\}$ which are assumed bounded above and measurable in a, and

(vi) a collection of discount factors $\{\beta_k,\ k = 1, 2, \ldots, K\}$, where $0 < \beta_k < 1$ for any $k = 1, 2, \ldots K$.

The discounted reward associated with the one-step reward r_k and discount factor β_k when the initial state is x and strategy π is used is given by

$$V_k(x;\pi) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} (\beta_k)^t r_k(x_t, a_t) \ . \tag{1.1}$$

The weighted discounted reward when the initial state is x and strategy π is used is now defined as

$$V(x;\pi) = \sum_{k=1}^{K} V_k(x;\pi) \ . \tag{1.2}$$

The value of this problem is given by

$$V(x) = \sup_\pi V(x;\pi). \tag{1.3}$$

Let ϵ be a nonnegative constant. A strategy π^* is called ϵ-optimal if, for all x,

$$V(x;\pi^*) \geq V(x) - \epsilon \ . \tag{1.4}$$

A 0-optimal strategy is called optimal.

We remark that an important practical motivation of this paper is to analyze a situation when a controller pursues a goal to satisfy short— and long-term interests simultaneously. In this situation, it is natural to consider a linear combination of total discounted and average per unit time rewards; cf. Krass, Filar, and Sinha (1990). However, from a mathematical viewpoint, this criterion is awkward; even for models satisfying the strongest possible ergodic conditions and with finite state and action sets, optimal strategies may not exist; cf. example 3.1 in Krass, Filar, and Sinha (1990). Criterion (1.2), with $K = 2$, $\beta_1 > \beta_2$, and β_1 close to 1, models essentially the same situation. Therefore, in view of the results of section III, it is more natural to consider criteria of the form $\lambda_1(1-\beta_1)V_1 + \lambda_2(1-\beta_2)V_2$, where λ_1 and λ_2 are positive numbers. In our case, since r_k are arbitrary, we can omit λ_k by re-defining r_k. This as well as the omission of the normalizing term $1-\beta_k$ do not change the results (except for the values of the bounds and N, cf. Section III), and are adopted in the interest of simplifying the notation.

In section II we establish the existence of ϵ-optimal Markov (Theorem 2.1), ϵ-optimal (N,∞)-stationary (Theorem 2.4), and optimal Markov (Theorems 2.2, 2.6) strategies for this criterion. In section III we establish the existence of optimal (N,∞)-stationary strategies for models with finite state and action sets and describe an effective algorithm for the computation of these strategies.

However, some interesting features of the weighted average criterion are displayed in the following examples.

Example 1.1. For all $\epsilon > 0$ small enough, there exists no ϵ-optimal randomized stationary strategy for a model with finite state and action sets. Moreover, the best strategy among the randomized stationary ones is indeed randomized.

Consider the model with $\mathbf{X} = \{x, y\}$ and $\mathbf{A} = \{a, b\}$. Let

$$p(x \mid z, a) = 1 = p(y \mid z, b), \quad z = x, y \ .$$

We will take the simplest case where $K = 2$, $r_1 = r_2 = r$, and

$$r(x, a) = 1, \quad r(y, a) = r(x, b) = 0, \quad r(y, b) = 2.$$

Now fix $1 > \beta > 0$ and consider the standard discounted Markov decision process with criterion V_β as defined in (1.1) with $\beta_k = \beta$. For this maximization problem there exists an optimal stationary (non-randomized) strategy. It is clear that action b is optimal at y. Thus the two candidates for optimal stationary strategy are

g': stay where you are, i.e. use a at x and b at y, or

g'': go to y and stay there, i.e. use only action b.

A simple calculation gives

$$\begin{aligned} V_\beta(x; g') &= \sum_{t=0}^{\infty} \beta^t = \frac{1}{1-\beta}, \\ V_\beta(x; g'') &= 0 + \sum_{t=1}^{\infty} 2\beta^t = \frac{2\beta}{1-\beta} \ . \end{aligned} \tag{1.5}$$

Thus g' is optimal if $\beta \leq \frac{1}{2}$, while g'' is optimal if $\beta \geq \frac{1}{2}$.

We shall now compute the strategy which is best among all stationary randomized strategies for the weighted-discounted criterion V as defined in (1.1)–(1.3) with some $\beta_1 < \frac{1}{2}$ and $\beta_2 > \frac{1}{2}$. A randomized stationary strategy π is defined through $\pi(a \mid x) = \alpha$, $\pi(a \mid y) = \delta$. The discounted cost $V_\beta(\pi; z)$, $z = x, y$, is the unique solution of the system of linear equations

$$\begin{cases} V_\beta(x; \pi) = \alpha(1 + \beta V_\beta(x; \pi)) + (1-\alpha)\beta V_\beta(y; \pi) \\ V_\beta(y; \pi) = \delta\beta V_\beta(x; \pi) + (1-\delta)(2 + \beta V_\beta(y; \pi)) \end{cases}$$

Solving this system, we have

$$\begin{pmatrix} V_\beta(x;\pi) \\ V_\beta(y;\pi) \end{pmatrix} = \begin{pmatrix} 1-\alpha\beta & -(1-\alpha)\beta \\ -\delta\beta & 1-(1-\delta)\beta \end{pmatrix}^{-1} \begin{pmatrix} \alpha \\ 2(1-\delta) \end{pmatrix}$$

so that

$$\begin{aligned} V_\beta(x;\pi) &= \frac{(1-(1-\delta)\beta)\alpha + 2(1-\delta)(1-\alpha)\beta}{(1-\alpha\beta)(1-(1-\delta)\beta) - (1-\alpha)\beta^2\delta} \\ &= \frac{(\alpha+2\beta-3\alpha\beta)+(3\alpha-2)\beta\delta}{(1-\beta)(1-\alpha\beta+\beta\delta)} \ . \end{aligned} \tag{1.6}$$

Denote $V_\beta(x;\alpha,\delta) = V_\beta(x;\pi)$. It is intuitively clear that $V_\beta(x;\alpha,0) \geq V_\beta(x;\alpha,\delta)$ for any δ in $[0,1]$, since b is optimal at y, and this for all β and α. This can also be seen more formally, as follows. From (1.6) we obtain after some algebra

$$\frac{\partial V_\beta(x;\alpha,\delta)}{\partial\delta} = \frac{\beta(\alpha-1)(2\beta(1-\alpha)+2-\alpha\beta)}{(1-\beta)(1-\alpha\beta+\beta\delta)^2} \leq 0$$

for any $\alpha,\delta \in [0,1]$ and any $\beta \in (0,1)$. Therefore,

$$V_\beta(x;\alpha,0) = \frac{(1-\beta)\alpha+2(1-\alpha)\beta}{(1-\beta)(1-\alpha\beta)} \geq V_\beta(x;\alpha,\delta)$$

for any $\alpha,\delta \in [0,1]$ and any $\beta \in (0,1)$.

Now fix $\beta_1 \neq \beta_2$ and define V through (1.2). Let $V(x;\alpha,\delta) = V_{\beta_1}(x;\alpha,\delta) + V_{\beta_2}(x;\alpha,\delta)$. Then by the previous argument, $V(x;\alpha,0) \geq V(x;\alpha,\delta)$ for any $\alpha,\delta \in [0,1]$. For $\beta_1 = \frac{1}{5}$ and $\beta_2 = \frac{3}{5}$,

$$V(x;\alpha,0) = \frac{5\alpha^2 - 120\alpha + 175}{6\alpha^2 - 40\alpha + 50}$$

for all $\alpha \in [0,1]$. Taking the derivative with respect to α and equating to zero, we get

$$\alpha^* = \frac{20-5\sqrt{3}}{13}$$

and it is easy to verify that α^* maximizes $V(x;\alpha,0)$ for $\alpha \in [0,1]$. The reward associated with the strategy π^* defined through $\alpha = \alpha^*$ and $\delta = 0$ is

$$V(x;\alpha^*,0) = \sup_{0\leq\alpha,\delta\leq 1} V(x;\alpha,\delta) \approx 3.767949 \ .$$

Therefore, given the initial state x and $\beta_1 = \frac{1}{5}$, $\beta_2 = \frac{3}{5}$, the randomized stationary strategy π^* is best among all randomized stationary strategies.

Define the strategy f by

$$f_n = \begin{cases} g', & \text{if } n = 0; \\ g'', & \text{if } n \geq 1. \end{cases}$$

Direct calculation yields

$$V(x; f) = \left(1 + \frac{2\beta_1^2}{1-\beta_1}\right) + \left(1 + \frac{2\beta_2^2}{1-\beta_2}\right) = 3.9,$$

$$V(x; g') = \frac{1}{1-\beta_1} + \frac{1}{1-\beta_2} = 3.75,$$

$$V(x; g'') = 3.5.$$

We conclude that, for $\epsilon < 3.9 - 3.76795$ there does not exist an ϵ-optimal randomized stationary strategy, and that the best randomized stationary strategy is strictly better than the best stationary (non-randomized) strategy.

Remark 1.2. Example 1.1 can be modified so that the decision process is ergodic (i.e. the process is an ergodic Markov chain under all stationary strategies). Since the reward is continuous in the transition probabilities, the conclusions will continue to hold under such a small change. Thus the source of the non-stationarity is indeed the structure of the criterion.

If the state and action spaces are finite, then so is the number of stationary strategies. Therefore, for any given initial state there is a best stationary strategy. Example 1.3 shows that, in contrast to standard discounted (or average) problems, the best stationary strategy may depend on the initial state.

Example 1.3. Consider the model from Example 1.1, but with an additional state s, $\mathbf{A}(s) = \{a\}$ and $p(x \mid s, a) = 1$. We let $r_1(s, a) = r_2(s, a) = 0$ and, as in Example 1.1, we set $\beta_1 = \frac{1}{5}$, $\beta_2 = \frac{3}{5}$. Since the set of actions at s is a singleton, we retain the same notation for stationary strategies as in Example 1.1. As shown in Example 1.1, the stationary strategy g' is the only stationary strategy which is best for both initial states x and y. However,

$$V(s, g') = \frac{\beta_1}{1-\beta_1} + \frac{\beta_2}{1-\beta_2} = \frac{7}{4},$$

$$V(s, g'') = \frac{2\beta_1^2}{1-\beta_1} + \frac{2\beta_2^2}{1-\beta_2} = \frac{19}{10}.$$

Therefore, $V(s,g'') > V(s,g')$ and we conclude that the best stationary strategy depends on the initial state. Direct computation along the lines of Example 1.1 shows that the best randomized stationary strategy also depends on the initial state.

Note that the optimization of this model starting at state s is equivalent to the optimization of the model starting at x, but with the reward functions changed from r_1 and r_2 to $\beta_1 r_1$ and $\beta_2 r_2$.

II. The structure of optimal strategies.

We establish the existence of optimal Markov strategies by embedding our model into a standard stationary discounted Markov decision process. Consider the Markov decision process with state space $\mathbf{X} \times \mathbb{N}$, where $\mathbb{N} = \{0, 1, \ldots\}$. Denote the generic state variable by $\hat{x} = (x,t)$. The action space remains unchanged, and the set of action available at state $\hat{x} = (x,t)$ is $\mathbf{A}(x)$. The new transition probabilities are defined through

$$\hat{p}((y,t') \mid (x,t), a) = \begin{cases} p(y \mid x, a), & \text{if } t' = t+1, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to check that conditions (i)–(iv) of §I hold for the new model. It is also clear that for any strategy π of the original model there corresponds a strategy $\hat{\pi}$ in the extended model, which is uniquely specified for (extended) histories such that $\hat{x}_0 = (x,0)$. Now assume without loss of generality that the discount factors are ordered so that $\beta_1 > \beta_2 > \ldots > \beta_K$. Under assumption (v) we can write the reward (1.1)–(1.2) as

$$V(x;\pi) = \mathbb{E}_x^\pi \sum_{t=0}^\infty (\beta_1)^t \left(\sum_{k=1}^K \left(\frac{\beta_k}{\beta_1}\right)^t r_k(x_t, a_t) \right) .$$

Then letting

$$r((x,t),a) = \sum_{k=1}^K \left(\frac{\beta_k}{\beta_1}\right)^t r_k(x,a) \tag{2.1}$$

we have

$$\hat{V}((x,0);\hat{\pi}) = V(x;\pi) = \mathbb{E}_x^\pi \sum_{t=0}^\infty (\beta_1)^t r(\hat{x}_t, a_t) \tag{2.2}$$

and the resulting model is stationary. Since β_1 is the largest discount factor, assumption (v) of §I implies that r is bounded above and measurable. But then we have a standard countable state discounted Markov decision process with rewards that are bounded above.

Theorem 2.1. *For any $\epsilon > 0$ there exists an ϵ-optimal Markov strategy for the weighted discounted problem (1.1)–(1.3).*

Proof. Given $\epsilon > 0$, a stationary ϵ-optimal strategy exists for a discounted Markov decision model, if the one-step reward function is bounded above; see Dynkin and Yushkevich (1979). Since the reward functions r_k are bounded above, the one-step reward function r in the extended model (2.1) is bounded above on $\mathbf{X} \times \mathbb{N} \times \mathbf{A}$. For a given $\epsilon > 0$, let $\hat{\phi}$ be a stationary ϵ-optimal strategy for the extended model. By our construction, this strategy corresponds to a (non-stationary) Markov strategy ϕ for the original model through

$$\phi_n(x) = \hat{\phi}(x, n) \ . \tag{2.3}$$

ϵ-optimality follows from the embedding and (2.2). ■

Theorem 2.2. *There exists an optimal Markov strategy for the weighted discounted problem if $\mathbf{A}(x)$ are compact subsets of a Borel space, $r_k(x, \cdot)$ are upper semi-continuous and $p(y \mid x, \cdot)$ are continuous for each x, y and k.*

Proof. Under the hypotheses, the results on standard dynamic programming (Schäl 1981, Theorem 7.5) imply that there exists an optimal stationary strategy $\hat{\phi}$ for the extended model. The Markov strategy ϕ, defined through (2.3) is optimal for the original model. ■

Corollary 2.3. *If $\mathbf{A}(x)$ is finite for each $x \in \mathbf{X}$, then there exists an optimal Markov strategy for the weighted discounted problem.*

Theorem 2.4. *If the functions r_k are bounded for all $k = 1, \ldots, K$, except possibly one, then for any $\epsilon > 0$ there exist a finite N and (N, ∞)-stationary ϵ-optimal strategy for the weighted discounted problem.*

Proof. Recall that in our model, all functions r_k are bounded above. There exist ϵ-optimal stationary strategies for discounted problems with rewards bounded above; see Dynkin and Yushkevich (1979). Therefore, we shall consider the case $K > 1$. Suppose that the functions $r_1, \ldots, r_{m-1}, r_{m+1}, \ldots, r_K$ are bounded, with $1 \leq m \leq K$.

We fix some $\epsilon > 0$. Let ϕ be a stationary strategy such that ϕ is $(\epsilon/4)$-optimal for the criterion V_m. Let σ be a Markov $(\epsilon/4)$-optimal strategy for the weighted problem.

Let $|r_k(\cdot, \cdot)| \leq R < \infty$ for $k = 1, \ldots, m-1, m+1, \ldots, K$. Choose $N \in \mathbb{N}$ such that

$$\frac{R\beta_k^N}{1 - \beta_k} \leq \frac{\epsilon}{4(K-1)}, \quad k = 1, \ldots, m-1, m+1, \ldots, K.$$

Define the (N,∞)-stationary strategy γ :

$$\gamma(x) = \begin{cases} \sigma_n(x), & \text{if } n < N; \\ \phi(x), & \text{if } n \geq N, \end{cases}$$

and a Markov strategy σ^N :

$$\sigma_n^N(\cdot) = \sigma_{n+N}(\cdot), \qquad n = 0, 1, \ldots .$$

Then for any $x \in \mathbf{X}$

$$\begin{aligned} V(x;\sigma) - V(x;\gamma) = \mathbb{E}_x^\sigma \sum_{t=N}^{\infty} \sum_{\substack{k=1 \\ k\neq m}}^{K} (\beta_k)^t r_k(x_t, a_t) \\ - \mathbb{E}_x^\phi \sum_{t=N}^{\infty} \sum_{\substack{k=1 \\ k\neq m}}^{K} (\beta_k)^t r_k(x_t, a_t) + (\beta_k)^N \mathbb{E}_x^\phi \Big(V_m(x_N;\sigma^N) - V_m(x_N;\phi) \Big) \\ \leq \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{3\epsilon}{4}. \end{aligned}$$

Therefore, for any $x \in \mathbf{X}$

$$V(x;\gamma) \geq V(x,\sigma) - \frac{3\epsilon}{4} \geq V(x) - \epsilon.$$

■

Remark 2.5. Theorems 2.1 and 2.2 hold for more general models. In fact, consider the conditions under which the standard discounted problem (i)–(iv) with reward (1.1) possesses an optimal stationary strategy. If this holds for each $k = 1, 2, \ldots, K$, then (under all currently available conditions) the conclusion of Theorem 2.2 holds. See, e.g. Schäl (1975), Whittle (1982) and references therein. Theorem 2.1 holds for $K = \infty$ provided that $\sup_k \beta_k < 1$ and that the sum in (2.1) is bounded for $t = 0$. In fact, this approach is applicable even when the sum in (1.2) is replaced by an integral with respect to some measure, over a continuum of values of β.

Theorem 2.6. *If there exist an optimal strategy for the weighted discounted problem then there exists a Markov optimal strategy.*

Proof. The following result holds for a nonstationary Markov decision model with total expected rewards, when a value function is finite (Feinberg 1982a, Theorem 3). Given an initial distribution, for any strategy there exists a Markov strategy with greater or equal total expected rewards. If the state space is finite or countable, one may consider an initial distribution such that for any $x \in \mathbf{X}$ the probability that x is an initial state is positive. Thus, if π is an optimal strategy then there exists a Markov strategy which is not worse than π, and hence is optimal as well. Therefore, if for a nonstationary Markov decision model there exist an optimal strategy then there exists an optimal Markov strategy.

Consider now the weighted discounted problem. The existence of an optimal strategy for the original model implies the existence of an optimal strategy for the extended model. Therefore, for the extended model there exists an optimal Markov strategy $\hat{\phi}$. The Markov strategy ϕ defined by $\phi_n(x) = \hat{\phi}_n(x, n)$ is optimal for the weighted discounted model. ■

Remark 2.7. The results of this section may be extended easily to uncountable state models, since the optimization of the original model with state space $\mathbf{X}$ is equivalent to the optimization of the extended model with state space $\mathbf{X} \times \mathbb{N}$.

Remark 2.8. Filar and Vrieze (1989) studied a weighted discounted problem for zero-sum two-person games with finite state and action sets. They proved the existence of the value, the existence of optimal randomized Markov strategies, and the existence of ϵ-optimal randomized (N, ∞)-stationary strategies for N large enough. The main idea of the proofs of these results in Filar and Vrieze (1989) is the reduction of the original model to an auxiliary model that is more complicated than the extended model with the state space $\mathbf{X} \times \mathbb{N}$ defined above. For example, even in the case of finite state and action spaces, the auxiliary model defined in Filar and Vrieze (1989) has an uncountable state space. The extended model with state space $\mathbf{X} \times \mathbb{N}$ defined at the beginning of Section II may also be used as an auxiliary model for Markov games with an arbitrary number of players. If $L < \infty$ is the number of players, the only difference is that the functions r_k and r in (2.1) become vector-valued and $a = (a^1, a^2, \ldots, a^L)$, where a^i is an action chosen by player i. Then the extended model is a standard discounted Markov game with discount factor β_1. If the extended model has an equilibrium randomized stationary or Markov strategy, then there exists an equilibrium randomized Markov strategy for the weighted discounted problem. For example, equilibrium randomized stationary strategies exist for Markov games with countable state sets, compact action sets, and continuous reward and transition functions (Whitt 1980, Theorem 6.2, and Federgruen 1983, Theorem 6.3.4). Therefore, there exist equilibrium randomized Markov strategies for such models with weighted discounted criteria.

III. The structure of optimal strategies in finite models.

In this section we consider a finite model (finite state and action spaces). As in §II we assume without loss of generality that $\beta_1 > \beta_2 \ldots > \beta_K$. For a finite model, there always exists a stationary optimal strategy for the standard discounted problem, and it is optimal *if and only if* it solves the optimality equation (see e.g. Bertsekas 1987, Dynkin and Yushkevich 1979, Heyman and Sobel 1984, Ross 1984). By Corollary 2.3 there exists an optimal Markov strategy for the weighted problem.

Consider now the discounted problem associated with r_k and β_k. Let $V_k(x)$ denote the optimal (maximal) value of the discounted problem with this reward and discount factor, and let $V_k^-(x)$ denote the minimal value, attained over all strategies. Let $\Gamma_1(x)$ denote the set of conserving actions of the discounted problem with reward r_1 and discount β_1 at state x, that is

$$\Gamma_1(x) = \left\{ a \in \mathbf{A}(x);\ V_1(x) = r_1(x,a) + \beta_1 \sum_{z \in \mathbf{X}} p(z|x,a) V_1(z) \right\}.$$

If the action sets $\mathbf{A}(\cdot)$ are reduced to $\Gamma_1(\cdot)$, then any strategy in the new model is optimal in the initial model for the criterion V_1 because of the following well-known result.

Lemma 3.1. *$V_1(x,\pi) = V_1(x)$ for some $x \in \mathbf{X}$ if and only if $a_t \in \Gamma_1(x_t)$ ($\mathbb{P}_x^\pi$-a.s.) for all $t = 0, 1, \ldots$.*

Proof. For a randomized Markov strategy π, the lemma follows from the optimality equation and the thrifty and equalizing properties; cf. Hordijk (1974), van der Wal (1991), van der Wal and Wessels (1984). Let π be an arbitrary strategy and let an initial sate x be fixed. The lemma by Derman-Strauch-Veinott (Derman 1970, Section 7, Theorem 1, and van der Wal 1981, Lemma 2.1) constructs a randomized Markov strategy σ such that all two-dimensional distributions of (x_t, a_t) under the strategies π and σ coincide for all $t = 0, 1, \ldots$. This result and the correctness of Lemma 3.1 for randomized Markov strategies imply Lemma 3.1. ■

For a strategy $\pi \in \Pi$ and a history $\tilde{h}_n = x_0 a_0 x_1 \ldots x_n a_n \in (X \times A)^{n+1}$, $n = 0, 1, \ldots$, we define the shifted strategy $\tilde{h}_n \pi$ as the strategy which uses, in response to a history $h'_m = x'_0 a'_0 \ldots x'_m$, the action that π would use at epoch $(n+m)$, if the history $\tilde{h}_n h'_m = x_0 a_0 \ldots x_n a_n x'_0 a'_0 \ldots x'_m$ is observed. Namely, $\tilde{h}_n \pi = \sigma$ if for any $m = 0, 1, \ldots$, and for any history $h'_m = x'_0 a'_0 \ldots x'_m \in (X \times A)^m \times \mathbf{X}$, we have $\sigma(\cdot|h'_m) = \pi(\cdot|\tilde{h}_n h'_m)$. We also define $\tilde{h}_{-1} = \emptyset$ and $\emptyset\pi = \pi$.

A strategy π is called persistently optimal, if for any $\tilde{h}_n \in (X \times A)^{n+1}$, $n = 0, 1, \ldots$, the strategy $\tilde{h}_n \pi$ is optimal; cf. Dubins and Sudderth (1977), Feinberg (1986).

Lemma 3.1 implies the following result.

Corollary 3.2. *A policy π is persistently optimal for criterion V_1 if and only if $\pi(\Gamma_1(x_n)|h_n) = 1$ for any $h_n = x_0 a_0 \dots x_n \in H_n$, $n = 0, 1, \dots$.*

Let $X_1 = \{x \in \mathbf{X};\ \mathbf{A}(x) \neq \Gamma_1(x)\}$ and

$$\epsilon_1 = \begin{cases} \min_{x\in\mathbf{X}_1}\Big(V_1(x) - \max_{a\in\mathbf{A}(x)\setminus\Gamma_1(x)}\Big(r_1(x,a) + \beta_1 \sum_{y\in\mathbf{X}} p(y \mid x,a)V_1(y)\Big)\Big), & \text{if } X_1 \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

Recall that $\beta_1 > \beta_k$ for all k. If $\epsilon_1 > 0$, we define

$$N_1 = \min\left\{ n \in \{0,1,\dots\}; \epsilon_1 > \sum_{k=2}^{K} \left(\frac{\beta_k}{\beta_1}\right)^n \max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right)\right\}. \tag{3.1}$$

If $\epsilon_1 = 0$, we define $N_1 = 0$.

Lemma 3.3. *Let $\mathbf{X}$ and $\mathbf{A}$ be finite. If σ is an optimal Markov strategy for the weighted discounted problem and $t \geq N_1$, then $\sigma_t(x_t) \in \Gamma_1(x_t)$ ($\mathbb{P}_x^\sigma$-a.s.) for any $x \in \mathbf{X}$.*

Proof. If $X_1 = \emptyset$ then $\Gamma_1(x) = \mathbf{A}(x)$ for any $x \in \mathbf{X}$, and Lemma 3.3 is trivial. Therefore, we consider the case $X_1 \neq \emptyset$. In this case $\epsilon_1 > 0$ by lemma 3.1 and N_1 is defined by (3.1). The result will be established by contradiction.

First, we prove that for any stationary strategy ϕ, epoch $t = 0, 1, \dots$, and states $x, z \in \mathbf{X}$ such that $\mathbb{P}_x^\phi\{x_t = z\} > 0$, one has

$$\mathbb{E}_x^\sigma \left\{ \sum_{s=t}^{\infty} \sum_{k=1}^{K} \beta_k^s r_k(x_s, a_s) | x_t = z \right\} \geq \mathbb{E}_x^\phi \left\{ \sum_{s=t}^{\infty} \sum_{k=1}^{K} \beta_k^s r_k(x_s, a_s) | x_t = z \right\}. \tag{3.2}$$

To prove (3.2) by contradiction we define a (non-randomized) strategy π through

$$\pi(x_0 a_0 \dots x_n) = \begin{cases} \phi(x_n), & \text{if } n \geq t \text{ and } x_t = z, \\ \sigma(x_n), & \text{otherwise.} \end{cases}$$

If (3.2) does not hold then $V(x;\pi) > V(x;\sigma)$. This contradicts the optimality of σ. Therefore, inequality (3.2) is proved.

For a Markov strategy $\gamma = (\gamma_0, \gamma_1, \dots)$ we consider the shifted strategies $\gamma^n = (\gamma_n, \gamma_{n+1}, \dots)$, $n = 0, 1, \dots$. We can rewrite (3.2) in the following way

$$\sum_{k=1}^{K} \beta_k^t V_k(z;\sigma^t) \geq \sum_{k=1}^{K} \beta_k^t V_k(z;\phi).$$

Therefore,

$$\sum_{k=1}^{K} \beta_k^t \Big(V_k(z;\sigma^t) - V_k(z;\phi) \Big) \geq 0. \tag{3.3}$$

To continue the proof by contradiction we assume that for some x and z in $\mathbf{X}$ there exists some $t \geq N_1$ such that $\sigma_t(z) \notin \Gamma_1(z)$, with $\mathbb{P}_x^\sigma(x_t = z) > 0$. Let ϕ be a stationary strategy such that $\phi(y) \in \Gamma_1(y)$, for any $y \in \mathbf{X}$. By Lemma 3.1,

$$V_1(y;\phi) = V_1(y), \qquad y \in \mathbf{X}. \tag{3.4}$$

We have finally

$$\epsilon_1 > \sum_{k=2}^{K} \frac{\beta_k^t}{\beta_1^t} \left(V_k(z) - V_k^-(z) \right) \geq \sum_{k=2}^{K} \frac{\beta_k^t}{\beta_1^t} \left(V_k(z;\sigma^t) - V_k(z;\phi) \right) \geq V_1(z) - V_1(z;\sigma^t)\ , \tag{3.5}$$

where the first inequality follows from the definition of N_1, the second one follows from the definitions of the value functions, and the third one follows from (3.3) and (3.4).

On the other hand,

$$\begin{aligned} V_1(z;\sigma^t) &= r_1(z,\sigma_t(z)) + \sum_{y\in\mathbf{X}} p(y|z,\sigma_t(z)) V_1(y;\sigma^{t+1}) \\ &\leq r_1(z,\sigma_t(z)) + \sum_{y\in\mathbf{X}} p(y|z,\sigma_t(z)) V_1(y), \end{aligned}$$

so that, from the definition of ϵ_1

$$V_1(z) - V_1(z;\sigma^t) \geq V_1(z) - \left(r_1(z,\sigma_t(z)) + \sum_{y\in\mathbf{X}} p(y|z,\sigma_t(z)) V_1(y) \right) \geq \epsilon_1 \tag{3.6}$$

since $\sigma_t(z) \notin \Gamma_1(z)$. Inequalities (3.5) and (3.6) contradict each other. ■

Corollary 3.4. *Let* $\mathbf{X}$ *and* $\mathbf{A}$ *be finite. There exists a Markov optimal strategy* ϕ *such that* $\phi_t(x) \in \Gamma_1(x)$ *for any* $x \in \mathbf{X}$ *and for any* $t \geq N_1$.

Proof. Let σ be an optimal Markov strategy. If $t \geq N_1$ and $\mathbb{P}_y^\sigma\{x_t = x\} > 0$ for some $x, y \in \mathbf{X}$, then $\sigma_t(x) \in \Gamma_1(x)$ in view of Lemma 3.3. We consider a Markov strategy ϕ such that $\phi_t(x) = \sigma_t(x)$ if $\mathbb{P}_y^\phi\{x_t = x\} > 0$ for some $y \in \mathbf{X}$ and $\phi_t(x) \in \Gamma_1(x)$ if $\mathbb{P}_y^\phi\{x_t = x\} = 0$ for any $y \in \mathbf{X}$. Then $\phi_t(\cdot) \in \Gamma_1(\cdot)$ when $t \geq N = N_1$. Since $\mathbb{P}_y^\phi = \mathbb{P}_y^\sigma$ for any $y \in \mathbf{X}$, the strategy ϕ is optimal. ■

If $\Gamma_1(x)$ is a singleton for each $x \in \mathbf{X}$, then Corollaries 2.3 and 3.4 imply that there is an optimal (N,∞)-optimal strategy ϕ such that $\phi_t(x) = \Gamma_1(x)$, for any $x \in \mathbf{X}$, and for any $t \geq N_1$. Optimal actions $\phi_t(x)$, $x \in \mathbf{X}$ may be found as a solution of N_1-step dynamic programming model with state space $\mathbf{X}$, action space $\mathbf{A}(x)$, $x \in \mathbf{X}$, transition probabilities p, one-step rewards $\tilde{r}_t = \sum_{k=1}^{K} \beta_k^t r_k$, $t = 0, \ldots, N_1 - 1$, and terminal rewards $\sum_{k=1}^{K} \beta_k^{N_1} V_k(x;\psi)$, where $\psi = \phi^{N_1}$ is a stationary strategy.

If $\Gamma_1(x)$ is not a singleton for some $x \in \mathbf{X}$ then, as Corollary 3.4 states, the action sets $\mathbf{A}(x_t)$ for $t \geq N_1$ may be actually reduced to $\Gamma_1(x_t)$. Moreover, by Lemma 3.1, for any strategy π using actions from sets $\Gamma_1(\cdot)$, one has $V_1(x;\pi) = V_1(x)$, $x \in \mathbf{X}$. Therefore, the expected rewards for criterion V_1 from epoch N_1 to ∞ are the same for any Markov strategy ϕ such that $\phi_t(x) \in \Gamma_1(x)$ for any $t \geq N_1$ and for any $x \in \mathbf{X}$. Thus, if our goal is to construct an optimal strategy from moment N_1 onward, then we have reduced the problem with K reward functions $r_1, r_2, \ldots, r_K$ to the problem with $K-1$ reward functions $r_2, \ldots, r_K$.

Fix $k = 2, 3, \ldots, K$. We denote $\Gamma_0(\cdot) = \mathbf{A}(\cdot)$. For $l = k, k+1, \ldots, K$ let $V^{(kl)}$ denote the value function for the discounted problem with state space $\mathbf{X}$, action sets $\Gamma_{k-1}(x)$, $x \in \mathbf{X}$, rewards r_l, transition probabilities p and discount factor β_l. Let $V_-^{(kl)}$ denote the minimal value for the respective minimization problem and let $\Gamma_k(x)$, $x \in \mathbf{X}$, be the sets of conserving actions for the maximization problem:

$$\Gamma_k(x) = \left\{ a \in \Gamma_{k-1}(x);\ V^{(kk)}(x) = r_k(x,a) + \beta_k \sum_{z \in \mathbf{X}} p(z|x,a) V^{(kk)}(z) \right\}. \tag{3.7}$$

We denote $\Gamma(x) = \Gamma_K(x)$, $x \in \mathbf{X}$.

We say that a strategy π is lexicographically optimal at $x \in \mathbf{X}$ if the vector $(V_1(x,\pi), \ldots, V_K(x,\pi))$ is lexicographically greater than or equal to the vector $(V_1(x,\sigma), \ldots, V_K(x,\sigma))$ for any strategy σ (in the lexicographical ordering: that is, $V_l(x,\sigma) > V_l(x,\pi)$ implies $V_i(x,\sigma) < V_i(x,\pi)$ for some $i < l$.) A strategy π is called lexicographically persistently optimal if $\tilde{h}\pi$ is lexicographically optimal at x for any $x \in \mathbf{X}$ and any $\tilde{h} \in \cup_{n=0}^{\infty} (X \times A)^n$.

Lemma 3.5. *A policy π is lexicographically optimal at x if and only if $a_t \in \Gamma(x_t)$ ($\mathbb{P}_x^\pi$-a.s.) for any $t = 0, 1, \ldots$.*

Lemma 3.6. *A policy π is lexicographically persistently optimal if and only if $\pi(\Gamma(x_n)|h_n) = 1$ for all $h_n = x_0 a_0 \ldots x_n \in H_n$, $n = 0, 1, \ldots$.*

Proof. Lemma 3.5 follows from Lemma 3.1 and Lemma 3.5 follows from Corollary 3.2. ∎

Define $X_k = \{x \in \mathbf{X};\ \Gamma_{k-1}(x) \neq \Gamma_k(x)\}$ and

$$\epsilon_k = \begin{cases} \min\limits_{x \in \mathbf{X}_k} \left(V^{(kk)}(x) - \max_{a \in \Gamma_{k-1}(x) \setminus \Gamma_k(x)} \left(r_k(x,a) + \beta_k \sum\limits_{y \in \mathbf{X}} p(y|x,a) V^{(kk)}(y) \right) \right), & \text{if } X_k \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

If $\epsilon_k > 0$, we define

$$N_k = \min \left\{ n \in \{N_{k-1}, N_{k-1}+1, \ldots\};\ \epsilon_k > \sum_{l=k+1}^{K} \left(\frac{\beta_l}{\beta_k} \right)^n \max_x \left\{ V^{(kl)}(x) - V_-^{(kl)}(x) \right\} \right\}.$$

If $\epsilon_k = 0$, we define $N_k = N_{k-1}$. We denote $V^{(1l)} = V_l$, $V_-^{(1l)} = V_l^-$, $l = 1, \ldots, K$.

We are now ready to state an algorithm for the computation of optimal strategies.

Algorithm 3.7.

0. Set $k = 1$.
1. Compute $V^{(kk)}(\cdot)$, $\Gamma_k(\cdot)$, $V^{(kl)}(\cdot)$, $V_-^{(kl)}(\cdot)$, $l = k+1, \ldots, K$, ϵ_k, and N_k .
2. If $\Gamma_k(x)$ is a singleton for all x or $k = K$, set $N = N_k$ and $\Gamma(x) = \Gamma_k(x)$, $x \in \mathbf{X}$ and continue to step 3. Otherwise increase k by 1 and repeat from 1.
3. Fix a stationary strategy ψ such that $\psi(x) \in \Gamma(x)$, $x \in \mathbf{X}$.
4. Compute $V_k(x;\psi)$, $\quad k = 1, 2, \ldots, K$, $x \in \mathbf{X}$.
5. Compute an N-stage optimal Markov strategy σ by solving the N-stage Markov decision problem with state space $\mathbf{X}$, action sets $\mathbf{A}(x)$, $x \in \mathbf{X}$, transition probabilities p, rewards $\tilde{r}_t = \sum\limits_{k=1}^{K} \beta_k^t r_k$, and terminal rewards

$$\sum_{k=1}^{K} \beta_k^N V_k(x;\psi) \ .$$

6. Construct an (N,∞)-stationary optimal strategy ϕ :

$$\phi_t(x) = \begin{cases} \sigma_t(x), & \text{if } t < N, \\ \psi(x), & \text{if } t \geq N. \end{cases}$$

Theorem 3.8. *If the state and action spaces are finite, then there exists an (N,∞)-stationary optimal strategy ϕ for the weighted problem, with $N < \infty$. Algorithm 3.7 finds such a strategy. The stationary strategy ϕ^N which an optimal strategy ϕ uses when the time parameter is greater* *than or equal to N ($\phi^N = \psi$; see Algorithm 3.7) coincides with a stationary strategy which is lexicographically optimal for the problem with discounted criteria $V_1, V_2, \ldots, V_K$.*

Proof. We apply Lemma 3.3 iteratively at most K times. After the kth iteration, $k = 1, \ldots, K-1$, we replace the original model by the model that starts at moment N_k. This means that the initial rewards r_j, $j = 1, \ldots, K$, are replaces by $(\beta_j)^{N_k} r_j$. Lemma 3.3 allows us to reduce the action sets to $\Gamma_k(\cdot)$ after the kth iteration. Since for the kth new model the criteria $V_1, \ldots, V_k$ are insensitive to strategies, we can replace r_j by 0 when $j = 1, \ldots, k$. After a finite number of iterations we have from Lemma 3.3 that the stationary strategy ψ defined in step 3 of Algorithm 3.7 describes some optimal strategy from time N onward, where N is the last value of N_k in the algorithm. Any solution of a finite stage dynamic programming problem described in step 5 of Algorithm 3.7 provides an optimal strategy at moments $0, \ldots, N-1$.

We note that $\prod_{x \in \mathbf{X}} \Gamma(x)$, where the sets $\Gamma(\cdot)$ are computed in Algorithm 3.7, is the set of lexicographically optimal stationary strategies for the problem with discounted criteria $V_1, \ldots, V_K$. ■

Remark 3.9. The computational complexity of solving the finite weighted discounted problem is of the same order of magnitude as that of solving a standard discounted problem. We need to solve a sequence of such problems, but at each step the size of the action space decreases. Finally, we need to solve a finite problem, whose size depends on the ratios of the discount factors (through N).

Remark 3.10. In order to compute N_k we need to compute $V^{(kl)}$ and $V_-^{(kl)}$ for $l = k, k+1, \ldots, K$. Note that these need to be recomputed at every stage since the action set changes at every stage. This requires the solution of $2(K-k)$ discounted problems at stage k. To avoid this computation, one may replace $V_-^{(kl)}$ by a lower bound $R_{(kl)}/(1-\beta_l)$, where $R_{(kl)} = \inf\{r_l(x,a);\ x \in \mathbf{X}, a \in \Gamma_{k-1}(x)\}$, $k = 1, 2, \ldots, K$, $l = k+1, \ldots, K$. Similarly, $V^{(kl)}$ may be replaced by the upper bound $R^{(kl)}/(1-\beta_l)$, where $R^{(kl)} = \sup\{r_l(x,a);\ x \in \mathbf{X}, a \in \Gamma_{k-1}(x)\}$, $k = 1, 2, \ldots, K$, $l = k+1, \ldots, K$. But the computation of $V^{(kk)}$ cannot be avoided, if the algorithm reaches step 1 with this k. This results in a considerable reduction in the complexity of computing N_k. However, this cruder estimate will lead to a larger final value for N, and hence increase the complexity of the finite problem in step 5 of Algorithm 3.7. It is also possible to use $V^{(kl)}$ as upper bounds of $V^{(ml)}$ and $V_-^{(kl)}$ as lower bounds of $V_-^{(ml)}$, $k < m$, in order to reduce the number of discounted dynamic programming problems that should be solved at step 1 of the algorithm.

The next two lemmas characterize the optimal policies for the finite-time part of the problem (step 5 of the algorithm), along the lines of lemma 3.1. For $n = 0, \dots, N-1$ let

$$R(x, n, a) = \sum_{k=1}^{K} \beta_k^n r_k(x, a).$$

Given some arbitrary numbers $V^N(x)$, $x \in \mathbf{X}$, consider the criterion

$$W(x, \pi) = \mathbb{E}_x^\pi \left\{ \sum_{n=0}^{N-1} R(x_n, n, a_n) + V^N(x_N) \right\}.$$

For $n = 0, 1, \dots, N-1$ and for $x \in \mathbf{X}$ we define

$$V^n(x) = \max_{a \in \mathbf{A}(x)} \left\{ R(x, n, a) + \sum_{y \in \mathbf{X}} p(y|x, a) V^{n+1}(y) \right\}. \tag{3.8}$$

Finally for $n = 0, 1, \dots, N-1$ and for $x \in \mathbf{X}$ define the sets of optimal actions

$$\mathbf{A}_n(x) = \left\{ a \in \mathbf{X} : V^n(x) = R(x, n, a) + \sum_{y \in \mathbf{X}} p(y|x, a) V^{n+1}(y) \right\}. \tag{3.9}$$

Lemma 3.11. *A strategy π is optimal for the criterion W with an initial state $x \in \mathbf{X}$ if and only if $a_t \in \mathbf{A}(x_t)$ ($\mathbb{P}_x^\pi$-a.s.) for any $t = 0, \dots, N-1$.*

Lemma 3.12. *A strategy π is persistently optimal for the criterion W if and only if $\pi(\mathbf{A}_t(x_n)|h_n) = 1$ for any $h_n = x_0 a_0 \dots x_n \in H_n$, and any $n = 0, \dots, N-1$.*

Proof. If π is a randomized Markov policy, Lemma 3.12 is the principle of optimality. This and the lemma by Derman-Strauch-Veinott (cf. the proof of Lemma 3.1) imply Lemma 3.12. Lemma 3.11 follows from Lemma 3.12. ■

We set

$$V^N(x) = \sum_{k=1}^{K} \beta_k^N V_k(x, \phi), \tag{3.10}$$

where the number N and the stationary strategy ϕ are computed in Algorithm 3.7. We consider the sets $\mathbf{A}_n(x)$, $n = 0, \dots, N-1$, defined by (3.6)–(3.10) and the sets $\Gamma(x)$, $x \in \mathbf{X}$.

The following two theorems provide a description of the set of optimal strategies.

Theorem 3.13. *Let* $\mathbf{X}$ *and* $\mathbf{A}$ *be finite. One has* $V(x,\pi) = V(x)$ *for a strategy* π *and initial state* $x \in \mathbf{X}$ *if and only if* $a_t \in \mathbf{A}_t(x_t)$ ($\mathbb{P}^\pi_x$*-a.s.) for all* $t = 0,\ldots,N-1$ *and* $a_t \in \Gamma(x_t)$ ($\mathbb{P}^\pi_x$*-a.s.) for all* $t \geq N$.

Theorem 3.14. *Let* $\mathbf{X}$ *and* $\mathbf{A}$ *be finite. A policy* π *is persistently optimal for the weighted total discounted criterion if and only* $\pi(\mathbf{A}_n(x_n)|h_n) = 1$ *for* $n = 0,\ldots,N-1$ *and* $\pi(\Gamma(x_n)|h_n) = 1$ *for* $n \geq N$ *for any* $h_n = x_0a_0\ldots x_n \in H_n$.

Proof. Theorem 3.13 follows from Lemma 3.5, Theorem 3.8, and Lemma 3.11. Theorem 3.14 follows from Lemma 3.6, Theorem 3.8, and Lemma 3.12. ■

Remark 3.15. The existence result of Theorem 3.4 holds when the number of discount factors and reward functions is countable, provided two conditions hold:

(i) $1 > \beta_1 > \beta_2 > \ldots$;

(ii) $\sum_{k=1}^{\infty} |r_k(x,a)| < \infty$ for all x and a.

Since the state and action spaces are finite, $\Gamma_k(x) = \Gamma_{k+1}(x)$ for all k large enough and all x. If the sets $\Gamma_k(x)$ become singletons for all x after a finite number of steps, then the algorithm will identify the optimal strategy within a finite number of steps.

In conclusion, we give examples which demonstrate that if either a state space or an action space is infinite, (N,∞)-stationary optimal strategies may not exist.

Example 3.16. Let $\mathbf{X}$ be a singleton. We therefore omit the state variable from the notation. Let $\mathbf{A} = [0,1]$, $\beta_1 = 2^{-1}$, $\beta_2 = 4^{-1}$, $r_1(a) = 1 - a^2$, and $r_2(a) = a - a^2$. Then we have,

$$V(x,\pi) = \sum_{t=0}^{\infty}(2^{-t}(1-a_t^2) + 4^{-t}(a_t - a_t^2)).$$

Let a_t^* be an optimal decision at epoch $t = 0,1,\ldots$. We have that a_t^* is unique for every t and $a_t^* = 4^{-t}/2(4^{-t} + 2^{-t})$. We note that $a_t^* < a_s^*$ for $t > s$. Therefore, there is no finite N such that there exists an optimal (N,∞)-optimal strategy.

Example 3.17. Let $\mathbf{X} = \{0\} \cup \{\infty\} \cup \{1,2,\ldots\}^2$, $\mathbf{A} = \{b,c\}$,

$$\mathbf{A}(x) = \begin{cases} \{b,c\}, & \text{if } x = (1;m),\ m = 1,2,\ldots, \\ \{b\}, & \text{otherwise.} \end{cases}$$

For any $a \in \mathbf{A}$ let

$$p(y|x,a) = \begin{cases} 1, & \text{if } x = (l+1;m),\ y = (l;m),\ l,m = 1,2,\ldots, \\ 1, & \text{if } x = (1;m),\ y = \infty,\ m = 1,2,\ldots, \\ 1, & \text{if } x = y = \infty, \\ 2^{-(l+m)}, & \text{if } x = 0,\ y = (l;m),\ l,m = 1,2,\ldots, \\ 0, & \text{otherwise.} \end{cases}$$

The process starts at state 0. Then it moves to a state $(l;m)$. Then it moves to $(l-1;m),(l-2;m),\ldots$ until it reaches state $(1;m)$. From this state, it jumps to ∞ and remains there.

Let $\beta_1 = 4^{-1}$, $\beta_2 = 2^{-1}$. Let $r_1(x,b) = r_2(x,b) = 0$ if $x \neq (1;m)$ for some m. If $x = (1;m)$ then $r_1(x,b) = 1$ and $r_2(x,c) = 3 \times 2^{-m}$. So, if at epoch $t = 0,1,\ldots$ the process is at state $(1;m)$, $m = 1,2,\ldots$, the reward is 4^{-t}, if decision b is chosen, and the reward is $3 \times 2^{-(m+t)}$, if decision c is chosen. We have that an optimal Markov policy σ is unique and it is defined by

$$\sigma_t((1;m)) = \begin{cases} b, & \text{if } t < m - \log_2 3, \\ c, & \text{if } t > m - \log_2 3. \end{cases}$$

Therefore, there is no $N < \infty$ such that there exists a stationary (N,∞)-optimal policy.

Acknowledgments

The authors wish to thank Professor Matthew J. Sobel for many useful discussions and suggestions concerning this paper and Professors David C. Nachman, Donald S. Siegel, and Eric A. Stubbs for their comments concerning possible applications to finance.

The research of the second author was performed in part while he was visiting the Systems Research Center, University of Maryland at College Park, and the Department of Networks and Systems, Mathematical Research Center, AT&T Bell Laboratories, Murray Hill. This research was supported in part by the fund for promotion of research at the Technion.

References

Bertsekas, D. (1987). *Dynamic Programming: Deterministic and Stochastic Models.* Prentice-Hall, N.J.

Blackwell, D. (1962). Discrete Dynamic Programming. *Ann. Math. Statist.* **33**, 719–726.

________ (1965). Discounted Dynamic Programming. *Ann. Math. Statist.* **36**, 226–235.

Brealey, R. A. and Myers, S. C. (1988). *Principles of Corporate Finance.* McGraw Hill, New York.

Denardo, E. V. (1967). Contraction Mappings in the Theory Underlying Dynamic Programming. *SIAM Rev.* **9**, 165–177.

________ (1971). Markov Renewal Programs with Small Interest Rates. *Ann. Math. Stat.* **42**, 477–496.

Derman, C. (1970). *Finite State Markovian Decision Processes.* Academic Press, New York.

Dubins, L. and Sudderth, W. D. (1977). Persistently ϵ-Optimal Strategies. *Math. Oper. Res.* **2**, 125–134.

Dynkin, E. B. and Yushkevich, A. A. (1979). *Controlled Markov Processes.* Springer-Verlag, New York.

Federgruen, A. (1983). *Markovian Control Problems.* Mathematical Centre Tracts **97**, Amsterdam.

Feinberg, E. A. (1982a). Non-Randomized Markov and Semi-Markov Strategies in Dynamic Programming. *Theory of Probability and its Applications* **27**, 116–126.

__________ (1982b). Controlled Markov Processes with Arbitrary Numerical Criteria. *Theory of Probability and its Applications* **27**, 486-503.

__________ (1986). The Structure of Persistently Nearly-Optimal Strategies in Stochastic Dynamic Programming Problems. *Lecture Notes in Control and Inform. Sci.* **81**, 22–31, Springer-Verlag, New York.

__________ and Shwartz, A. (1992) Constrained Markov decision models with weighted discounted rewards. Preprint.

Filar, J. A. and Vrieze, O. J. (1989). Weighted Reward Criteria in Competitive Markov Decision Processes. *Proc. IFAC Symp. on Dynamic Modeling and Control of National Economies*, Scotland, 93–98.

Ghosh, M. K. and Marcus, S. I. (1991). Infinite Horizon Controlled Diffusion Problems with Some Nonstandard Criteria. *J. of Mathematical Systems, Estimation and Control* **1**, 45–69.

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

Hinderer, K. (1970). *Foundations of Non Stationary Dynamic Programming with Discrete Time Parameter*. Lecture Notes in Operations Research **33**, Springer-Verlag, New York.

Hordijk, A. (1974). *Dynamic Programming and Markov Potential Theory.* Mathematical Centre Tracts **51**, Amsterdam.

Keeney, R.L. and Raiffa H. (1976). *Decisions with Multiple Objectives.* Wiley, New York.

Krass, D., Filar, J. A. and Sinha, S. S. (1990). A Weighted Markov Decision Process. Preprint.

Ross, S. M. (1984). *Introduction to Stochastic Dynamic Programming.* Academic Press, New York.

Schäl, M. (1975). Conditions for Optimality in Dynamic Programming and for the Limit of n-Stage Optimal Policies to be Optimal. *Z. Wahr. verw. Gebiete* **32**, 179–196.

_____ (1981). An Operator-Theoretical Treatment of Negative Dynamic Programming. In *Dynamic Programming and its Applications* (ed. M. L. Puterman), Academic Press, New York, 351–368.

Shapley, L. S. (1953). Stochastic Games. *Proc. Nat. Acad. Sci. USA* **39**, 1095–1100.

Sobel, M. J. (1991). Discounting and Risk Neutrality. Preprint.

van der Wal, J. (1981). *Stochastic Dynamic Programming.* Mathematical Centre Tracts **139**, Amsterdam.

__________ and Wessels, J. (1984). On the Use on Information in Markov Decision Processes. *Statistics & Decisions* **2**, 1–21.

Veinott, Jr., A. F. (1966). On Finding Optimal Policies in Discrete Dynamic Programming with No Discounting. *Ann. Math. Stat.* **37**, 1284–1294.

Whitt, W. (1980). Representation and Approximation of Noncooperative Sequential Games. *SIAM J. Control Optim.* **18**, 33-48.

Whittle, P. (1982). *Optimization Over Time; Dynamic Programming and Stochastic Control.* Wiley, New York.

FEINBERG: 303 HARRIMAN HALL, SUNY AT STONY BROOK, STONY BROOK, NY 11794-3775

SHWARTZ: DEPARTMENT OF ELECTRICAL ENGINEERING, TECHNION—ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL