Death and Discounting

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
URL: http://www.ee.technion.ac.il/~adam

Submitted May 2000, revised October 2000.

Abstract

Markov Decision Processes with discounted cost are equivalent to processes with a finite, random duration, and hence the discount factor models a (random) time horizon for the life of the process. We elaborate on this idea, but show that an objective function which is a linear combination of several discounted costs (each with a different discount factor) does not, in general, model processes with several time scales, but rather processes with partial information.

*This research was supported in part by the fund for promotion of research at the Technion, and in part by the fund for promotion of sponsored research at the Technion.

1 Introduction

It is well known [2] that a Markov Decision Process with a discounted cost criterion can be viewed as a model of a process with total cost and finite (random) duration. Consequently, the discount factor is often given the interpretation of a measure life-span of the process. This raises a question of whether Markov Decision Processes that exhibit phenomena on several different time scales (e.g. session time vs. project time) can be modelled through an objective function which is a linear combination of several discounted cost functionals, each with a different discount factor. This seems appealing, since a theory, as well as algorithms for such Markov decision processes exist [3, 4]. Feinberg and Shwartz [5] describe some models in which several discount factors arise naturally.

In Section 2 we couple a discounted MDP with a total-cost MDP with finite (random) duration, establishing that the models are equivalent in a strong sense. However, models with several time scales and models with several discount factors are not equivalent, as we show in Examples 5 and 7. The "mixed discounts" model is shown to be equivalent to a model with several time scales, in which the "death," or the end of one time scale, cannot be detected. Finally, we show that it is always possible to provide an equivalence between the largest discount (in a mixed-discount model) and the (probabilistically) longest random duration.

1.1 Discounted Markov Decision Processes

A Markov Decision Process (MDP) is characterized by a state space $\mathbf{X}$, an action space $\mathbf{A}$ with actions $\mathbf{A}(x)$ available when in state x, and transition matrix $P(a) = \{p(y \mid x, a)\}_{x,y}$ when action a is used. A policy π assigns to each history $h_n := x_0 a_0 x_1 ... x_n$ of length n a probability distribution $\pi(\cdot \mid h_n)$ on $\mathbf{A}$ so that $\pi(\mathbf{A}(x_n) \mid h_n) = 1$. Each policy π and initial state x induce a probability measure $\mathbb{P}^\pi_{\mathrm{x}}$ on the canonical space $H_\infty = (\mathbf{X} \times \mathbf{A})^\infty$, with corresponding expectation operator $\mathbb{E}^\pi_{\mathrm{x}}$.

Fix $0 < \beta < 1$ and let $c(x, a)$ be an immediate cost for using action a in state x. The β-discounted cost of using policy π and starting at x is

$$V(x; \pi) := \mathbb{E}_{\mathrm{x}}^{\pi} \sum_{n=0}^{\infty} \beta^n c(x_n, a_n) \ . \tag{1.1}$$

We denote this Markov Decision Process by $\mathbb{M} := \{\mathbf{X}, \mathbf{A}, \mathbf{A}(\cdot), p, c, V\}$.

Assumption 1 *The state space is countable and the costs are non-negative. The action spaces* $\mathbf{A}$, $\mathbf{A}(x)$ *are compact metric spaces.*

The results hold verbatim under much weaker conditions, such as measurability and existence of expectations, of the type required for any Markov decision process. Generalizing to, say, Borel state and action spaces requires only some measurability requirements and a change in notation—from sums to integrals. However, our aim here is not generality, but basic structure.

2 Discounted MDPs and cemeteries

We now define a related MDP by adding to the state space (in the nomenclature of Markov processes) an absorbing state Δ (a "cemetery.") Define

$$p_\beta(y \mid x, a) := \beta \cdot p(y \mid x, a), \qquad \text{for } y \neq \Delta,\ x \neq \Delta, \tag{2.1}$$

$$p_\beta(\Delta \mid x, a) = 1 - \beta, \qquad \text{for } x \neq \Delta \tag{2.2}$$

$$p_\beta(\Delta \mid \Delta, a) = 1 \, . \tag{2.3}$$

The MDP $\mathbb{M}_\Delta := \{\mathbf{X}_\Delta, \mathbf{A}, \mathbf{A}(\cdot), p_\beta, c_\Delta, V_\Delta\}$ has state space $\mathbf{X}_\Delta := \mathbf{X} \cup \{\Delta\}$. The action space $\mathbf{A}$ as well as $\mathbf{A}(x)$ are unchanged for $x \neq \Delta$, and we let $\mathbf{A}(\Delta) := \{a_\Delta\}$ for some (arbitrary) $a_\Delta \in \mathbf{A}$. Let

$$c_\Delta(\Delta, a_\Delta) := 0, \quad c_\Delta(x, a) := c(x, a) \text{ otherwise.} \tag{2.4}$$

Let $\tilde{\mathbb{P}}_x^{\tilde{\pi}}$ and $\tilde{\mathbb{E}}_x^{\tilde{\pi}}$ be the probability (expectation) corresponding to the modified chain. Note that policies in the new model are different since there is an additional state, and possibly additional information. Finally, let

$$V_\Delta(x;\tilde{\pi}) := \tilde{\mathbb{E}}_x^{\tilde{\pi}} \sum_{n=0}^{\infty} c_\Delta(x_n, a_n) \ . \tag{2.5}$$

This total cost is well-defined since the costs are positive.

To obtain a sample-path description of the modified process, denote

$$\tau := \inf\{n : x_n = \Delta\}\,. \tag{2.6}$$

Then, by (2.1)–(2.3), τ has a geometric distribution. Let $\{\zeta_n\}$ be a sequence of $\{0,1\}$ valued, i.i.d. random variables with mean β, which are independent of all other random variables in our model. We can then represent τ as

$$\tau = \inf\left\{n : \prod_{t=0}^{n} \zeta_t = 0\right\}\ , \tag{2.7}$$

so that $x_n = \Delta$ if and only if $\zeta_t = 0$ for some $0 \le t \le n$. This allows us to construct the two MDPs on the same probability space, as follows. As in Schäl [6], we include the $\{\zeta_t\}$ in our probability space. The stochastic process of MDP $\mathbb{M}$ evolves as before, while that of $\mathbb{M}_\Delta$ evolves in the following way. If at time t the process is in state $x \neq \Delta$ and action a is chosen, then if $\zeta_t = 1$ the process moves to state y with probability $p(y \mid x, a)$. If $\zeta_t = 0$, the process moves to state Δ. The MDP of $\mathbb{M}_\Delta$ now differs from $\mathbb{M}$ in two ways. First, it has an additional state, so in particular there is more information for choosing a control. Second, it moves to state Δ at time τ, and remains there forever. Setting $\prod_{t=0}^{-1}\zeta_t := 1$ as a convention, we have for MDP $\mathbb{M}$

Lemma 2 *For any immediate cost c, any x, π and $n \geq 0$,*

$$\mathbb{E}_{\mathrm{x}}^{\pi}\,\beta^n c(x_n, a_n) = \mathbb{E}_{\mathrm{x}}^{\pi}\left[\left(\prod_{t=0}^{n-1}\zeta_t\right) c(x_n, a_n)\right] \tag{2.8}$$

and in particular

$$V(x,\pi) = \mathbb{E}_{\mathrm{x}}^{\pi} \sum_{n=0}^{\infty}\left(\prod_{t=0}^{n-1}\zeta_t\right) c(x_n, a_n) \tag{2.9}$$

$$= \mathbb{E}_{\mathrm{x}}^{\pi} \sum_{n=0}^{\tau-1} c(x_n, a_n)\,. \tag{2.10}$$

Proof. The policy π is not a function of the $\{\zeta_t\}$. Due to their definition, they are independent of the state and the action processes. Therefore,

$$\mathbb{E}_x^\pi \left(\prod_{t=0}^{n-1} \zeta_t\right) c(x_n, a_n) = \mathbb{E}_x^\pi \left\{ \mathbb{E}_x^\pi \left[\left(\prod_{t=0}^{n-1} \zeta_t\right) c(x_n, a_n) \,\middle|\, x_n, a_n \right] \right\} \tag{2.11}$$

$$= \mathbb{E}_x^\pi \left\{ c(x_n, a_n)\, \mathbb{E}_x^\pi \left[\left(\prod_{t=0}^{n-1} \zeta_t\right) \,\middle|\, x_n, a_n \right] \right\} \tag{2.12}$$

$$= \beta^n\, \mathbb{E}_x^\pi\, c(x_n, a_n)\,. \tag{2.13}$$

Summing over n yields (2.9), and (2.10) follows by (2.7). ∎

Policies in $\mathbb{M}_\Delta$, however, rely on richer information: by observing whether or not state Δ has been visited by time $0 \le t \le n$, we know the values of $\{\zeta_t,\, 0 \le t \le n\}$. Let

$$\tilde{h}_n := x_0 \zeta_0 a_0 x_1 \ldots x_{n-1} a_{n-1} \zeta_{n-1} x_n\,. \tag{2.14}$$

Policies in $\mathbb{M}_\Delta$ are functions of $\tilde{h}_n$. For each π in $\mathbb{M}$, define a $\tilde{\pi}$ in $\mathbb{M}_\Delta$ by

$$\tilde{\pi}(\cdot \mid \tilde{h}_n) := \begin{cases} \pi(\cdot \mid h_n) & \text{if } \zeta_t = 1 \text{ for all } t = 0, 1, \ldots, n-1 \\ \delta_{a_\Delta} & \text{otherwise.} \end{cases} \tag{2.15}$$

That is, $\tilde{\pi}$ agrees with π before τ, and from time τ on chooses only a_Δ.

Conversely, given a policy $\tilde{\pi}$ in $\mathbb{M}_\Delta$ we can define a policy π in $\mathbb{M}$ through

$$\pi(\cdot \mid x_0 a_0 x_1 \ldots x_n) := \tilde{\pi}(\cdot \mid x_0 1 a_0 x_1 \ldots 1 x_n)\,, \tag{2.16}$$

that is, π chooses actions at time n in the same way $\tilde{\pi}$ would if $\zeta_t = 1,\ t = 0, 1, \ldots, n-1$. Since in $\mathbb{M}_\Delta$ we know the values of the ζ_t, there are more policies in $\mathbb{M}_\Delta$. On the other hand, this knowledge is not very useful, since as soon as $\zeta_t \ne 1$, actions become irrelevant, both in terms of transitions and in terms of cost. The theorem below establishes the equivalence of a discounted-cost MDP, the MDP with total cost and random duration, and the related MDP with an absorbing state—a cemetery, with a total cost criterion. This generalizes the result in Derman [2, pp. 31–32].

Theorem 3 *The models* $\mathbb{M}$ *and* $\mathbb{M}_\Delta$ *are equivalent in the following sense. Fix a policy* π *in* $\mathbb{M}$ *and define* $\tilde{\pi}$ *through* (2.15). *Then, for every cost function* c*, every* $x \neq \Delta$ *and* $n \geq 0$,

$$\mathbb{E}^{\pi}_{\mathrm{x}}\, \beta^n c(x_n, a_n) = \tilde{\mathbb{E}}^{\tilde{\pi}}_x c_\Delta(x_n, a_n)\,. \tag{2.17}$$

Conversely, given $\tilde{\pi}$ *in* $\mathbb{M}_\Delta$*, define* π *through* (2.16). *Then* (2.17) *holds. Consequently, under either* (2.15) *or* (2.16)*, for every* $x \neq \Delta$,

$$V(x, \pi) = V_\Delta(x, \tilde{\pi})\,. \tag{2.18}$$

Proof. If the same actions are used, both processes agree until τ, and by the argument following (2.7), they have the same distribution, namely

$$\begin{aligned} &\mathbb{P}^{\pi}_{\mathrm{x}}\left(x_n = x,\, a_n = a \text{ and } \zeta_t = 1,\, t = 0, 1, \dots, n-1\right) \\ &\qquad\qquad = \tilde{\mathbb{P}}^{\tilde{\pi}}_x\left(x_n = x,\, a_n = a \text{ and } \zeta_t = 1,\, t = 0, 1, \dots, n-1\right)\,, \end{aligned} \tag{2.19}$$

while if $\zeta_t = 0$ for some $t \leq n$ then the both products below equal 0. Using Lemma 2 and the independence of the ζ_t,

$$\mathbb{E}^{\pi}_{\mathrm{x}}\, \beta^n c(x_n, a_n) = \mathbb{E}^{\pi}_{\mathrm{x}}\left[\left(\prod_{t=0}^{n-1} \zeta_t\right) c(x_n, a_n)\right] \qquad \text{by Lemma 2} \tag{2.20}$$

$$= \tilde{\mathbb{E}}^{\tilde{\pi}}_x\left[\left(\prod_{t=0}^{n-1} \zeta_t\right) c_\Delta(x_n, a_n)\right] \qquad \text{by (2.19)} \tag{2.21}$$

$$= \tilde{\mathbb{E}}^{\tilde{\pi}}_x\left[c_\Delta(x_n, a_n)\right] \tag{2.22}$$

since the product of the ζ_t equals 1 before τ, while after τ the state is Δ and $c_\Delta(\Delta, a_\delta) = 0$. This establishes (2.17), and (2.18) is now immediate. ∎

In particular, the models are equivalent for standard optimization:

Theorem 4 *Consider the models* $\mathbb{M}$ *and* $\mathbb{M}_\Delta$*. For every cost function* c

$$\sup_\pi V(x, \pi) = \sup_{\tilde{\pi}} V_\Delta(x, \tilde{\pi}) \quad \textit{for } x \neq \Delta, \tag{2.23}$$

where in each case the suprema are over all available policies.

We thus established the equivalence of the three models: discounting, geometric life (defined through τ or through $\{\zeta_t\}$), and a cemetery. By Theorem 4 these models are equivalent for the purposes of optimization.

3 Mixed discount MDPs

Feinberg and Shwartz [5] describe some models in which several discount factors appear (for a theoretical analysis of these models, as well as references to related models, see Feinberg and Shwartz [3, 4]). The criteria in these models involve several cost functionals simultaneously, each with a different discount factor and different immediate cost. So, let $0 < \beta_i < 1$, $i = 1, \dots, I$, fix some cost functions $c_i(x, a)$, $i = 1, \dots, I$ and define

$$V^i(x, \pi) := \mathbb{E}_{\mathrm{x}}^{\pi} \sum_{n=0}^{\infty} \beta_i^n c_i(x_n, a_n) \qquad i = 1, 2, \dots, I. \tag{3.1}$$

Let $\mathbb{M}^I$ denote the model with I discount factors and costs, with objective

$$V^I(x, \pi) := \sum_{i=1}^{I} V^i(x, \pi). \tag{3.2}$$

Let $\{\zeta_t^i,\ i = 1, \dots, I\ , t = 0, \dots\}$ be $\{0, 1\}$ valued random variables with means β_i respectively. Denote $\zeta_t := \{\zeta_t^i,\ i = 1, \dots, I\}$ and assume that $\{\zeta_t,\ t = 0, \dots\}$ are statistically independent. Define τ^i through (2.7). Then by Lemma 2, the discounts can be exchanged for geometric time-horizons:

$$\begin{aligned} V^i(x, \pi) &= \mathbb{E}_{\mathrm{x}}^{\pi} \sum_{n=0}^{\infty} \left(\prod_{t=0}^{n-1} \zeta_t^i \right) c_i(x_n, a_n) && (3.3) \\ &= \mathbb{E}_{\mathrm{x}}^{\pi} \sum_{n=0}^{\tau^i - 1} c_i(x_n, a_n)\,. && (3.4) \end{aligned}$$

No assumptions are made concerning the statistical dependencies between the components of $\zeta_t = \{\zeta_t^i,\ i = 1, \dots, I\}$. Define $\mathbb{M}_\Delta^I$ as in Section 2.

3.1 Counter examples

It is tempting to regard the discounts β_i as indications of the natural time scale of each reward process. However, in contrast with the single discount case, here information on ζ_t^i may change the value function:

Example 5 *Consider a single state x, action set $\{0,1\}$, and $I = 2$. Let*

$$\mathbb{P}(\zeta_t^1 = 1) = \beta_1 > \beta_2 = \mathbb{P}(\zeta_t^2 = 1) \tag{3.5}$$

where $\{\zeta_t^1,\, t \geq 0\}$ is independent of $\{\zeta_t^2,\, t \geq 0\}$. Set

$$c_1(0) = 2 \qquad c_2(0) = 0 \tag{3.6}$$

$$c_1(1) = 0 \qquad c_2(1) = 1. \tag{3.7}$$

Let $\tilde{\pi}$ be the policy that chooses $a = 0$ until the first time that $\zeta_t^1 = 0$ and then switches to $a = 1$. Then there is an $\epsilon > 0$ so that for any policy π which does not depend on the $\{\zeta_t^i\}$ we have

$$V_\Delta^1(\tilde{\pi}) + V_\Delta^2(\tilde{\pi}) > V^1(\pi) + V^2(\pi) + \epsilon. \tag{3.8}$$

To show this, define τ^i as before, and denote $x^+ := \max\{x, 0\}$. Then

$$V_\Delta^1(\tilde{\pi}) + V_\Delta^2(\tilde{\pi}) = \sum_{t=0}^{\infty} 2\,\mathbb{E}^{\tilde{\pi}}\,\mathbf{1}[\tau_1 = t] + 1 \cdot \mathbb{E}\,(\tau_2 - \tau_1)^+ \tag{3.9}$$

$$= 2\sum_{t=0}^{\infty} \beta_1^t + \mathbb{E}\,(\tau_2 - \tau_1)^+ \tag{3.10}$$

$$:= \frac{2}{1-\beta_1} + \epsilon\ , \tag{3.11}$$

and $\epsilon > 0$. If the ζ_t^i are not observed, since $2\beta_1^t > \beta_2^t$ for all t the maximum of $V^1(\pi) + V^2(\pi)$ is $\frac{2}{1-\beta_1}$, obtained by choosing $a_t = 0$ for all t. ∎

Remark 6 *Consider the two extreme modelling situations: in the first, the $\{\zeta_t^i : t = 0, 1, \dots\}, i = 1, 2$ are independent. In the second, let $\{\tilde{\zeta}_t^i : t = 0, 1, \dots\}$ be independent collections of i.i.d. random variables, with*

$\mathbb{P}(\tilde\zeta_t^1 = 1) = \beta_1$, $\mathbb{P}(\tilde\zeta_t^2 = 1) = \frac{\beta_2}{\beta_1}$. *Set* $\zeta_t^1 = \tilde\zeta_t^1$ *and* $\zeta_t^2 = \tilde\zeta_t^1 \cdot \tilde\zeta_t^2$. *In both cases, the random variables satisfy the assumptions of the example with the discount factors* (3.5)*. However, from a modelling point of view, the cases are different, and in the second case* $\tau_2 \le \tau_1$ *so that* $\epsilon = 0$*. Mixed-discounting (with different discount factors) does not distinguish between the two, and therefore cannot be a reasonable model of different time scales.*

Even when the τ_i and dependent, observing the ζ_t^i may be beneficial.

Example 7 *Consider Example 5, but with the following modification. Suppose* $2\beta_1 > \beta_2 > \beta_1$ *and let* $\zeta_t^1 = \zeta_t^2 \cdot \tilde\zeta_t^1$ *where* $\mathbb{P}(\tilde\zeta_t^1 = 1) = \frac{\beta_1}{\beta_2}$*. Then, without information about the* ζ_t^i*, the cost is the same as in Example 5, by the same arguments. In addition,* $\mathbb{E}(\tau_2 - \tau_1)^+ > 0$ *so that* $\epsilon > 0$*.*

Thus, information on the random duration is essential both in the dependent and in the independent case.

3.2 Equivalence in the unobserved case

As long as τ_i (equivalently $\{\zeta_t^i\}$) cannot be observed, the equivalences hold.

Lemma 8 *Let* $c_{ij}(n, x, a)$, $1 \le i \le I$, $j = 1, 2, \ldots$ *be (time-dependent) cost functions so that* $c_{ij}(n, \Delta, a) = 0$*. There exists a policy* π *in* $\mathbb{M}^I$ *so that*

$$\mathbb{E}_x^\pi \beta_i^n c_{ij}(n, x_n, a_n) = \alpha(x, n, i, j) \quad \text{for all } x, n, i \text{ and } j \tag{3.12}$$

for some function α *if and only if* π *satisfies* (3.12) *in* $\mathbb{M}_\Delta^I$*.*

Proof. This follows from Lemma 2 and Theorem 3, since the proof there does not depend on the choice of immediate cost function. ∎

The following conclusion is now immediate:

Theorem 9 *The mixed-discount model and the model with several durations are equivalent as long as policies do not depend on the* ζ_t^i*.*

The representation through a "time of death" extends easily:

Corollary 10 *There are random variables τ_i so that*

$$V^i(x,\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\tau_i - 1} c_i(x_n, a_n) \quad \textit{for all } i. \tag{3.13}$$

Proof. As before, we use independent, $\{0,1\}$-valued random variables $\{\zeta_t^i,\ i = 1, \ldots, t = 0, 1, \ldots\}$ with $\mathbb{E}\,\zeta_t^i = \beta_i$ to construct τ_i. As before, due to the independence the additional random variables need not be included in the information h_n, and the result follows. ∎

We obtain the same analogy: discounting corresponds to a random time horizon, with geometric distribution. Moreover, we can model several "time scales τ_i" simultaneously by using appropriate discount factors. Note, however, that the time horizon is statistically independent of the dynamics, by definition and since we did not allow policies to depend on the $\{\zeta_t^i\}$ (or, equivalently, on $\{\tau_i < n\}$). If death is detectable, the models differ.

Thus, modelling the mixed-discount model as an exit problem (by introducing a "cemetery") is not possible. Intuitively, the reason is that in the new model there are several "states of life," corresponding to several indices i. In a standard MDP model we assume that the state is observed at each stage. Therefore, if we model the multiple-discount problem by adding "states of life," these would be observable and could be used to design policies. This was not an issue in the single discount case since, in that case, when Δ is reached the process dies, in the sense that there are neither choice of actions nor costs.

4 Extensions and concluding remarks

In general, we can replace the longest time scale with a random duration, as follows. Without loss of generality, assume $0 < \beta_I < \cdots < \beta_1 < 1$. Fix β so that $\beta_1 \le \beta < 1$ and let $\tilde{\beta}_i := \beta_i/\beta_1 \le 1$. With this value of β we can now follow the construction (2.1)–(2.3) of Section 2, which does not depend on the costs. Define τ through (2.6), so that (2.7) holds. Set $c_i^\Delta(x,a) := c_i(x,a)$ for $x \neq \Delta$ and $c_i^\Delta(\Delta, a) := 0$ for all i. Define $\mathbb{M}^I_{\Delta_1}$ by adding a cemetery to the model of Section 3.

Corollary 11 *The models $\mathbb{M}^I$ and $\mathbb{M}^I_{\Delta_1}$ are equivalent. Namely, fix a policy π in $\mathbb{M}^I$ and define $\tilde{\pi}$ through* (2.15)*. Then for every $x \neq \Delta$,*

$$V^I(x,\pi) = V^I_\Delta(x,\tilde{\pi})\,. \tag{4.1}$$

Proof. Follows immediately from Theorem 3 by redefining the cost functions and representing each of the random variables ζ_t^i as a product of ζ_t^1 and an independent, $\{0,1\}$ valued random variable. ∎

This defines a time scale for the process (which is limited by the largest discount), and the other discounts are scaled appropriately.

The model with random duration can be given the following representation. Enlarge the state space to $\mathbb{M}_\times := 2^{\{\Delta_i,\ 1\leq i\leq I\}} \times \mathbf{X}$, that is, the first coordinate is a subset of the $\{\Delta_i\}$. Start the process with the empty subset. Then, as soon as $\zeta_t^i = 1$ for some i, add Δ_i to the first coordinate of the process. Once added, it cannot be removed. Then the first coordinate determines which time scales expired. Thus, any problem with a number of random durations is equivalent to a problem with a collection of cemeteries.

Finally, note that all the results of this paper extend immediately to stochastic games. In particular, Theorem 4 holds for a zero-sum game if (2.23) is replaced with equality of the value of the games.

In conclusion, any mixed discount problem can be transformed into a problem with random duration, but where the durations cannot be observed. In the other direction, any MDP with several random durations and total cost can be reduced to a similar one but with a discounted cost. In particular, choosing $\beta = \max_i \beta_i$ allows us to reduce the number of random durations by one, through the introduction of a discount. If the random durations are not observed, so that they cannot be used as part of the control, then the problem is equivalent to a mixed discount problem.

References

[1] W.L. Cooper, "Negative binomial sums of random variables and discounted reward processes," *J. Appl. Prob.*, to appear.

[2] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, 1970.

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

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

[5] E.A. Feinberg and A. Shwartz, "Constrained dynamic programming with two discount factors: applications and an algorithm," *IEEE Transactions on Automatic Control* **TAC-44** 628–630, 1999.

[6] M. Schäl, "Asymptotic results for sequential Markov decision models under uncertainty," *Statistics and decisions* **2**, pp. 39–62, 1984.