Eventually-stationary policies for Markov decision models with non-constant discounting

(Invited paper)

Yair Carmon
Department of Electrical Engineering
Technion — IIT
Haifa 32000, Israel

Adam Shwartz[*]
Department of Electrical Engineering
Technion — IIT
Haifa 32000, Israel
adam@ee.technion.ac.il

ABSTRACT

We investigate the existance of simple policies in finite discounted cost Markov Decision Processes, when the discount factor is not constant. We introduce a class called "exponentially representable" discount functions. Within this class we prove existence of optimal policies which are eventually stationary—from some time N onward, and provide an algorithm for their computation. Outside this class, optimal policies with this structure in general do not exist.

Keywords

Markov Decision Processes, Discounted Cost, Mixed Discounting, Hyperbolic Discounting, General Discounting Function

1. INTRODUCTION

In Markov decision models (MDPs), discounting is used to model the fact that the further in the future something happens, the less important it is. Simple discounting, where the reward is multiplied by a constant discount factor at each epoch, arises naturally from economic considerations, when constant rates of interest (bonds) or constant inflation are assumed. Such models are relatively easy to analyze. In particular, in this case there exists an optimal policy which is stationary, namely deterministic and independent of the time and of past states [5]. Because it can be intuitively understood and handily analyzed, simple discounting has been thoroughly researched and applied to countless models—from machine learning, computer networks to game theory and psychology.

However, in general the decrease in value of the future need not be geometric. In models of "learning curves", the cost of "getting to know" the system is added to the original criterion. The discounting in this criterion typically has a power-law form, though some models have geometric learning curves. Additional geometric decreases arise from models such as "Moore's law," in which a cost (in this case of a unit of computation power) decreases geometrically fast. The addition of an exponentially-decreasing component (with a rate different from the discount factor) to a discounted Markov decision model results in a weighted discounted criterion—that is, a criterion that is the sum of several standard discounted criteria. A theory for finite models with weighted discounted criteria was developed by Feinberg and Shwartz [2]. The main results are that for such criteria there are optimal policies that are stationary from some finite time onwards, called N-stationary policies, and an algorithm for the computation of these policies is given.

In models of human preferences, it makes sense to use discounting with a decreasing rate. Intuitively, while tomorrow may be less important than today, a year and a day from now is just about the same as a year from now. "Hyperbolic discount functions", of the form $(1+\alpha n)^{-\gamma/\alpha}$ with $\alpha, \gamma > 0$, feature a decreasing discounting rate, and are reported to effectively model psychological preferences (see [3] for presentation, and [6] for critique). Even Moore's Law seems to be breaking down, leading to non-constant discounting.

Because of the difficulty of analyzing decision processes with general discount functions, most theoretical results are obtained with "toy functions", for example $f = [1, \delta\beta, \delta\beta^2, \delta\beta^3, \ldots]$, which in a sense has a decreasing discounting rate for $0 < \delta < 1$, and often serves as a replacement of the hyperbolic function mentioned above (see, e.g. [4]). Following the lines of the weighted discounted theory, we define the class of "exponentially representable" functions, prove that when they are used as discount functions there exist N-stationary optimal policies, and describe a computation algorithm. These functions display the decreasing discount rate of the hyperbolic discount functions. However, the hyperbolic discount functions are not exponentially representable, and moreover—exponentially representable discount functions cannot be used to model power-law learning curves.

Below we define the model and the exponentially representable functions, and state our main result. In section 2 we develop the algorithm for the computation of the optimal policy and through it prove our result. In section 3 we further discuss the meaning of exponential representability and which functions belong in that class. Finally, in section 4 we show that the N-stationary property is not always assured.

[*]Corresponding Author

ValueTools 2008, October 21–23, 2008, Athens, GREECE.
Copyright ©2008 ICST ISBN # 978-963-9799-31-8 .

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4392
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4392

1.1 Markov Decision Processes

Consider a discrete time process with a finite state space $\mathbf{X}$, where $x_n \in \mathbf{X}$ denotes the state at time $n = 0, 1, 2, \ldots$. At each time, an action is chosen from a finite action set $\mathbf{A}(x)$. Let $\mathbf{A} = \bigcup_{x \in \mathbf{X}} \mathbf{A}(x)$ be the (finite) action space—the set of possible actions and $a_n \in \mathbf{A}(x_n)$ the chosen action at time n. The state and chosen action at time n determine the probability distribution of the next state through the transition probability $p\left(x_{n+1} | x_n, a_n\right)$, assumed to be independent of the time and of further information on the past. Additionally, the state and action determine the immediate reward $r\left(x_n, a_n\right)$ given at time n.

The rule used to choose the action at each time is called a policy. We call $h_n = x_0 a_0 \cdots x_{n-1} a_{n-1} x_n$ the history at time n. Since the choice of action is required to be causal, h_n contains all the information available for making it. The most general policy is therefore a mapping of every history h_n to a probability measure $\pi\left(\cdot | h_n\right)$ on $\mathbf{A}(x_n)$. Policies in which only one action is possible for each given history are called deterministic, so that $a_n = \pi(h_n)$. Policies which depend only on the current state and the time, i.e. $\pi\left(\cdot | h_n\right) = \pi\left(\cdot | x_n, n\right)$ are called Markov policies, and Markov policies which do not depend on the time are called stationary.

The discounted criterion assigns to each policy and initial state numerical value

$$V^{\beta}\left(x; \pi\right) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \beta^n r\left(x_n, a_n\right)$$

where $\mathbb{E}_x^{\pi}$ is the expectation induced by policy π, given that $h_0 = x$, and $0 < \beta < 1$ is the discount factor. Since $r\left(x, a\right)$ is bounded, the value is always finite.

We discuss a more general discounted criterion, in which β^n is replaced with $f(n)$:

$$V^{\mathrm{gen}}\left(x; \pi\right) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} f(n) r\left(x_n, a_n\right) \,. \tag{1}$$

A sufficient condition for the above summation to be well defined is $|f(n)| \leq K \beta^n$ for some $0 < \beta < 1$ or, equivalently, $f(n) = \beta^n g(n)$ for $0 < \beta < 1$ and some bounded function $g(n)$. We call a function that satisfies this condition *exponentially bounded.*

The weighted discounted criterion is a sum of a finite number of standard discounted criteria, each with a possibly different immediate reward function:

$$\begin{aligned} &V^{\mathrm{wd}}\left(x; \pi\right) \\ &= \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \sum_{k=1}^{K} \beta_k^n r_k\left(x_n, a_n\right) \text{ with } \beta_1 > \beta_2 > \cdots > \beta_K \,. \end{aligned} \tag{2}$$

Let us define the maximal and minimal values of an MDP, respectively:

$$V(x) \equiv \sup_{\pi} V(x; \pi) \qquad V^{-}(x) \equiv \inf_{\pi} V(x; \pi) \,. \tag{3}$$

An optimal policy is a policy for which $V\left(x; \pi\right) = V(x)$, for all $x \in \mathbf{X}$.

Definition 1.1. *A Markov policy π is called* N-stationary *if*

$$\pi(x, n) = \pi(x, N) \quad \forall x \in \mathbf{X}, n \geq N \,.$$

Definition 1.2. *A function $f : \{0, 1, \ldots\} \to \mathbb{R}$ is called* exponentially representable *(ER) if there exist sequences $\{c_k\}_{k=1}^{\infty}$ and $\{\beta_k\}_{k=1}^{\infty}$ such that:*

- $\{\beta_k\}_{k=1}^{\infty}$ is positive, strictly decreasing and $\beta_1 < 1$.
- $f(n) = \sum_{k=1}^{\infty} c_k \beta_k^n$, the sum converges absolutely after some $N < \infty$.

Example 1.3. *The function*

$$f(n) = \frac{1}{e-1} \sum_{k=1}^{\infty} \frac{1}{k!} \left(\beta_0^k\right)^n = \frac{e^{\beta_0^n} - 1}{e-1}$$

is ER for $0 < \beta_0 < 1$ and logarithmically convex ($(\log f\left(x\right))'' > 0$)—hence with a decreasing rate of discounting, since the rate is inversely proportional to $f\left(n+1\right)/f\left(n\right)$. This is the required property in aforementioned the human preferences models.

1.2 The main result

Our starting point will be the following result on the structure of optimal policies under criteria (2) and (1).

Theorem 1.4. *In a weighted discounted MDP (2), there exists an optimal policy which is Markov and deterministic. This holds also in MDPs with general discounting (1) when the discount function is exponentially bounded.*

For a weighted discounted MDP, a proof (under more general conditions) is given in [2], Thm. 2.2. The idea is to embed the process in an ordinary discounted MDP with a countable state space $\tilde{\mathbf{X}} = \left(\mathbf{X} \times \mathbb{N}_+\right)$, where time is added so that the new state is $\tilde{x}_n = \left(x_n, n\right)$, and use the standard result that discounted criteria have deterministic and stationary optimal policies. The same embedding can be carried out in the case of a single general discount function which is exponentially bounded. This theorem also extends straightforwardly to a criterion that is a sum of several criteria with general discount functions, as long as those functions are exponentially bounded. Since this is a straightforward extension, we omit the details.

From now on we focus on ER discount functions. Since they are exponentially bounded, using Theorem 1.4 we restrict our policies to be Markov and deterministic.

Theorem 1.5. *Consider a finite MDP with an ER discount function. There exists an N-stationary optimal policy for this problem, with $N < \infty$. This policy can be found using Algorithm 2.6.*

2. OPTIMAL POLICIES

The generalized discounted criterion in (1), with $f\left(n\right)$ ER, is an infinite version of the weighted discounted criterion. To see this, find $\{c_k\}_{k=1}^{\infty}$ and decreasing $\{\beta_k\}_{k=1}^{\infty}$ such that $f(n) = \sum_{k=1}^{\infty} c_k \beta_k^n$, and rewrite the criterion as

$$\begin{aligned} V^{\mathrm{gen}}\left(x; \pi\right) &= \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} f(n) r\left(x_n, a_n\right) \\ &= \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \sum_{k=1}^{\infty} \beta_k^n c_k r\left(x_n, a_n\right) \end{aligned} \tag{4}$$

which is an infinite weighted discounted criterion with $r_k\left(x_n, a_n\right) = c_k r\left(x_n, a_n\right)$.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4392
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4392

We now adapt the algorithm from part 3 of [2] to infinite weighted discounted criteria induced by an ER discount function. We will also prove that this algorithm halts after a finite number of iterations, and provide a bound on that number. Let

$$V_k\left(x;\pi\right) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \beta_k^n c_k r\left(x_n, a_n\right) \tag{5}$$

denote the value of the k^{th} summand in (4), and let $V_k\left(x\right)$ and $V_k^-\left(x\right)$ be the maximal and minimal value for initial state x. Define a "conserving set":

$$\Gamma_1(x) \equiv$$
$$\left\{ a \in \mathbf{A}(x) \mid V_1(x) = c_1 r(x,a) + \beta_1 \sum_{y\in\mathbf{X}} p\left(y|x,a\right) V_1(y) \right\} . \tag{6}$$

It is easy to see that a policy is optimal for this criterion if and only if it chooses actions from the (conserving) set $\Gamma_1\left(x\right)$ when in state x: see Lemma 3.1 in [2].

Let $\mathbf{X}_1 = \{x \in \mathbf{X} \mid \Gamma_1(x) \neq \mathbf{A}(x)\}$ be the set of states for which suboptimal actions for criterion V_1 exist. If $\mathbf{X}_1 \neq \emptyset$, define:

$$\varepsilon_1 \equiv \min_{x\in\mathbf{X}_1, a\in\mathbf{A}(x)\setminus\Gamma_1(x)}$$
$$\left(V_1(x) - c_1 r(x,a) - \beta_1 \sum_{y\in\mathbf{X}} p\left(y|x,a\right) V_1(y) \right) . \tag{7}$$

ε_1 is the value of the smallest "mistake" one can make in the choice of a single action, with regard to criterion V_1. If $\mathbf{X}_1 = \emptyset$ define $N_1 \equiv 0$. Otherwise define:

$$N_1 = \min\left\{ n \geq 0 \mid \varepsilon_1 > \sum_{k=2}^{\infty} \left(\frac{\beta_k}{\beta_1}\right)^n \max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right) \right\} . \tag{8}$$

Lemma 2.1. *If $f(n)$ is ER, N_1 is well defined and finite.*

Proof. Define $S(n) = \sum_{k=2}^{\infty} \left(\beta_k/\beta_1\right)^n \max_{x\in\mathbf{X}} \left(V_k(x)\right.$ $\left.-V_k^-(x)\right)$. Denote the span semi-norm of $r\left(x,a\right)$ by $M = \max_{x\in\mathbf{X},a\in\mathbf{A}(x)} r\left(x,a\right) - \min_{x\in\mathbf{X},a\in\mathbf{A}(x)} r\left(x,a\right)$. Then

$$\forall k \ : \ \max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right) \leq |c_k| \frac{M}{1-\beta_k} \leq |c_k| \frac{M}{1-\beta_1} \tag{9}$$

$$S(n) \leq \frac{\beta_1^{-n} M}{1-\beta_1} \sum_{k=2}^{\infty} \beta_k^n \left|c_k\right| . \tag{10}$$

This proves N_1 is finite—see [1]. □

Remark 2.2. *Let $\mathbf{X} = \{x_0\}$ and $\mathbf{A} = \{a_1, a_2\}$ with $r\left(x, a_1\right) = 1$ and $r\left(x, a_2\right) = 0$. Then, $\max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right) = \left|c_k\right|/1 - \beta_k$ for any k, so $S(n) = \beta_1^{-n} \sum_{k=2}^{\infty} \beta_k^n \left|c_k\right|$. Here $S(n) \underset{n\to\infty}{\longrightarrow} 0$ only if $\sum_{k=2}^{\infty} \beta_k^n c_k$ converges absolutely for some $N < \infty$, i.e. only if $f(n)$ is ER. Thus for a given discount function f, the bound N_1 is well-defined for* any model *if* and only if *the discount function is ER.*

Using the definitions (6) and (8) of $\Gamma_1(x)$ and N_1 respectively,

Lemma 2.3. *Consider a finite MDP with an ER discount function. If σ is an optimal Markov policy then $n \geq N_1$ and $\mathbb{P}_x^{\sigma}\left\{x_n = z\right\} > 0$ imply $\sigma(z,n) \in \Gamma_1(z)$.*

Proof. The proof extends that of Lemma 3.3 in [2], using the same basic ideas. It is simple to show (see [2]) that the optimality of σ implies that for any stationary policy ϕ, time m and state $z \in \mathbf{X}$ such that $\mathbb{P}_x^{\sigma}\left\{x_m = z\right\} > 0$:

$$\mathbb{E}_x^{\sigma} \left\{ \sum_{n=m}^{\infty} f(n) r\left(x_n, a_n\right) | x_m = z \right\}$$
$$\geq \mathbb{E}_x^{\phi} \left\{ \sum_{n=m}^{\infty} f(n) r\left(x_n, a_n\right) | x_m = z \right\} .$$

By substituting $f(n) = \sum_{k=1}^{\infty} c_k \beta_k^n$ in the above equation, denoting the time-shifted optimal policy by $\sigma^m\left(x,n\right) = \sigma\left(x, n+m\right)$, and using V_k as defined in (5), the above inequality can be rewritten ask

$$\sum_{k=1}^{\infty} \beta_k^m V_k\left(z;\sigma^m\right) \geq \sum_{k=1}^{\infty} \beta_k^m V_k\left(z;\phi\right)$$

and therefore

$$\sum_{k=2}^{\infty} \beta_k^m \left(V_k\left(z;\sigma^m\right) - V_k\left(z;\phi\right)\right) \geq \beta_1^m \left(V_1\left(z;\phi\right) - V_1\left(z;\sigma^m\right)\right) . \tag{11}$$

We now suppose by contradiction that for some time $m \geq N_1$ and for some state $z \in \mathbf{X}$ such that $\mathbb{P}_x^{\sigma}\left\{x_m = z\right\} > 0$, $\sigma\left(z,m\right) \notin \Gamma_1(z)$. Let ϕ be the optimal stationary policy for criterion V_1, so that $V_1(x,\phi) = V_1(x)$ for all $x \in \mathbf{X}$. Since $m \geq N_1$, and by the definition of N_1 (which is meaningful due to the fact that f is ER) we have:

$$\begin{aligned} \varepsilon_1 &> \sum_{k=2}^{\infty} \left(\frac{\beta_k}{\beta_1}\right)^m \left[V_k(z) - V_k^-(z)\right] \\ &\geq \sum_{k=2}^{\infty} \left(\frac{\beta_k}{\beta_1}\right)^m \left[V_k(z;\sigma^m) - V_k^-(z;\phi)\right] \\ &\geq V_1\left(z;\phi\right) - V_1\left(z;\sigma^m\right) \end{aligned} \tag{12}$$

where the second inequality comes from the definitions and the last from (11). On the other hand, since $\sigma\left(z,m\right) \notin \Gamma_1(z)$ and from the definition of ε_1:

$$V_1(z) - V_1\left(z;\sigma^m\right) \geq V_1(z)$$
$$- \left(r\left(z, \sigma\left(z,m\right)\right) + \beta_1 \sum_{y\in\mathbf{X}} p\left(y|z, \sigma\left(z,m\right)\right) V_1(y) \right) \geq \varepsilon_1 \tag{13}$$

Inequalities (12) and (13) contradict each other, thus proving this lemma. □

If $\Gamma_1(x)$ is a singleton for all $x \in \mathbf{X}$, then the lemma requires any optimal policy to be N_1-stationary, and determines the stationary part of the policy. If it is not a singleton, we know that after time N_1 our action sets reduce to $\Gamma_1(x)$ and for every admissible policy, V_1 will attain its maximum value and therefore be irrelevant. Our task therefore becomes finding the optimal policy for the weighted sum starting from the second discount factor, with the action sets restricted to Γ_1. To iterate the above process, define recursively for $k > 1$, the restricted action sets in iteration k — $\mathbf{A}_k\left(x\right) = \Gamma_{k-1}\left(x\right)$,

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4392
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4392

the m^{th} value function restricted to the k^{th} action set — $V_m^{\mathbf{A}_k}(x)$, and similarly the minimal value function $V_m^{-,\mathbf{A}_k}(x)$. Additionally:

$$\Gamma_k(x) \equiv \Big\{ a \in \mathbf{A}_k(x) \mid V_k^{\mathbf{A}_k}(x) = c_k r(x,a) + \beta_k \sum_{y \in \mathbf{X}} p\left(y|x,a\right) V_k^{\mathbf{A}_k}(y) \Big\} \quad (14)$$

$$\mathbf{X}_k = \{ x \in \mathbf{X} \mid \Gamma_k(x) \neq \mathbf{A}_k(x) \} \quad (15)$$

$$\varepsilon_k \equiv \min_{x \in \mathbf{X}_k, a \in \mathbf{A}_k(x) \setminus \Gamma_k(x)} \left(V_k^{\mathbf{A}_k}(x) - c_k r(x,a) - \beta_k \sum_{y \in \mathbf{X}} p\left(y|x,a\right) V_k^{\mathbf{A}_k}(y) \right) \quad (16)$$

$$N_k = \min \Bigg\{ n \geq N_{k-1} \mid \varepsilon_k > \sum_{m=k+1}^{\infty} \left(\frac{\beta_m}{\beta_k} \right)^n \max_{x \in \mathbf{X}} \left(V_m^{\mathbf{A}_k}(x) - V_m^{-,\mathbf{A}_k}(x) \right) \Bigg\} \quad (17)$$

where ε_k is set to ∞ in the case that $\mathbf{X}_k = \emptyset$. Again, N_k is well defined when $f\left(n\right)$ is ER. With these definitions, the following is evident:

Lemma 2.4. *Consider a finite MDP with ER discount function. If σ is an optimal Markov policy, then for every $k \geq 1$, $n \geq N_k$ and state $z \in \mathbf{X}$ such that $\mathbb{P}_x^{\sigma}\left\{x_n = z\right\} > 0$, we have $\sigma(z,n) \in \Gamma_k(z)$.*

Proof. By induction using Lemma 2.3 and the above definitions. □

We will now prove that iterating this procedure does indeed provide us with an N-stationary policy after a finite and bounded number of computations.

Lemma 2.5. *Consider a finite MDP with an ER discount function. Let $S = |\mathbf{X}|$. Then for all $k \geq 2S-1$ and every $x \in \mathbf{X}$, $\Gamma_k\left(x\right) = \Gamma_{2S-1}\left(x\right)$.*

Proof. If $\Gamma_{2S-1}\left(x\right)$ is a singleton for all $x \in \mathbf{X}$, then the lemma is immediate. Otherwise, let $\Phi = \{\phi_1, \phi_2, ..., \phi_L\}$ be the set of stationary policies such that $\phi_i\left(x\right) \in \Gamma_{2S-1}\left(x\right)$ for all $x \in \mathbf{X}$, $i = 1, 2, ..., L$. For $\phi \in \Phi$, define $f_\phi : [0,1) \rightarrow \mathbb{R}^S$ as

$$\left[f_\phi\left(\beta\right)\right]_s = \mathbb{E}_{x^s}^{\phi} \sum_{n=0}^{\infty} \beta^n r\left(x_n, a_n\right) \ ,$$

so that $V_k\left(x^s; \phi\right) = c_k \left(f_\phi\left(\beta_k\right)\right)_s$. Let $[P_\phi]_{m,n} \equiv p(x^n|x^m, \phi\left(x^s\right))$ and $\left[r_\phi\right]_s = r\left(x^s, \phi\left(x^s\right)\right)$ be the state transition matrix and reward vector induced by ϕ_i. Then

$$f_\phi\left(\beta\right) = r_\phi + \beta P_\phi f_\phi\left(\beta\right) \Rightarrow f_\phi\left(\beta\right) = \left(I - \beta P_\phi\right)^{-1} r_\phi \ . \quad (18)$$

Since P_ϕ is stochastic, by the Perron–Frobenius theorem $I - \beta P_\phi$ is invertible for $\beta \in [0,1)$ and singular for $\beta = 1$. Since $M^{-1} = \operatorname{adj}\left(M\right) / \det\left(M\right)$, by (18) every entry (coordinate) of f_ϕ is a rational function of β, with numerator degree $S-1$ and denominator degree S, with a pole at $\beta = 1$, which possibly cancels with a zero in some of the entries. Since $\phi \in \Phi$ if and only if it is optimal for all criteria V_k for $k = 1, 2, ..., 2S-1$ (under different action sets for each k), all policies in Φ must have the same values for $\beta_1, \beta_2, ..., \beta_{2S-1}$. Consequently, for every $i, j \leq L$:

$$f_{\phi_i}\left(\beta_k\right) = f_{\phi_j}\left(\beta_k\right) \ , \ \forall k = 1, 2, ..., 2S-1 \ . \quad (19)$$

Fix i and j and consider each entry of $f_{\phi_i}\left(\beta\right) - f_{\phi_j}\left(\beta\right) = 0$. It is a polynomial equation of degree $2S-2$ (since the common poles at $\beta = 1$ cancel). However, according to (19), this polynomial has $2S-1$ distinct roots—and is therefore identically zero. We conclude that $f_{\phi_i}\left(\beta\right) = f_{\phi_j}\left(\beta\right)$ for all $\beta \in [0,1)$ and every two policies $\phi_i, \phi_j \in \Phi$, and accordingly $V_k(x;\phi)$ is the same over all $\phi \in \Phi$, for each $x \in \mathbf{X}$ and $k \geq 2S-1$. This means that for $k \geq 2S-1$, all possible policies have identical values, and will therefore all be optimal. Since the set of optimal policies remains constant, so do the conserving sets. □

The proof of Theorem 1.5 now follows—see [1] for details.

The computation of $\{N_k\}_{k=1}^{2S-1}$ involves evaluations of infinite sums, which are usually not feasible. To avoid this, we can instead find upper bounds $\hat{N}_k \geq N_k$ for each k, and compute an $\hat{N}_{2S-1}$-stationary optimal policy with a stationary part determined by the conserving sets. One way to find $\hat{N}_k$ is to use the semi-norm bounds in (10). In each iteration, the semi-norm of the reward function should be computed with respect to the restricted action set, and therefore decrease.

Algorithm 2.6.

1. Find $\{\beta_k\}_{k=1}^{\infty}$ and $\{c_k\}_{k=1}^{\infty}$ of Definition 1.2, set $S = |\mathbf{X}|$ and $k = 1$.
2. Compute $\Gamma_k\left(x\right)$ for all $x \in \mathbf{X}$, ε_k and N_k or an appropriate upper bound.
3. If $\Gamma_k\left(x\right)$ is a singleton for every $x \in \mathbf{X}$, or $k = 2S+1$, set $N = N_k$ and continue. Else set $\mathbf{A}_{k+1}\left(\cdot\right) = \Gamma_k\left(\cdot\right)$, increment k by 1 and go back to step 2.
4. Fix a stationary policy ψ, such that $\psi\left(x\right) \in \Gamma_k\left(x\right)$ for all $x \in \mathbf{X}$.
5. Compute an optimal Markov policy σ for the N-step MDP with immediate rewards $r_n\left(x,a\right) = f\left(n\right) r\left(x,a\right)$, for $n = 0, 1, ...N-1$ and terminal reward
$$\mathbb{E}_{x_N}^{\psi} \sum_{n=0}^{\infty} f\left(n+N\right) r\left(x_{n+N}, a_{n+N}\right) = \sum_{k=1}^{\infty} \beta_k^N V_k\left(x_N; \psi\right) \ .$$
6. The N-stationary optimal policy is defined by $\pi\left(x,n\right) = \sigma\left(x,n\right)$ for $n < N$ and $\pi\left(x,n\right) = \psi\left(x\right)$ for $n \geq N$.

Step 5 is standard—see [5].

Remark 2.7. *Our results can be extended to criteria of the form:*

$$V\left(x;\pi\right) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \sum_{k=1}^{K} f_k\left(n\right) r_k\left(x_n, a_n\right) \ , \ f_k\left(n\right) = \sum_{i=1}^{\infty} c_{i,k} \beta_{i,k}^n$$

where for each k, $f_k\left(n\right)$ is ER and the additional condition

$$\beta_{i,k} > \beta_{1,k+1} \ , \ \forall i, k. \quad (20)$$

Lemmas 2.3 and 2.4 can be extended by changing the definitions of the N_k's to include the rest of the discount functions, with condition (20) making sure they remain well defined. The N-stationary optimal policy can then be obtained

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4392
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4392

by finding $\Gamma_{2S-1,1}(x)$ for the first discount function. In the case it is not a singleton, the action space will be restricted appropriately, and the procedure will be applied to f_2. This may continue until $\Gamma_{2S-1,K}(x)$ is computed, from which we may choose the stationary part of the optimal policy arbitrarily. Finally, we remark that if $r_k(\cdot) = b_k r(\cdot)$ for some function $r(x,a)$, the procedure will end in the computation of $\Gamma_{2S-1,1}(x)$, since afterwards all permissible policies for the stationary part will have the same value.

3. STRUCTURE OF ER FUNCTIONS

ER functions behave asymptotically as exponential functions:

Lemma 3.1. *Let $f(n)$ be an ER function. Then there exist $0 < \beta < 1$ such that*

$$\lim_{n\to\infty} \beta^{-n} f(n) = c \neq 0 \text{ and } c < \infty \, .$$

Proof. Write $f(n) = \sum_{k=1}^{\infty} c_k \beta_k^n$. Without loss of generality, we may assume that $c_1 \neq 0$. Since f is ER, we have absolute convergence from some time $N < \infty$. Therefore, for $n > N$ and some $C < \infty$:

$$\beta_1^{-n} \left| \sum_{k=2}^{\infty} c_k \beta_k^n \right| \leq \beta_1^{-n} \sum_{k=2}^{\infty} |c_k| \beta_k^n$$
$$< \frac{\beta_2^{n-N}}{\beta_1^n} \sum_{k=2}^{\infty} |c_k| \beta_k^N = C \left(\frac{\beta_2}{\beta_1}\right)^n \underset{n\to\infty}{\rightarrow} 0 \, .$$

Consequently, $\lim_{n\to\infty} \beta_1^{-n} \sum_{k=2}^{\infty} c_k \beta_k^n = 0$ and choosing $\beta \equiv \beta_1$ we have,

$$\lim_{n\to\infty} \beta^{-n} f(n) = \lim_{n\to\infty} c_1 + \beta_1^{-n} \sum_{k=2}^{\infty} c_k \beta_k^n = c_1 \neq 0$$

and $c_1 < \infty$. □

Functions with power-law form, like $(1+n^2)^{-1}$ or the hyperbolic discount function do not satisfy the conclusion of Lemma 3.1, and are therefore not ER. The same holds for sub-exponential functions, like $1/n!$ and e^{-n^2}. Moreover, functions of the form $g(n)\beta^n$, where $g(n) \to 0$ or $g(n) \to \infty$ non-exponentially, are also not ER for the same reason. Examples are $n\beta^n$ and $\beta^n/(1+n)$ for some $0 < \beta < 1$.

4. A DISCOUNT FUNCTION WITH NO N-STATIONARY OPTIMAL POLICY

When a discount function decreases monotonically it seems natural that it should produce a behavior that is monotonic, or stationary. However, this is not true: we provide an example of a discount function and a model for which there is no N-stationary optimal policy. By our results, the discount function is not ER.

Consider the function $f(n) = \beta^n h(n)$, with some $0 < \beta < 1$ and

$$h(n) = \begin{cases} 2 & n \bmod 6 = 0 \\ 1 & \text{otherwise} \end{cases} = [2,1,1,1,1,1,2,1,1,1,\ldots]$$

which is periodic with period 6. The condition of Lemma 3.1 does not hold for $f(n)$, and it is therefore not ER.

Now consider the following (deterministic) model:

$$\begin{gathered} \mathbf{X} = \{1,2,3,4,5\} \ , \ \mathbf{A}(1) = \{a_1, a_2\} \ , \\ \mathbf{A}(2) = \mathbf{A}(3) = \mathbf{A}(4) = \mathbf{A}(5) = \{a\} \\ p(2|1,a_1) = p(3|1,a_2) = p(4|3,a) \\ = p(5|4,a) = p(1|5,a) = p(1|2,a) = 1 \end{gathered} \tag{21}$$

with the immediate reward function

$$\begin{aligned} & r(1,a_1) = 3 \ , \ r(1,a_2) = 4 \ , \\ & r(2,a) = r(3,a) = r(4,a) = r(5,a) = 0 \, . \end{aligned}$$

Let σ be a hypothetical N-stationary, optimal and possibly randomized policy for this process. Since state 1 is recurrent, there exists a time $M_0 \geq N$ such that $\mathbb{P}_1^{\sigma}(x_{M_0} = 1) > 0$. Consequently, $\mathbb{P}_1^{\sigma}(x_{M_0+4} = 1) > 0$ and $\mathbb{P}_1^{\sigma}(x_{M_0+8} = 1) > 0$, since at those times there must be a positive probability that a single action is used repeatedly. We know M_0 is even because every return to state 1 takes either 2 or 4 steps, and therefore, either M_0, M_0+4 or M_0+8 divides by 6. We may thus choose $M \geq N$ such that M is a multiple of 6, and $\mathbb{P}_1^{\sigma}(x_M = 1) > 0$.

Define the shifted value criterion:

$$V^M(1;\pi) \equiv \mathbb{E}^{\pi} \left\{ \sum_{n=M}^{\infty} \beta^n h(n) r(x_n, a_n) \, | x_M = 1 \right\} \, .$$

Criterion $V^M(1;\pi)$ comprises the part of $V(1;\pi)$ that involves times M and onwards. If an optimal Markov policy for criterion $V(1;\pi)$ has any chance of reaching state 1 in time M, it must also optimize criterion $V^M(1;\pi)$, when taken from times M onwards — this is a form of the principle of optimality. Therefore, and considering that σ is an optimal Markov policy for $V(1;\pi)$, and $\mathbb{P}_1^{\sigma}(x_M = 1) > 0$ by our choice of M, the stationary policy $\sigma^M(\cdot) = \sigma(\cdot, n+M)$ must be of optimal for criterion V^M. However,

$$V^M(1;\pi) = \mathbb{E}_1^{\pi} \sum_{n=0}^{\infty} \beta^{n+M} h(n+M) r(x_n, a_n) = \beta^M V(1;\pi)$$

where the last equality follows from the periodicity of $h(n)$ and our choice of M. Since $V^M(1;\cdot)$ is proportional to $V(1;\cdot)$, σ^M is optimal for the original criterion as well, when starting from state 1. Moreover, if σ^M is randomized, a stationary and deterministic policy $\hat{\sigma}^M$ with $V(1;\sigma^M) = V(1;\hat{\sigma}^M) = V(1)$ can be obtained. This is done by using, for every state, an action that has positive probability under σ^M. In conclusion, if this problem has an N-stationary optimal policy, a stationary and deterministic policy must maximize $V(1;\cdot)$.

Let $\sigma_1(1) = a_1, \sigma_2(1) = a_2$ be the two stationary policies in this model, and consider the periodic Markov policy:

$$\pi(1,n) = \begin{cases} a_2 & n \bmod 6 = 0 \\ a_1 & n \bmod 6 = 4 \end{cases}$$

Let $\beta = 0.45$. The values of the 3 policies will then be:

$$V(1;\pi) = \frac{8+3\beta^4}{1-\beta^6} \approx 8.19 \tag{22}$$

$$V(1;\sigma_1) = \frac{6+3\beta^2+3\beta^4}{1-\beta^6} \approx 6.79 \tag{23}$$

$$V(1;\sigma_2) = \frac{8+4\beta^4+4\beta^8}{1-\beta^{12}} \approx 8.17 \, . \tag{24}$$

Evidently, both stationary policies are suboptimal when $x_0 = 1$. Therefore, there cannot be an N-stationary optimal policy, since our considerations have shown that it will result in an optimal, stationary and deterministic policy when starting from state 1. Since we chose $\beta < 1/2$, $h(n) > \beta h(n+1)$ for every n, making the discount function monotonically decreasing, as required.

Acknowledgments: Adam Shwartz holds The Julius M. and Bernice Naiman Chair in Engineering. His research was supported in part by the fund for promotion of research at the Technion, by the fund for promotion of sponsored research at the Technion, and by the B. and I. Green Research Fund.

5. REFERENCES

[1] Y. Carmon and A. Shwartz. "Markov decision processes with exponentially representable discounting," CCIT Report 675, *I*nternal Report, Technion 2008.

[2] E.A. Feinberg and A. Shwartz. Markov decision models with weighted discounted criteria. *M*athematics of Operations Research, 19(1):152–168, 1994.

[3] G. Loewenstein and D. Prelec. Anomalies in intertemporal choice: Evidence and an interpretation. *T*he Quarterly J. Economics, 107(2):573–597, 1992.

[4] D. Laibson. Golden eggs and hyperbolic discounting. *T*he Quarterly J. Economics, 112(2):443–477, 1997.

[5] M.L. Puterman. *M*arkov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, N.Y., 1994.

[6] A. Rubinstein. Economics and psychology? the case of hyperbolic discounting. *I*nternational Economic Review, 44(4):1207–1216, 2003.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4392
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4392