MARKOV DECISION PROCESSES WITH EXPONENTIALLY REPRESENTABLE DISCOUNTING

YAIR CARMON, ADAM SHWARTZ

Faculty of Electrical Engineering, Technion - IIT[1]

April 2008

Abstract. We generalize the geometric discount of finite discounted cost Markov Decision Processes to "exponentially representable" discount functions, prove existence of optimal policies which are stationary from some time N onward, and provide an algorithm for their computation. Outside this class, optimal "N-stationary" policies in general do not exist.

Keywords: Markov Decision Processes, Discounted Cost, Mixed Discounting, Hyperbolic Discounting, general discounting function

[1]Corresponding author: Adam Shwartz, Faculty of Electrical Engineering, Technion - IIT, Technion City, Haifa 32000 Israel. adam@ee.technion.ac.il

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 in economics when considering constant rates of interest or inflation. Such models are relatively easy to analyze, and it is well known [5] that for this case there exists an optimal policy which is stationary, namely independent of the time and of past states. 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.

Difficulties arise in models where the rate of discounting is not constant. The most immediate example concerns interest rates, but it is definitely not the only one. In models of human preferences, it makes sense to use discounting with a decreasing rate. An intuitive reason for this is the fact that while tomorrow may be considerably less important than today, a year and a day from now is just about the same as a year from now. "Hyperbolic discount functions", which are 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).

Because of the difficulty of analyzing decision processes with general discount functions, most theoretical results are obtained with "toy functions". A noticeable example is the function $f = \left[1, \delta\beta, \delta\beta^2, \delta\beta^3, \dots\right]$, 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, for example, [4]).

In models of "learning curves", the cost of "getting to know" the system is added to the original criterion. The discounting in the added learning curve criterion typically has a power-law form, though some models have geometric learning curves. The addition of an exponentially-decreasing learning curve 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.

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 may display the decreasing discount rate of the hyperbolic discount functions. However, we show that the hyperbolic discount functions are not exponentially representable, and moreover—that exponentially representable discount functions cannot be used to model power-law learning curves.

In the rest of this section we give definitions of MDPs, relevant value criteria, N-stationary policies and exponentially representable functions. We then state our main result formally. 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 give an example of a monotonically decreasing discount function for which there is no N-stationary optimal policy, in order to demonstrate that the existence of the N-stationary property is not always assured.

1.1. Markov Decision Processes.

Consider a discrete time process with a finite state space $\mathbf{X}$, finite action sets $\mathbf{A}(x)$ with $\mathbf{A} = \bigcup_{x\in\mathbf{X}} \mathbf{A}(x)$. Let x_n (a_n) denotes the state (resp. chosen action) at time n. The transition probability are $p\left(x_{n+1}|x_n, a_n\right)$. The immediate reward at time n is $r\left(x_n, a_n\right)$.

We call $h_n = x_0 a_0 \cdots x_{n-1} a_{n-1} x_n$ the history at time n. A policy is 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. $\pi(\cdot|h_n) = \pi(\cdot|x_n, n)$ are called Markov policies, and Markov policies which do not depend on the time are called stationary.

For every initial state $x \in \mathbf{X}$, the discounted criterion assigns to each policy a value

$$V^{\beta}(x;\pi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \beta^n r(x_n, a_n) \tag{1}$$

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

In this work we discuss a more general discounted criterion, in which β^n is replaced with a discount function $f(n)$:

$$V^{\text{gen}}(x;\pi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} f(n) r(x_n, a_n) \ . \tag{2}$$

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.*

A third relevant criterion is the weighted discounted criterion, which is a sum of a finite number of standard discounted criteria, each with a possibly different immediate reward function:

$$V^{\text{wd}}(x;\pi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \sum_{k=1}^{K} \beta_k^n r_k(x_n, a_n) \ \text{ with } \beta_1 > \beta_2 > \cdots > \beta_K \, . \tag{3}$$

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

$$V(x) \equiv \sup_{\pi} V(x;\pi) \qquad V^{-}(x) \equiv \inf_{\pi} V(x;\pi) \, . \tag{4}$$

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

1.2. **Two further definitions.** Before stating our main result, two more concepts need to be defined.

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

$$\pi(x,n) = \pi(x,N) \quad \forall x \in \mathbf{X}, n \geq N \, . \tag{5}$$

Definition 1.2. A function $f : \{0,1,\dots\} \to \mathbb{R}$ is called *exponentially representable* 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 $\beta_1 < 1$.
- $f(n) = \sum_{k=1}^{\infty} c_k \beta_k^n$, and the sum converges absolutely, starting from some time $N < \infty$.

Example 1.3. The function

$$f(n) = \frac{1}{e-1} \sum_{k=1}^{\infty} \frac{1}{k!} \left(\beta_0^k\right)^n = \frac{e^{\beta_0^n} - 1}{e-1} \tag{6}$$

is exponentially representable for $0 < \beta_0 < 1$. It is logarithmically convex ($\left(\log f(x)\right)'' > 0$), which is equivalent to a decreasing rate of discounting, since the rate is inversely proportional to $f(n+1)/f(n)$. This is the required property in the human preferences models mentioned in the introduction.

1.3. **The main result.** Our starting point will be the following result on the structure of optimal policies under criteria (3) and (2).

Theorem 1.4. *In a weighted discounted MDP (3), there exists an optimal policy which is Markov and deterministic. This holds also in MDPs with general discounting (2) when the discount function is exponentially bounded.*

For the case of a weighted discounted MDP, a full proof (under more general conditions) is given in [2], Theorem 2.2. The idea of the proof is to embed the process in an ordinary discounted MDP with a countable state space, where the time is added to the state, 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.

From now on our discussion will focus on exponentially representable discount functions. Since those functions are exponentially bounded, in light of Theorem 1.4 we can and shall restrict our policies to be Markov and deterministic unless specifically mentioned otherwise. We can now state our main result.

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

In what follows, we prove our result by construction.

2. Optimal policies for exponentially representable discount functions

The generalized discounted criterion in (2), with $f(n)$ exponentially representable, 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

$$V^{\text{gen}}(x;\pi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} f(n) r(x_n, a_n) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \sum_{k=1}^{\infty} \beta_k^n c_k r(x_n, a_n) \tag{7}$$

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

In the rest of this section we adapt the algorithm described in part 3 of [2] to the case of infinite weighted discounted criteria induced by an exponentially representable discount function. To this end, we will review the construction of the algorithm, and add to the proofs as necessary. We will also prove that this algorithm halts after a finite number of iterations, and provide a bound on that number. Let

$$V_k(x;\pi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \beta_k^n c_k r(x_n, a_n) \tag{8}$$

denote the value of the k^{th} summand in (7), and let $V_k(x)$ and $V_k^-(x)$ be the maximal and minimal value for initial state x, respectively. For each $x \in \mathbf{X}$, we define a "conserving set":

$$\Gamma_1(x) \equiv \left\{ a \in \mathbf{A}(x) \mid V_1(x) = c_1 r(x,a) + \beta_1 \sum_{y \in \mathbf{X}} p(y|x,a) V_1(y) \right\} . \tag{9}$$

It is easy to see that $\Gamma_1(x)$ is the set of optimal actions in state x for criterion V_1, and thus a policy is optimal for this criterion if and only if it chooses actions from the set $\Gamma_1(x)$ 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:

$$\varepsilon_1 \equiv \min_{x \in \mathbf{X}_1, a \in \mathbf{A}(x) \setminus \Gamma_1(x)} \left(V_1(x) - c_1 r(x,a) - \beta_1 \sum_{y \in \mathbf{X}} p(y|x,a) V_1(y) \right) . \tag{10}$$

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:

$$N_1 = \min \left\{ n \geq 0 \mid \varepsilon_1 > \sum_{k=2}^{\infty} \left(\frac{\beta_k}{\beta_1} \right)^n \max_{x \in \mathbf{X}} \left(V_k(x) - V_k^-(x) \right) \right\} . \tag{11}$$

Clearly, N_1 "suffers" from the transition to an infinite sum, and we need to show that it is well defined and finite.

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

Proof. Define:

$$S(n) = \sum_{k=2}^{\infty} \left(\frac{\beta_k}{\beta_1} \right)^n \max_{x \in \mathbf{X}} \left(V_k(x) - V_k^-(x) \right) . \tag{12}$$

Let $M = \max_{x \in \mathbf{X}, a \in \mathbf{A}(x)} r(x,a) - \min_{x \in \mathbf{X}, a \in \mathbf{A}(x)} r(x,a)$ denote the span semi-norm of $r(x,a)$. Using this definition, it is straightforward that

$$\forall k : \max_{x \in \mathbf{X}} \left(V_k(x) - V_k^-(x) \right) \leq |c_k| \frac{M}{1-\beta_k} \leq |c_k| \frac{M}{1-\beta_1} \tag{13}$$

and therefore

$$S(n) \leq \frac{\beta_1^{-n} M}{1-\beta_1} \sum_{k=2}^{\infty} \beta_k^n |c_k| . \tag{14}$$

Since $f(n)$ is exponentially representable, $\sum_{k=2}^{\infty} \beta_k^n |c_k| < \infty$ starting from some $N < \infty$. Therefore, for $n > N$ we may write $S(n) \leq (\beta_2/\beta_1)^{n-N} S(N) \underset{n \to \infty}{\longrightarrow} 0$. This means that there exist $\tilde{N}$ such that $\varepsilon_1 > S(\tilde{N})$, which proves N_1 is finite. □

Remark 2.2. Consider a very simple model with only one state x_0 and only two actions a_1 and a_2 for which $r(x,a_1) = 1$ and $r(x,a_2) = 0$. Then, $\max_{x \in \mathbf{X}} \left(V_k(x) - V_k^-(x) \right) = |c_k| / 1 - \beta_k$ for any k, and thus $S(n) = \beta_1^{-n} \sum_{k=2}^{\infty} \beta_k^n |c_k|$. In this model, $S(n) \underset{n \to \infty}{\longrightarrow} 0$ only if $\sum_{k=2}^{\infty} \beta_k^n c_k$ converges absolutely for some $N < \infty$, i.e. only if $f(n)$ is exponentially representable. It follows that, for a given discount function f, the bound N_1 is well-defined for *any model* if *and only if* the discount function is exponentially representable.

Having made sure that all the basic definitions of the weighted discounted theory are still meaningful, we are ready to rephrase the theory's main lemma. Using the definitions (9) and (11) of $\Gamma_1(x)$ and N_1 respectively,

Lemma 2.3. *Consider a finite Markov Decision Process with an exponentially representable discount function. If σ is an optimal Markov policy for this problem, then for every $n \geq N_1$ and every state $z \in \mathbf{X}$ such that $\mathbb{P}_x^\sigma \{x_n = z\} > 0$, we have $\sigma(z,n) \in \Gamma_1(z)$.*

Proof. Due to Lemma 2.1, the proof is essentially the same as that of Lemma 3.3 in [2], and is therefore omitted. See [1] for details. □

If the set $\Gamma_1(x)$ 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. Clearly, we may iterate the above process. For this purpose define recursively for $k > 1$, the restricted action sets in iteration k — $\mathbf{A}_k(x) = \Gamma_{k-1}(x)$, 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:

$$\Gamma_k(x) \equiv \left\{ a \in \mathbf{A}_k(x) \mid V_k^{\mathbf{A}_k}(x) = c_k r(x,a) + \beta_k \sum_{y \in \mathbf{X}} p(y|x,a) V_k^{\mathbf{A}_k}(y) \right\} \tag{15}$$

$$\mathbf{X}_k = \{x \in \mathbf{X} \mid \Gamma_k(x) \neq \mathbf{A}_k(x)\} \tag{16}$$

$$\varepsilon_k \equiv \min_{x \in \mathbf{X}_k, a \in \mathbf{A}_k(x) \setminus \Gamma_k(x)} \left(V_k^{\mathbf{A}_k}(x) - c_k r(x,a) - \beta_k \sum_{y \in \mathbf{X}} p(y|x,a) V_k^{\mathbf{A}_k}(y) \right) \tag{17}$$

$$N_k = \min \left\{ n \geq N_{k-1} \mid \varepsilon_k > \sum_{m=k+1}^{\infty} \left(\frac{\beta_m}{\beta_k} \right)^n \max_{x \in \mathbf{X}} \left(V_m^{\mathbf{A}_k}(x) - V_m^{-,\mathbf{A}_k}(x) \right) \right\} \tag{18}$$

where ε_k is taken to be ∞ in the case that $\mathbf{X}_k = \emptyset$. Similarly to N_1, N_k is well defined when $f(n)$ is exponentially representable. Using the above definitions, the following is evident:

Lemma 2.4. *Consider a finite Markov Decision Process with an exponentially representable discount function. If σ is an optimal Markov policy for this problem, then for every $k \geq 1$, $n \geq N_k$ and state $z \in \mathbf{X}$ such that $\mathbb{P}_x^\sigma \{x_n = z\} > 0$, 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 Markov Decision Process with an exponentially representable discount function, and let $S = |\mathbf{X}|$. Then for all $k \geq 2S-1$ and every $x \in \mathbf{X}$, $\Gamma_k(x) = \Gamma_{2S-1}(x)$.*

Proof. If $\Gamma_{2S-1}(x)$ 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 $\phi_i(x) \in \Gamma_{2S-1}(x)$ for all $x \in \mathbf{X}$, $i = 1, 2, ..., L$. For $\phi \in \Phi$, define $f_\phi : [0,1) \to \mathbb{R}^S$ as

$$[f_\phi(\beta)]_s = \mathbb{E}_{x^s}^\phi \sum_{n=0}^{\infty} \beta^n r(x_n, a_n) , \tag{19}$$

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

$$f_\phi(\beta) = r_\phi + \beta P_\phi f_\phi(\beta) \Rightarrow f_\phi(\beta) = (I - \beta P_\phi)^{-1} r_\phi \,. \tag{20}$$

Since P_ϕ is a stochastic matrix, by the Perron–Frobenius theorem $I - \beta P_\phi$ is invertible for $\beta \in [0,1)$ and singular for $\beta = 1$. For a square invertible matrix M, $M^{-1} = \operatorname{adj}(M) / \det(M)$. Applying this relation to (20) reveals that every entry (coordinate) of f_ϕ is a rational function of β, with numerator degree $S-1$ and denominator degree S. We also know that every entry of f_ϕ has 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 $i, j \leq L$:

$$f_{\phi_i}(\beta_k) = f_{\phi_j}(\beta_k) \;,\; \forall k = 1, 2, ..., 2S-1 \,. \tag{21}$$

Fix i and j and consider each entry of the vector equation $f_{\phi_i}(\beta) - f_{\phi_j}(\beta) = 0$ separately. We find it is a polynomial equation of degree $2S-2$ (since the common poles at $\beta = 1$ cancel). However, according to (21), this polynomial has $2S-1$ distinct roots—and is therefore identically zero. We conclude that $f_{\phi_i}(\beta) = f_{\phi_j}(\beta)$ 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. □

We are now able to prove our main result.

Proof. [of Theorem 1.5] Suppose the given Markov Decision Process has state space of size S, a finite action space, an exponentially representable discount function $f(n) = \sum_{k=1}^{\infty} c_k \beta_k^n$, and immediate reward function $r(x,a)$. Rewrite the criterion as an infinite weighted discounted criterion, with discount factors $\{\beta_k\}_{k=1}^{\infty}$ and reward functions $r_k(x,a) = c_k r(x,a)$. Compute N_{2S-1} as defined in (18). Let $\pi(x,n)$ denote an optimal Markov policy for this problem. Applying Lemma 2.4, we may assume without loss of generality that

$$\pi(z,n) \in \Gamma_{2S-1}(z) \text{ for all } z \in \mathbf{X} \text{ and } n \geq N_{2S-1} \tag{22}$$

since when $\mathbb{P}_x^\pi\{x_n = z\} = 0$ we can change π so that $\pi(z,n) \in \Gamma_{2S-1}(z)$ without changing its (optimal) value. Write the value of the MDP as

$$\begin{aligned} V(x;\sigma) &= \mathbb{E}_x^\sigma \sum_{n=0}^{\infty} f(n) r(x_n, a_n) = \mathbb{E}_x^\sigma \sum_{n=0}^{N_{2S-1}-1} f(n) r(x_n, a_n) \\ &\quad + \mathbb{E}_x^\sigma \left\{ \mathbb{E}_x^\sigma \left\{ \sum_{n=N_{2S-1}}^{\infty} f(n) r(x_n, a_n) \,|x_{N_{2S-1}} \right\} \right\} \\ &= \mathbb{E}_x^\sigma \sum_{n=0}^{N_{2S-1}-1} f(n) r(x_n, a_n) + \sum_{z \in \mathbf{X}} \mathbb{P}_x^\sigma\left(x_{N_{2S-1}} = z\right) \sum_{k=1}^{\infty} \beta_k^{N_{2S-1}} V_k\left(z; \sigma^{N_{2S-1}}\right) \end{aligned} \tag{23}$$

The expression $\sum_{k=1}^{\infty} \beta_k^{N_{2S-1}} V_k\left(z;\sigma^{N_{2S-1}}\right)$ in (23) can be optimized separately, since it depends only on $\sigma(x,n)$ for $n \geq N_{2S-1}$, while the expressions $\mathbb{E}_x^\sigma \sum_{n=0}^{N_{2S-1}-1} f(n) r(x_n, a_n)$ and $\mathbb{P}_x^\sigma\left(x_{N_{2S-1}} = z\right)$ depend only on the policy at times $n < N_{2S-1}$. On the other hand,

by Lemma 2.5, for any policy π satisfying (22), $\pi^{N_{2S-1}}(x,m) = \pi(x, n+N_{2S-1}) \in \Gamma_k(x)$ for any $x \in \mathbf{X}$, $k \geq 1$ and $m \geq 0$. Therefore $V_k\left(z; \pi^{N_{2S-1}}\right)$ is constant over all such policies, for each $z \in \mathbf{X}$, and so is $\sum_{k=1}^{\infty} \beta_k^{N_{2S-1}} V_k\left(z; \pi^{N_{2S-1}}\right)$. This means that we can choose the actions in $\pi^{N_{2S-1}}(x,n)$ arbitrarily from $\Gamma_{2S-1}(x)$, and might as well make the choice constant for all times n. This proves the existence of an N_{2S-1}-stationary optimal policy. □

Before describing a computation method for the optimal policy, there is one more issue that needs to be addressed. The computation of $\{N_k\}_{k=1}^{2S-1}$ involves evaluations of infinite sums, which are unlikely to be feasible for non-trivial models. In order 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 (14). In each iteration, the semi-norm of the reward function should be computed with respect to the restricted action set, and therefore decrease.

Finally, the computation algorithm is stated.

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 $\Gamma_k(x)$ for all $x \in \mathbf{X}$, ε_k and N_k or an appropriate upper bound.
3. If $\Gamma_k(x)$ is a singleton for every $x \in \mathbf{X}$, or $k = 2S+1$, set $N = N_k$ and continue. Otherwise set $\mathbf{A}_{k+1}(\cdot) = \Gamma_k(\cdot)$, increment k by 1 and go back to step 2.
4. Fix a stationary policy ψ, such that $\psi(x) \in \Gamma_k(x)$ for all $x \in \mathbf{X}$.
5. Compute an optimal Markov policy σ, for the N-step finite-horizon MDP, with state space, action space and transition probabilities as the original model, immediate reward function $r_n(x,a) = f(n) r(x,a)$, for $n = 0, 1, \ldots N-1$ and terminal reward
$$\mathbb{E}_{x_N}^{\psi} \sum_{n=0}^{\infty} f(n+N) r(x_{n+N}, a_{n+N}) = \sum_{k=1}^{\infty} \beta_k^N V_k(x_N; \psi) \ . \tag{24}$$
6. Output the N-stationary optimal policy
$$\pi(x,n) = \begin{cases} \sigma(x,n) & n < N \\ \psi(x) & n \geq N \end{cases} \tag{25}$$

The optimal N-stationary policy in times before N can be computed using standard Dynamic Programming methods. For more details, see [5].

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

$$V(x;\pi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \sum_{k=1}^{K} f_k(n) r_k(x_n, a_n) \ , \tag{26}$$

where for each k, $f_k(n)$ is exponentially representable with representation $f_k(n) = \sum_{i=1}^{\infty} c_{i,k} \beta_{i,k}^n$, and the additional condition

$$\beta_{i,k} > \beta_{1,k+1} \ , \ \forall i,k \tag{27}$$

for every i and every k. 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 (27) making sure they remain well defined. The N-stationary optimal policy can then be obtained 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 and limitations of exponentially representable functions

An important property of exponentially representable functions is that they behave asymptotically as exponential functions:

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

$$\lim_{n\to\infty} \beta^{-n} f(n) = c \neq 0 \text{ and } c < \infty \,. \tag{28}$$

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 exponentially representable, we have absolute convergence from some time $N < \infty$. Therefore, for $n > N$ and some $C < \infty$:

$$\beta_1^{-n} \left| \sum_{k=2}^{\infty} c_k \beta_k^n \right| \leq \beta_1^{-n} \sum_{k=2}^{\infty} |c_k| \beta_k^n < \frac{\beta_2^{n-N}}{\beta_1^n} \sum_{k=2}^{\infty} |c_k| \beta_k^N = C \left(\frac{\beta_2}{\beta_1} \right)^n \underset{n\to\infty}{\to} 0 \,. \tag{29}$$

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,

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

□

Functions with power-law form, like $(1+n^2)^{-1}$ or the hyperbolic discount function mentioned in the introduction do not satisfy the conclusion of Lemma 3.1, and are therefore not exponentially representable. 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 $g(n) \to \infty$ sub-exponentially, are also not exponentially representable for the same reason. Examples are $n\beta^n$ and $\beta^n/(1+n)$ for some $0 < \beta < 1$.

4. A cautionary note

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

Consider the function $f(n) = \beta^n h(n)$, with some $0 < \beta < 1/2$ and

$$h(n) = \begin{cases} 2 & n \bmod 6 = 0 \\ 1 & \text{otherwise} \end{cases} = [2,1,1,1,1,1,2,1,1,1,\ldots] \tag{30}$$

which is periodic with period 6. The condition of Lemma 3.1 does not hold for $f(n)$, and it is therefore not exponentially representable: it is, however, monotone decreasing.

Now consider the following (deterministic) model:

$$\mathbf{X} = \{1,2,3,4,5\}\ ,\ \mathbf{A}(1) = \{a_1, a_2\}\ ,\ \mathbf{A}(2) = \mathbf{A}(3) = \mathbf{A}(4) = \mathbf{A}(5) = \{a\} \tag{31}$$
$$p\,(2|1,a_1) = p\,(3|1,a_2) = p\,(4|3,a) = p\,(5|4,a) = p\,(1|5,a) = p\,(1|2,a) = 1$$

with the immediate reward function

$$r\,(1,a_1) = 3\ ,\ r\,(1,a_2) = 4\ , r\,(2,a) = r\,(3,a) = r\,(4,a) = r\,(5,a) = 0\ . \tag{32}$$

It can be shown [1] that stationary policies are suboptimal when $x_0 = 1$ and, moreover, there cannot be an N-stationary optimal policy.

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.

References

[1] Y. Carmon and A. Shwartz. Eventually-stationary policies for Markov decision models with non-constant discounting. *Internal Report*, Technion 2008.

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

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

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

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

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