MARKOV DECISION PROCESSES WITH GENERAL DISCOUNT FUNCTIONS

YAIR CARMON

Final Report for "Personal Topic" (044180) course in the faculty of Electrical Engineering, Technion IIT.

Supervisor: Prof. Adam Shwartz.

November 2007

Abstract. In Markov Decision Processes, the discount function determines how much the reward for each point in time adds to the value of the process, and thus deeply affects the optimal policy. Two cases of discount functions are well known and analyzed. The first is no discounting at all, which correspond to the total- and average-reward criteria. The second case is a constant discount rate, which leads to a decreasing exponential discount function. However, other discount functions appear in many models, including those of human decision-making and learning, making it interesting and possibly useful to investigate other functions.

We review results for a weighted sum of several discount functions with different cost functions, showing that finite models with this criterion have optimal policies which are stationary from a fixed time N, aptly called N-stationary. We review a proof for their existence and an algorithm for their computation, as well as remark on the structure of these policies as the discount factors vary.

We then discuss two attempts to generalize the results for weighted exponential discount functions. The first is a hypothesis for a sum of different general discount function with certain exponential bounds, in the spirit of the results for the exponential case. We show via counterexample that despite the intuitive appeal of the hypothesis, it is in fact not true, and make some remarks on why this is so.

Our second attempt at generalization is to represent a general discount function as an infinite sum of decreasing exponential functions with constant coefficients. We give convergence conditions on the sum under which the previously established results can be extended to enable us to find an optimal policy for it. We discuss two examples that clarify our results, and connect them to areas which require non-exponential discount functions. The work is concluded by an example of a model with a monotonic discount function that has no optimal N-stationary policy.

Contents

1. Introduction

1.1. **Markov Decision Processes.** A Markov Decision Process (MDP) comprises the following elements:

- A state space $\mathbf{X}$. For most purposes we will assume $\mathbf{X}$ to be countable or finite.
- An action space $\mathbf{A}$. To avoid technicalities, we will assume it to be finite, however, many basic results can be extended for an infinite action space.
- Action sets $\mathbf{A}(x)$, representing the possible actions for each state $x \in \mathbf{X}$.
- Action-dependent transition probabilities $p(y|x,a)$ for every $x, y \in \mathbf{X}$ and $a \in \mathbf{A}(x)$.
- An immediate reward function $r(x,a)$ defined over all states and their possible actions, and bounded above for each state.
- A discount function on discrete time, $f(n)$.

A policy is a rule for determining which action to use for any circumstance. Formally, we define for any time $n \in \{0,1,2,\ldots\}$ the history $h_n \in (\mathbf{A} \times \mathbf{X})^{n-1} \times \mathbf{X}$ as the collection of all previous states and actions, plus the current state. That is, $h_n = x_0 a_0 \cdots x_{n-1} a_{n-1} x_n$. The most general policy, denoted π, will then be a mapping for every history h_n to a probability measure $\pi(\cdot|h_n)$ on $\mathbf{A}(x)$. A policy in conjunction with the transition probabilities defined above generates a discrete time stochastic process $\{x_n, a_n\}_{n=0}^{\infty}$.

The final step in defining a MDP is to quantify how beneficial each policy is. One way to do it is by assigning each policy the following value function, also known as a "criterion":

$$V(x;\pi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} f(n) r(x_n, a_n) \tag{1.1}$$

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 assuming $V(x;\pi)$ is well-defined.

Let us define the maximal and minimal value of a MDP, respectively:

$$V(x) \equiv \sup_{\pi} V(x;\pi) \qquad V^{-}(x) \equiv \inf_{\pi} V(x;\pi) \tag{1.2}$$

For a given policy π, several important properties need to be defined:

- A policy is *optimal* if $V(x;\pi) = V(x)$, for all $x \in \mathbf{X}$.
- A policy is ε-*optimal* if $V(x) - V(x;\pi) \leq \varepsilon$, for all $x \in \mathbf{X}$.
- A policy is *deterministic* if it assigns a single action to every history, in which case we may write: $a_n = \pi(h_n)$.
- A *Markov* policy is a policy that depends only on the present state and the time, so we may write: $\pi(\cdot|h_n) \equiv \pi(x,n)$.
- A *stationary* policy is a Markov policy that does not depend on the time.

Note that using a Markov policy generates a Markov process, while a stationary policy generates a homogeneous Markov process.

1.2. **Discounted MDPs.** A criterion is called discounted when $f(x) = \beta^n$, for some $0 < \beta < 1$. For this criterion, the value of a reward decreases exponentially with the time it is obtained, making it a natural choice for modeling interest rates and many other effects.

It is a well known and basic result that for a discounted MDP there exists an optimal policy which is stationary and deterministic. There are several algorithms for computing this optimal policy. For more details and proof of this, see, for example, chapter 6 in [5]. The time-independence of the optimal policy may be considered a result of the "memory-less" property of exponential function.

1.3. **Motivation for different criteria.** As mentioned earlier, discounted models capture the effect constant degradation. However, many phenomena call for more complicated discounting schemes.

One example is a non-constant interest or inflation rate, to describe which we will need to change the constant discounting β^n to a more general function like $f(n)$.

Models of decision making also very often involve a non-exponential discount function. In order to model psychological effects of "greediness" in decision-making, economists use discount function where the rate of discounting increases with time, meaning that the sequence $f(n+1)/f(n)$ is increasing.

Because of the difficulty of analyzing decision processes with discount functions other than the simple exponential, most theoretical results are obtained with "toy functions", such as $f = \left[1, \delta\beta, \delta\beta^2, \delta\beta^3, \ldots\right]$ in [4]. Another class of discount functions prevalent in this context are called hyperbolic discount functions (see [2]) and are of the form: $(1+\alpha n)^{-\gamma/\alpha}$ with $\alpha, \gamma > 0$.

Another occurrence of non-exponential discounting is in models that involve learning, where the immediate costs/rewards vary with time according to a "learning curve". This usually means that early in the process, states and actions will tend to cost more (or reward less) than in the future, because the system is still maturing to full capacity in some sense.

When the penalty of learning does not depend on specific states and actions, we may write down the learning curve as a discount function, $g(n)$, that will typically vary monotonically to a nonzero limit. Common learning curves exhibit exponential ($g(n) = c_1 + c_2\beta_l^n$) or power law ($g(n) = c_1 + c_2\,(n+n_0)^{-\alpha}$) behavior, the latter being the more common (see [3]). If we also take into account constant discounting, the total discount function will be of the form $f(n) = \beta^n g(n)$.

A different cause for generalization is the case of several reward sources, which need to be discounted differently. An example for such a case may be the management of several projects, each on a different time scale. Another example is an investment portfolio with different cash flow streams.

The most natural way to model situations like these is to go from one reward function, to a sum of several reward functions, each with a different discount function. Criteria of this kind will be referred to as *general weighted criteria*, and will have the form:

$$V(x;\pi) \equiv \sum_{k=1}^{K} V_k(x;\pi)\ ,\ \text{where } V_k(x;\pi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} f_k(n) r_k(x_n, a_n) \tag{1.3}$$

When $f_k(n) = \beta_k^n$ and $1 > \beta_1 > \beta_2 > \ldots > \beta_K$, this will be referred to as the *weighted discounted criterion*. It will be discussed in Chapter 2, and is the starting point of this work.

1.4. **Embedding weighted MDPs in Discounted MDPs.** We call $f(n)$ *exponentially bounded* if there exists a $0 < \beta < 1$ and some $K \in \mathbb{R}$ such that $|f(n)| \leq K\beta^n$ for all $n \geq 0$. Equivalently and more usefully, $f(n)$ is exponentially bounded if $f(n) = \beta^n g(n)$ for some $0 < \beta < 1$ and bounded function $g(n)$.

Suppose we have a Markov Decision Process with a value function as defined in the previous section, but with all the discount functions exponentially bounded. Let us then write $f_k = \beta_k^n g_k$, and choose their order so that $1 > \beta_1 > \beta_2 > \ldots > \beta_K > 0$.

Define the following discounted Markov Decision Process, with discount factor β_1:

$$\begin{gathered}\tilde{\mathbf{X}} = \mathbf{X} \times \mathbb{N}\ ,\ \tilde{\mathbf{A}}(x) = \mathbf{A}(x)\ ,\ \tilde{p}\left((y,m)\,|\,(x,n)\,,a\right) = \delta_{m,n+1} p\left(y|x,a\right) \\ \tilde{r}\left((x,n)\,,a\right) = \textstyle\sum_{k=1}^{K} \left(\beta_k/\beta_1\right)^n g_k(n) r_k(x,a)\end{gathered} \tag{1.4}$$

Notice that the state space is still countable[1], that the action space is unchanged, that the rewards are still bounded from above, and that both processes have the same space of possible policies. Also, in the new process, the state contains information of the time. This allows us to define the new immediate reward in such a way that both processes have the same value function, if we start at $n = 0$:

$$\begin{aligned} V(x;\pi) &= \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \sum_{k=0}^{K} f_k(n) r_k(x_n, a_n) \\ &= \mathbb{E}_{(x,0)}^{\pi} \sum_{n=0}^{\infty} \beta_1^n \tilde{r}\left((x_n,n)\,,a_n\right) = \tilde{V}\left((x,0)\,,\pi\right) \end{aligned} \tag{1.5}$$

Therefore, both processes have therefore the same optimal policy, σ. According to the previous section the optimal policy for the discounted model is stationary and deterministic, that is $a_n = \sigma(x,n)$. Going back to the original space, we obtain the following result:

Theorem 1.1. *For a general weighted MDP with exponentially bounded discount functions there exist an optimal policy that is deterministic and Markov.*

In light of Theorem 1.1, all policies mentioned from now on will be assumed to be deterministic unless specifically mentioned otherwise.

[1]Were $\mathbf{X}$ not countable, but Borel measurable, then $\tilde{\mathbf{X}}$ would have been Borel measurable as well. Since the result on the existence of stationary optimal policies can be extended to such state spaces, Theorem 1.1 can also be extended.

2. Weighted discounted criteria

For the most part, this chapter will review the theory developed by Feinberg and Schwartz in [1]. This chapter and the results of the chapters following it will often refer to this paper, and cannot be considered complete without it.

2.1. Two definitions.

As mentioned in the introduction, *weighted discounted* MDPs have the following criterion:

$$V(x;\pi) \equiv \sum_{k=1}^{K} V_k(x;\pi) \text{ , where } V_k(x;\pi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \sum_{k=1}^{K} \beta_k^n r_k(x_n, a_n) \quad (2.1)$$
$$\text{and } 1 > \beta_1 > \beta_2 > \ldots > \beta_K > 0$$

Let be a π Markov policy. We call π *N-stationary* if:

$$\pi(x,n) = \pi(x,N) \quad \forall x \in \mathbf{X}, n \geq N \quad (2.2)$$

A 0-stationary policy is therefore stationary.

2.2. ε-optimal, N-stationary policies.

Theorem 2.1. *If the functions r_k are all bounded except for possibly one, then for any $\varepsilon > 0$ there exists a finite N and N-stationary, ε-optimal policy for the weighted discounted problem.*

A full proof can be found in [1], Theorem 2.4. The proof relies on the fact that we can find an $\varepsilon/4$-optimal Markov policy for the weighted discounted process, $\sigma(x,n)$, and an $\varepsilon/4$-optimal stationary policy $\phi(x)$ for the criterion V_m, where r_m is not bounded (from below). We choose N such that:

$$\frac{|r_k(x,a)|\,\beta_k^N}{1-\beta_k} \leq \frac{\varepsilon}{4(K-1)} \quad \forall x \in \mathbf{X}, a \in \mathbf{A}(x), k \neq m \quad (2.3)$$

And define the N-stationary policy:

$$\gamma(x,n) = \begin{cases} \sigma(x,n) & n < N \\ \phi(n) & n \geq N \end{cases} \quad (2.4)$$

The ε-optimality of can now be proven by evaluating $V(x;\sigma) - V(x;\gamma)$, keeping in mid that σ is $\varepsilon/4$-optimal.

Note that this proof is not constructive in the sense that it does not tell us how to compute γ (by not prescribing a computation of σ). If we assume that all the r_k are bounded, then there is a constructive result available.

Theorem 2.2. *For a general weighted MDP with exponentially bounded discount functions and bounded immediate reward functions, there exist exists a finite N and an N-stationary, ε-optimal policy, for any $\varepsilon > 0$.*

Proof. (by "brute force"). Write down the criterion as:

$$V(x;\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \sum_{k=1}^{K} \beta_k^n g(n) r_k(x_n, a_n) \text{ , with } 1 > \beta_1 > \ldots > \beta_K > 0 \tag{2.5}$$

Define $R = \sup_{k,x,a,n} |g(n) r_k(x,a)| < \infty$, and choose N such that:

$$\frac{R\beta_k^N}{1-\beta_k} \leq \frac{\varepsilon}{2K} \quad \forall k \in \{1, 2, \ldots, K\} \tag{2.6}$$

Now use Dynamic Programming[2] to find a Markov policy $\sigma(x,n)$ that is optimal for the following Finite-Horizon Markov Decision Process:

$$V_{FH}(x;\pi) \equiv \mathbb{E}_x^\pi \sum_{n=0}^{N-1} \sum_{k=1}^{K} \beta_k^n g(n) r_k(x_n, a_n) \tag{2.7}$$

Manufacturing the N-stationary, ε-optimal policy now involves simply using σ for times before N, and arbitrarily choosing a stationary policy for later times, for example:

$$\gamma(x,n) = \begin{cases} \sigma(x,n) & n < N \\ \sigma(x,N) & n \geq N \end{cases} \tag{2.8}$$

Suppose is $\pi(x,n)$ the optimal policy for the weighted discounted process, then:

$$\begin{aligned} V(x;\pi) - V(x;\gamma) = \underbrace{V_{FH}(x;\pi) - V_{FH}(x;\sigma)}_{\leq 0 \text{ from the optimality of } \sigma} + \mathbb{E}_x^\pi \sum_{n=N}^{\infty} \sum_{k=1}^{K} \beta_k^n g(n) r_k(x_n, a_n) \\ - \mathbb{E}_x^\gamma \sum_{n=N}^{\infty} \sum_{k=1}^{K} \beta_k^n g(n) r_k(x_n, a_n) \leq 2 \sum_{k=1}^{K} R\beta_k^N \leq \varepsilon \end{aligned} \tag{2.9}$$

Which proves the ε-optimality of the policy. □

Note that, asymptotically, in both Theorems 2.1 and 2.2, ε decreases exponentially with N.

[2]Dynamic Programming is a recursive method to find optimal policies for finite-horizon criteria as in eq. 2.7. For more information, see chapter 4 in [5].

2.3. **N-stationary policies for finite models.** We will now present results on the structure of optimal policies when both the state and action spaces are finite. In order to do so we need to make some more definition. The first is the conserving set:

$$\Gamma_1(x) \equiv \left\{ a \in \mathbf{A}(x) \mid V_1(x) = r_1(x,a) + \beta_1 \sum_{y \in \mathbf{X}} p\left(y|x,a\right) V_1(y) \right\} \tag{2.10}$$

Where $V_k(x), V_k^-(x)$ are the maximum and minimum values of criterion k, respectively, as defined in eq. 1.2. For every state $x \in \mathbf{X}$, $\Gamma_1(x)$ is the set of actions which may be taken in the optimal policy for criterion $V_1(x;\pi)$. This is proven in Lemma 3.1 of [1].

We now define a set of states with suboptimal actions: $\mathbf{X}_1 = \{x \in \mathbf{X} \mid \Gamma_1(x) \neq \mathbf{A}(x)\}$. If $\mathbf{X}_1 \neq \emptyset$, define:

$$\varepsilon_1 \equiv \min_{x \in \mathbf{X}_1, a \in \mathbf{A}(x) \setminus \Gamma_1(x)} \left(V_1(x) - r_1(x,a) - \beta_1 \sum_{y \in \mathbf{X}} p\left(y|x,a\right) V_1(y) \right) \tag{2.11}$$

ε_1 is the value of the smallest "mistake" one can make in the choice of a single action, in regard to criterion V_1.

If $\mathbf{X}_1 = \emptyset$ define $N_1 \equiv 0$. Otherwise define:

$$N_1 = \min \left\{ n \in \{0,1,2,\ldots\} \mid \varepsilon_1 > \sum_{k=2}^{K} \left(\frac{\beta_k}{\beta_1} \right)^n \max_{x \in \mathbf{X}} \left(V_k(x) - V_k^-(x) \right) \right\} \tag{2.12}$$

We may now write down the following:

Lemma 2.3. *Let* $\mathbf{X}$ *and* $\mathbf{A}$ *be finite. If* σ *is an optimal Markov policy for the weighted discounted problem, then for every* $n \geq N_1$, $\sigma(x,n) \in \Gamma_1(x)$.

This Lemma is proven by contradiction using the definitions above. For more details, see the proof of Lemma 3.3 in [1].

This means that after a finite period of time, any optimal policy for the weighted discounted problem is dominated by the optimal policy of the process with the slowest decreasing discount factor, in the sense that any action in the optimal policy must be also optimal for the first criterion alone.

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. If it is not a singleton, we know that after time N_1 our action sets reduce to $\mathbf{A}_2(x) \equiv \Gamma_1(x)$ and for every permissible policy, V_1 will attain its maximum value and thus be irrelevant. We may therefore solve a new weighted discounted problem, now without the first factor and with the reduced action space. Find $\Gamma_2(x), \varepsilon_2, N_2$, defined similarly to their predecessors[3] but for the new problem, and go back to the beginning of the paragraph.

[3]The definitions will require minor adjustments. For example, the definition of N_k should now read:

This process will end either when the conserving set $\Gamma_k(x)$ is a singleton for all $x \in \mathbf{X}$ at some time $k < K$, or when we reach time K. In the latter case we may choose a stationary policy arbitrarily from $\Gamma_K(x)$ for each $x \in \mathbf{X}$, so in both cases we will end up with N-stationary policies. The none-stationary part of the optimal policy may then be computed using Dynamic Programming.

Thus we have outlined the proof of a significant result:

Theorem 2.4. *If the state and action spaces are finite, then there exist an N-stationary optimal policy for the weighted discounted problem, with $N < \infty$.*

For a more formal presentation and a proof, see Theorem 3.8 in [1]. The paper also provides a more detailed version of the algorithm described (Algorithm 3.7), and discusses its computational complexity.

Note that if either $\mathbf{X}$ or $\mathbf{A}$ is infinite, ε_1 might be equal zero while $\mathbf{X}_1 \neq \emptyset$, and then $N = \infty$, nullifying the result. This explains why we must limit ourselves to the finite case.

2.4. **Notes on optimal policies in finite discounted models.** We would like to gain further insight on what happens to optimal policies of finite discounted problems, as the discount factor varies. Assume we have such problem with discount factor β, and a stationary policy ϕ. We may write:

$$V^{\beta}(x;\phi) = \mathbb{E}_x^{\pi} \sum_{n=0}^{\infty} \beta^n r\left(x_n, a_n\right) = r\left(x, \phi(x)\right) + \beta \sum_{y \in \mathbf{X}} p\left(y|x,a\right) V^{\beta}(y) \tag{2.14}$$

Defining vector and matrix notation:

$$\left(V_{\phi}(\beta)\right)_i \equiv V^{\beta}(i;\phi)\ ,\ \left(r_{\phi}\right)_i \equiv r\left(i, \phi(i)\right)\ ,\ \left(P_{\phi}\right)_{ij} \equiv p\left(j|i,\phi(i)\right) \tag{2.15}$$

We can rewrite eq. 2.14 and obtain the value function explicitly:

$$V_{\phi}(\beta) = r_{\phi} + \beta P_{\phi} V_{\phi}(\beta) \Rightarrow V_{\phi}(\beta) = \left(I - \beta P_{\phi}\right)^{-1} r_{\phi} \tag{2.16}$$

Lemma 2.5. *Given a finite discounted model, for any $0 < B < 1$, and any two deterministic stationary policies ϕ_1, ϕ_2, either $V_{\phi_1}(\beta) = V_{\phi_2}(\beta)$ only on a finite number of values of $\beta \in [0, B]$, or $V_{\phi_1}(\beta) = V_{\phi_2}(\beta)$ for every $\beta \in [0, 1)$.*

Because of possible singularities in $\beta = 1$, the number of intersections between values of different policies might approach infinity as β goes to 1. However, demanding β to lay in a closed interval contained in $[0, 1)$ solves that problem, as will be shown immediately.

$$N_k \equiv \min\left\{ n \in \{N_{k-1}, N_{k-1}+1, \ldots\} \mid \varepsilon_k > \sum_{i=2}^{K-k} \left(\frac{\tilde{\beta}_i}{\tilde{\beta}_1}\right)^n \max_{x \in \mathbf{X}} \left(\tilde{V}_i(x) - \tilde{V}_i^{-}(x)\right) \right\} \tag{2.13}$$

Where it is understood that $\tilde{\beta}_i, \tilde{V}_i, \tilde{V}_i^{-}$ refer to the new problem with $K-k$ discount factors (so that $\tilde{\beta}_i = \beta_{i+k-1}$), and appropriately restricted actions sets.

Proof. Consider the following vector-valued of functions of β:

$$f(\beta) = V_{\phi_1}(\beta) - V_{\phi_2}(\beta) = (I - \beta P_{\phi_1})^{-1} r_{\phi_1} - (I - \beta P_{\phi_2})^{-1} r_{\phi_2} \tag{2.17}$$

According to the Perron–Frobenius theorem, a stochastic matrix has no eigenvalues on $\left[B^{-1}, \infty\right)$, and therefore $f(\beta)$ is well-defined, with each element an analytic function of $\beta \in [0, B]$. It is a well known result that on a compact set, an analytic function either has a finite number of zeroes or is identically zero, which proves the lemma. □

Notice that in the case where $V_{\phi_1} = V_{\phi_2}$ for every $\beta \in [0, 1)$, both policies always yield the same results in the discounted case, making one of them somewhat redundant. Let us therefore make the following definition:

A finite discounted model for which no two stationary policies have identical value functions for every β is called *minimal*.

Theorem 2.6. *For a minimal model, and a discounted criterion with discount factor $0 < \beta \leq B < 1$, for all but a finite number values of β:*

(i) Different stationary policies have different values.

(ii) There is only one optimal stationary policy.

Proof. Since the model is finite, there is a finite number of stationary policies ($\prod_{x \in \mathbf{X}} |\mathbf{A}(x)|$). It then follows from the minimality of the model and Lemma 2.5 that the overall number of intersections between two value functions of different policies is finite. Hence the number of β's for which there is an intersection is finite, proving part (i).

When there is more then one optimal policy, the value of two different policies must be the same, and by (i) this can happen at most for a finite number of β's, proving (ii). □

Also, there is a simple sufficient (but not necessary) condition for minimality:

Condition 2.7. A finite MDP is minimal if the reward function is 1-1 valued for every $x \in \mathbf{X}$.

Proof. We prove this by contradiction. Suppose the model is not minimal, then there are two different policies, ϕ_1, ϕ_2, such that $V_{\phi_1}(\beta) = V_{\phi_2}(\beta)$ for every $\beta \in [0, 1)$. In particular, this holds for $\beta = 0$:

$$r_{\phi_1} = (I - 0 \cdot P_{\phi_1})^{-1} r_{\phi_1} = V_{\phi_1}(0) = V_{\phi_2}(0) = r_{\phi_2} \tag{2.18}$$

But since the policies are different, there must be some $x \in \mathbf{X}$ so that and $r\left(x, \phi_1(x)\right)$ $\neq r\left(x, \phi_2(x)\right)$, because of the 1-1 nature of the reward function. This contradicts our assumption that there are two different policy with the same value, and therefore the model is minimal. □

We may also find an equivalent condition to minimality, in terms of transition matrices and reward vectors:

Lemma 2.8. *A finite MDP is minimal if and only if for every two different stationary policies ϕ_1 and ϕ_2, there exist $n \geq 0$ such that $P^n_{\phi_1} r_{\phi_1} \neq P^n_{\phi_2} r_{\phi_2}$.*

Proof. Because the value functions are analytic in β (as discussed in the proof of Lemma 2.5), their series expansions are well-defined:

$$V_\phi(\beta) = (I - \beta P_\phi)^{-1} r_\phi = \sum_{n=0}^{\infty} \beta^n P^n_\phi r_\phi \tag{2.19}$$

By definition a model is minimal if for every two different stationary policies $\phi_1$and ϕ_2, $V_{\phi_1}(\beta_0) \neq V_{\phi_2}(\beta_0)$ for some $\beta_0 \in [0, 1)$. Because the functions are analytic, this is equivalent to demanding that the expansions of $V_{\phi_1}(\beta)$, $V_{\phi_2}(\beta)$ in β have at least one different coefficient - which in light of eq. 2.19 means that there exists $n \geq 0$ such that $P^n_{\phi_1} r_{\phi_1} \neq P^n_{\phi_2} r_{\phi_2}$. □

Whether it is possible to derive a simpler yet equivalent condition for minimality remains an open question.

2.5. **Application to the weighted discounted problem.** Suppose now that we have a finite model with a weighted discounted criterion of the form described in eq. 2.1. Assuming that the model is nice enough be minimal, we can apply Theorem 2.6 and observe that $\Gamma_k(x)$ is in fact a singleton for any $x \in \mathbf{X}$ and any β_1, outside a finite[4] set of "bad" value.

On first inspection this would mean that if our model happened to have such a "bad" value of β_1, so that $\Gamma_1(x)$ is not a singleton for all $x \in \mathbf{X}$, we can just replace it with $\beta_1 - \delta$, where δ is infinitesimal. Since the change is very slight, we expect to still get valid results, and now our job is much simpler since we only need to find the optimal policy of V_1 in order to have the stationary part of the optimal policy.

However, let us examine what happens to ε_1. With β_1, we had at least two optimal policies, ϕ_1, ϕ_2, and with $\beta_1 - \delta$, only ϕ_1 remained optimal. However, since δ is infinitesimal, $V_{\phi_1}(\beta - \delta) - V_{\phi_2}(\beta - \delta)$ is infinitesimal as well, and by definition $\varepsilon_1 \leq V_{\phi_1}(\beta - \delta) - V_{\phi_2}(\beta - \delta)$. This means that as δ goes to zero, ε_1 goes to zero as well, and N_1 tends to infinity.

In conclusion, for many weighted discounted models we may be able to avoid iteration of the optimal policy computation algorithm by "infinitesimally" changing β_1. However, the price of this infinitesimal modification will be an "infinite" increase in the N of the N-stationary optimal policy, corresponding to an "infinite" increase in the computational effort.

[4]Assuming the relevant values of β_1 are bounded from above by some arbitrary $\beta_0 < 1$.

3. General weighted criteria

3.1. **An intuitive hypothesis.** Consider a finite Markov Decision process with criterion:

$$V(x;\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \sum_{k=1}^{K} f_k(n) r_k(x_n, a_n) \tag{3.1}$$

Also, assume that the functions have the following bounds:

$$\bar{\beta}_k^n \geq f_k(n) \geq \underline{\beta}_k^n \text{ and, } 1 > \bar{\beta}_1 \geq \underline{\beta}_1 > \bar{\beta}_2 \geq \dots \tag{3.2}$$

Note that this is a stronger condition than the previously assumed exponential bounds, and we may write $f_k(n) = \bar{\beta}_k^n g_k(n)$, with $0 < g(n) \leq 1$ for every n.

Since each of thef_k is strictly larger than the functions which succeed it, and all the functions are exponentially bounded, it is natural to expect that like in the weighted exponential case, the (not necessarily stationary) optimal policy of f_1 will dominate the optimal policy of the entire process in the long run. Next in order of priority should come the optimal policy for f_2 in the restricted action space, and so on.

In order to formalize this hypothesis, we need to define an equivalent of the conserving set in the case where the optimal policy is not stationary. First we define the value function of the first criterion, shifted by time N:

$$V_1^N(x;\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} f_1(n+N) r_1(x_n, a_n) \tag{3.3}$$

And the appropriate shifted optimal value:

$$V_1^N(x) = \sup_\pi \mathbb{E}_x^\pi \sum_{n=0}^{\infty} f_1(n+N) r_1(x_n, a_n) \tag{3.4}$$

Notice that for $f_1(n) = \beta^n$ we have: $V_1^N(x) = \beta^N V_1(x)$. Also notice that there exists a (Markov) policy $\pi(x,n)$ such that $\pi(x, n+N)$ achieves $V_1^N(x)$ for any $N \geq 0$. Such policy is called *persistently optimal*. For more information, see **?**].

We define a time-dependent conserving set:

$$\Gamma_1(x,n) \equiv \left\{ a \in \mathbf{A}(x) \mid V_1^n(x) = r_1(x,a) + \sum_{y \in \mathbf{X}} p(y|x,a) V_1^{n+1}(y) \right\} \tag{3.5}$$

So similarly to the discounted case, the set $\Gamma_1(x,n)$ is the set of permissible actions for the optimal policy of criterion V_1, at time n. If $f_1(n) = \beta^n$ then $\Gamma_1(x,n) = \Gamma_1(x)$ for every $x \in \mathbf{X}$ and every time n.

We may now write down an equivalent to Lemma 2.3:

Conjecture 3.1. *Let σ be an optimal Markov policy for a finite model with a general weighted criterion, with discount functions that obey the bounds in eq. 3.2. There exist $N < \infty$ such $\sigma(x,n) \in \Gamma_1(x,n)$ that for every $x \in \mathbf{X}$, $n \geq N$.*

However, this hypothesis turns out to be false, as shall be demonstrated immediately.

3.2. **A counterexample.** Consider the following model:

$$\begin{gathered} \mathbf{X} = \{d,u\} \;\; \mathbf{A}(d) = \{s,m\} \;\; \mathbf{A}(u) = \{s\} \\ p\left(d|d,s\right) = p\left(u|d,m\right) = p\left(u|u,s\right) = 1 \end{gathered} \tag{3.6}$$

Since the is no choice of action in state u, the model allows only two stationary policies:

π_s	when *d*own, *s*tay there.
π_m	when *d*own, *m*ove *u*p.

We define the immediate reward functions r_1 and r_2:

$$\begin{array}{lll} r_1\left(d,s\right) = 6 & r_1\left(d,m\right) = 0 & r_1\left(u,s\right) = 8 \\ r_2\left(d,s\right) = 0 & r_2\left(d,m\right) = 0 & r_2\left(u,s\right) = 100 \end{array} \tag{3.7}$$

And the discount functions:

$$f_1(n) = \left(0.75^n + 0.25^n\right)/2 \;,\; f_2(n) = 0.25^n \tag{3.8}$$

We have: $0.75^n \geq f_1(n) \geq 0.5^n > 0.25^n \geq f_2(n) \geq 0.25^n$, so the functions obey the bounds specified in eq. 3.2. For the general weighted criterion, let us find $\Gamma_1(d,n)$ as defined in eq. 3.5. V_1 is clearly a weighted discounted criterion, and therefore must have an N-stationary optimal policy. This means that $\Gamma_1(d,n) = \Gamma_1(d,N)$ for all $n \geq N$ and some $N < \infty$. Moreover, we know how to find $\Gamma_1(d,N)$, Denote:

$$V_{1,1}(d;\pi) = \mathbb{E}_d^\pi \sum_{n=0}^{\infty} 0.75^n r_1\left(x_n,a_n\right) \;,\; V_{1,2}(d;\pi) = \mathbb{E}_d^\pi \sum_{n=0}^{\infty} 0.25^n r_1\left(x_n,a_n\right)$$

The numbers were rigged so $V_{1,1}(d;\pi_s) = V_{1,1}(d;\pi_m) = 12$, and therefore the first discount factor adds no restrictions to the stationary part of the optimal policy of V_1. However, $V_{1,2}(d;\pi_s) = 4 > 4/3 = V_{1,2}(d;\pi_m)$, which means that $\Gamma_1(d,n) = \{s\}$ for sufficiently large n. According to our hypothesis, for even more sufficiently large n, the optimal policy will have to have its actions belong to $\Gamma_1(d,n)$, which happens to be a singleton.

Therefore, if the hypothesis is correct, there must be an N-stationary optimal policy for this model and the general weighted criterion, with π_s as its stationary part. Suspiciously enough, this result does not depend at all on our choice of r_2.

Let us rewrite our criterion:

$$V(x;\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \sum_{k=1}^{K} f_k(n) r_k\left(x_n, a_n\right) \tag{3.9}$$
$$= \mathbb{E}_x^\pi \sum_{n=0}^{\infty} 0.75^n \frac{r_1\left(x_n, a_n\right)}{2} + \mathbb{E}_x^\pi \sum_{n=0}^{\infty} 0.25^n \left(r_2(x_n, a_n) + \frac{r_1\left(x_n, a_n\right)}{2} \right)$$

Evidently in this case the problem may be viewed as discounted itself. Again the first discount factor does not affect the stationary part of the optimal policy, but now the policy π_m is clearly preferable for the second factor. We therefore conclude this model does indeed have an optimal N-stationary policy, but that the stationary part of this policy is π_m. This contradicts the hypothesis and thus proves it wrong.

3.3. **Discussion.** Our counterexample shows that general weighted models with the bounds in eq. 3.2 fail to capture the essential "hierarchical" property of the weighted discounted models. In the generalized case, a discount function may contain parts that act as "tie-breakers", in that they create variations in the values of what would otherwise be identically-valued policies. When we have "tie-breakers" that decay faster, or as fast as the subsequent discount functions, the overly fine-tuned optimal policy for one discount function might not dominate when the rest of the functions are taken into account.

It is possible to find conditions on the discount functions that actually work, at the cost of their elegance. Suppose we have a general weighted criterion with the current bounds, and in analogy to the weighted discounted case define: $\mathbf{X}_1(n) = \{x \in \mathbf{X} \mid \Gamma_1(x,n) \neq \mathbf{A}(x)\}$ Assuming the non-degenerate case where $\mathbf{X}_1(n) \neq \emptyset$, define[5]:

$$\varepsilon_1(n) = \underline{\beta}_1^{-n} \min_{x \in \mathbf{X}_1, a \in \mathbf{A}(x) \setminus \Gamma_1(x)} \left(V_1^n(x) - r_1(x,a) - \sum_{y \in \mathbf{X}} p\left(y|x,a\right) V_1^{n+1}(y) \right) \tag{3.10}$$

$$N_1(n) = \min \left\{ N \in \{0,1,\ldots\} \mid \varepsilon_1(n) > \sum_{k=2}^{K} \frac{f_k(N)}{\underline{\beta}_1^N} \max_{x \in \mathbf{X}} \left(V_k(x) - V_k^-(x) \right) \right\} \tag{3.11}$$

$$N_1 = \max_n N_1(n) \tag{3.12}$$

It is straight-forward to see that N_1 is well-defined and finite when $\varepsilon_1(n) \geq \varepsilon_1$for some $\varepsilon_1 > 0$ and all n.[6] Under this condition, one can follow through the steps the proof of Lemma 3.3 in [1] and arrive at a proof of the hypothesis, with the slightly unseemly condition added.

[5]When $\mathbf{X}_1(n) = \emptyset$, define $\varepsilon_1(n) = \infty$.

[6]In the counterexample it is possible to show that $\varepsilon_1(n)$ behaves asymptotically as 0.25^n, and thus $N_1 = \infty$ for this model and criterion.

4. Results for a single generalized discount function

4.1. **Representation as an infinite sum of exponential functions.** Consider model with finite $\mathbf{X}$ and $\mathbf{A}$, and a weighted discounted criterion with $r_k(x,a) = c_k r(x,a)$, where the c_k are some real constants. Also, let the number of discount factors go to infinity, yielding the following criterion:

$$V(x;\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} f(n) r(x_n, a_n) \text{ with, } f(n) = \sum_{k=1}^{\infty} c_k \beta_k^n \text{ and, } \beta_1 > \beta_2 > \cdots \quad (4.1)$$

This is a MDP with a generalized (single) discount function criterion, which is well defined as long as the series $\sum_{k=1}^{\infty} c_k \beta_k^n$ converges for each n. Usually it also makes sense to normalize the discount function so that $f(0) = \sum_{k=1}^{\infty} c_k = 1$.

While it is possible to show, using function-theoretical arguments, that any time series $f(n)$ may be represented in this fashion, useful results will add stronger convergence condition on the defining sum.

4.2. **Application of results for weighted discounted criteria.** Going back to the weighted discounted representation, we would like to use the results of chapter 2 to characterize the optimal policy for this criterion. As in chapter 3, the first step would be to generalize Lemma 2.3.

Since they depend only on the first discount factors, the definitions of $\Gamma_1(x)$ and ε_1 remain unchanged. The definition of N_1 should now be:

$$N_1 = \min\left\{ n \in \{0,1,\ldots\} \mid \varepsilon_1 > \sum_{k=2}^{\infty} \left(\frac{\beta_k}{\beta_1}\right)^n \max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right) \right\} \quad (4.2)$$

Define:

$$S(n) \equiv \sum_{k=2}^{\infty} \left(\frac{\beta_k}{\beta_1}\right)^n \max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right) \quad (4.3)$$

Since $\max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right) \geq 0$ for all k, $S(n) = \infty$ or is finite for every n, and is thus well defined. N_1 is well defined if $S(N) \underset{N\to\infty}{\longrightarrow} 0$.

Lemma 4.1. *Either $S(n) = \infty$ for every $n \geq 0$ or $S(n) \underset{n\to\infty}{\longrightarrow} 0$.*

Proof. The first condition clearly excludes the the latter. Conversely, if $S(N_0) < \infty$ for some $N_0 < \infty$, then for a given $\varepsilon > 0$ we can choose K so that:

$$S_{K,\infty}(N_0) \equiv \sum_{k=K}^{\infty} \left(\frac{\beta_k}{\beta_1}\right)^{N_0} \max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right) < \frac{\varepsilon}{2} \quad (4.4)$$

Denote $S_K(n) \equiv S(n) - S_{K,\infty}(n)$. Since $S_K(n)$ is a finite sum of exponentially decreasing functions, $S_K(n) \underset{n\to\infty}{\longrightarrow} 0$ and we may choose N so that $S_K(N) < \varepsilon/2$. Since all the summands in $S(n)$ are positive, $S_{K,\infty}(N) \leq S_{K,\infty}(N_0)$ for $N \geq N_0$. Putting it all together, we get:

$$S(N) = S_K(N) + S_{K,\infty}(N) \leq S_K(N) + S_{K,\infty}(N_0) < \varepsilon$$

This proves that $S(n) \underset{n\to\infty}{\longrightarrow} 0$. □

Denote:

$$R^+ = \max_{x\in\mathbf{X}, a\in\mathbf{A}(x)} r(x,a)\ ,\ R^- = \min_{x\in\mathbf{X}, a\in\mathbf{A}(x)} r(x,a) \tag{4.5}$$

For $c_k > 0$, for each x and k we have: $V_k(x) \leq R^+/(1-\beta_k)$ and $V_k(x) \geq R^-/(1-\beta_k)$, and by applying similar considerations to the case $c_k < 0$, we conclude that:

$$\forall k\ :\ \max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right) \leq |c_k| \frac{R^+ - R^-}{1-\beta_k} \leq |c_k| \frac{R^+ - R^-}{1-\beta_1} \tag{4.6}$$

And therefore:

$$S(n) \leq \frac{\beta_1^{-n}}{1-\beta_1} \sum_{k=2}^{\infty} \beta_k^n |c_k| \tag{4.7}$$

Let us call a function $f : \{0,1,\ldots\} \to \mathbb{R}$ *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, 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 $N < \infty$.

From Lemma 4.1 and the above considerations it follows that N_1 is well-defined if $f(n)$ is exponentially representable. Moreover, we may easily find a model so that for any discounted criterion the difference between the maximal and minimal attainable values is 1, and therefore for any k, $|c_k| = \max_{x\in\mathbf{X}} \left(V_k(x) - V_k^-(x)\right)$. For such a model, $S(N) = \beta_1^{-N} \sum_{k=2}^{\infty} \beta_k^N |c_k|$ and we may therefore say that for a given discount function, N_1 is well-defined for *any model*, if *and only if* the function is exponentially representable.

When N_1 is well defined, we can go on to prove Lemma 2.3 for the infinite case, exactly like in the finite case. The resulting generalized Lemma 2.3 should read:

Lemma 4.2. *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 $n \geq N_1$,$\sigma(x,n) \in \Gamma_1(x)$, with as $\Gamma_1(x)$ defined in eq. 2.10, and N_1 as defined in eq. 4.2.*

It is clear that any exponentially representable function is exponentially bounded. However, the condition turns out to be considerably more restrictive. This will be further discussed in the Example 4.2.

The next natural step will be to iterate this result for the rest of the discount factors and to end up with an algorithm for the finding of the optimal policy of somewhat general criteria. However, here there is another pitfall, since the algorithm will not halt unless there is some $K < \infty$ so that the (appropriately restricted) conserving set $\Gamma_1(x)$ will be a singleton for all $x \in \mathbf{X}$.

However, the notion of minimality (defined in section 2.4) and our conclusions on the structure of conserving sets for minimal models will enable us to write:

Theorem 4.3. *Consider a finite and minimal 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 the algorithm in section 2.3, which is guaranteed to halt.*

Proof. Lemma 4.2 ensures us that the calculations in the algorithm are meaningful. In the k^{th} iteration of the algorithm, the restricted conserving set $\Gamma_k(x)$ is not a singleton for every $x \in \mathbf{X}$ only if there are two different stationary policies that attain the same value for the criterion $V(x;\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \beta_k^n r(x_n, a_n)$. According to Theorem 2.6, this can happen only for a finite number of β_k's. Therefore, at some time $K < \infty$ the conserving set must become a singleton and the algorithm halts, providing an N_K-stationary optimal policy. □

4.3. **Examples.** We now review three examples that will clarify the results of this chapter.

Example 4.1. *Decision making.*

As mentioned in the introduction, discount functions for which $f(n+1)/f(n)$ is increasing are useful in models of decision making. We give an example of a function of the required type, that submits to the conditions of Theorem 4.3. For some $0 < \beta_0 < 1$, choose $\beta_k = \beta_0^k$ and $c_k = 1/(e-1)\,k!$. Then:

$$f(n) = \frac{1}{e-1} \sum_{k=1}^{\infty} \frac{1}{k!} \left(\beta_0^k\right)^n = \frac{e^{\beta_0^n} - 1}{e-1} \tag{4.8}$$

For this function $f(n+1)/f(n)$ is increasing. Perhaps the simplest way to see it is to confirm that $[\log f(x)]'' > 0$ for every $x > 0$.[7] Also, since the c_k are positive, Theorem 4.3 may be applied, with possibly interesting results.

Note that the function is exponentially bounded ($(e-1)^{-1}\beta_0^n \le f(n) \le \beta_0^n$)[8], as is expected. Also, note that the hyperbolic functions mentioned in the introduction

[7] This is a simple exercise in calculus: $[f(x+1)/f(x)]' > 0 \Leftrightarrow f'(x+1)/f(x+1) > f'(x)/f(x)$ $\Leftrightarrow [\log f(x+1)]' > [\log f(x)]' \Leftrightarrow [\log f(x)]'' > 0$.

[8] The lower bound is derived from $e^{\beta_0^n} \le 1 + \beta_0^n$, which is easily obtained by considering the Taylor series of e^x. For the upper bound, notice that $f(0) = \beta_0{}^0 = 1$ and that $f'(x) < (\beta_0^x)'$, for every $x > 0$.

are not exponentially bounded, and we will therefore not be able to analyze them using our results.

Example 4.2. *Learning curves*

We saw in the introduction that a criterion with constant discounting and learning gives a discount function of the form $f(n) = \beta^n g(n)$, where $g(n)$ is the learning curve. It follows that a simple exponential learning curve can be analyzed as an ordinary weighted discounted criterion. Moreover, any bounded function of the form $g(n) = \beta^{-n} f(n)$, with f exponentially representable can be shown to behave asymptotically as $c_1 + c_2\beta_l^n$, where c_1, c_2 are real constants and $0 < \beta_l < 1$. n

This unfortunately means we will not be able to analyze the more common case of power law learning. However, we are still able to generate a variety of possible learning curves. For example, let $c_k = (1-p)\,p^{k-1}$ and $\beta_k = \beta\frac{1+b^k}{1+b}$ with $0 < p, \beta < 1$. Then:

$$\begin{aligned} f(n) &= (1-p)\,\beta^n(1+b)^{-n}\sum_{k=1}^{\infty} p^{k-1}\left(1+b^k\right)^n \\ &= \beta^n(1+b)^{-n}\sum_{i=0}^{n}\left(\begin{array}{c} n \\ i \end{array}\right)\sum_{k=1}^{\infty}(1-p)\,p^{k-1}\left(b^i\right)^k \\ &= \beta^n(1+b)^{-n}\sum_{i=0}^{n}\left(\begin{array}{c} n \\ i \end{array}\right)\frac{1-p}{b^{-i}-p} \end{aligned} \tag{4.9}$$

And so:

$$g(n) = (1+b)^{-n}\sum_{i=0}^{n}\left(\begin{array}{c} n \\ i \end{array}\right)\frac{1-p}{b^{-i}-p} \tag{4.10}$$

This leaning curve decreases monotonically from 1 to $1-p$. For small values of b it does so almost exactly like $(1-p)+p(1+b)^{-n}$, but for larger values the functions are distinct.

Example 4.3. *A monotonic discount function with no N-stationary optimal policy.*

So far all the functions we discussed had an N-stationary optimal policy. It is reassuring to know that there are functions for which, under some models, the optimal policy will have no stationary part. Moreover, when a discount function decreases monotonically it seems natural for it to produce a behavior that is monotonic, or stationary, in some sense. However, this intuition is not true, and we provide an example of such a case.

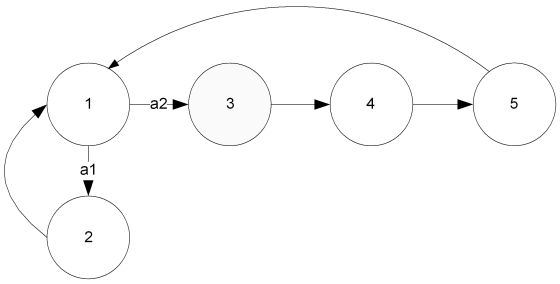
Consider the function $f(n) = \beta^n h(n)$, with some $0 < \beta < 1$ and:

$$h(n) = \begin{cases} 2 & n \bmod 6 = 0 \\ 1 & \text{otherwise} \end{cases} = [2,1,1,1,1,1,2,1,1,1,\ldots] \tag{4.11}$$

$h(n)$ is periodic with period 6. From our discussion in the previous example we already know that $f(n)$ is not exponentially representable, because $h(n)$ does not have asymptotic exponential behavior.

Now consider the following (deterministic) model:

$$\mathbf{X} = \{1,2,3,4,5\} \ , \ \mathbf{A}(1) = \{a_1, a_2\} \ , \ \mathbf{A}(2) = \mathbf{A}(3) = \mathbf{A}(4) = \mathbf{A}(5) = \{a\} \quad (4.12)$$
$$p\left(2|1,a_1\right) = p\left(3|1,a_2\right) = p\left(4|3,a\right) = p\left(5|4,a\right) = p\left(1|5,a\right) = p\left(1|2,a\right) = 1$$



An illustration of the state space and possible transitions.

With the immediate reward function:

$$r\left(1,a_1\right) = 3 \ , \ r\left(1,a_2\right) = 4 \ , r\left(2,a\right) = r\left(3,a\right) = r\left(4,a\right) = r\left(5,a\right) = 0 \qquad (4.13)$$

Suppose this MDP has an N-stationary optimal policy, $\sigma(1,n)$ (the rest of the states do not require decisions). Without loss of generality we may assume $\sigma$is deterministic[9], and therefore that there exists a time $M_0 \geq N$ such that $x_{M_0} = 1$ w.p. 1, because of the deterministic nature of the model. Also, $x_{M_0+4} = x_{M_0+8} = 1$ w.p. 1, since at those times the policy is stationary and must repeatedly use only one of the actions. For $x_0 = 1$ 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 $x_M = 1$ w.p. 1 and $h(n+M) = h(n)$ for every $n \geq 0$.

Let $\sigma^M(1) \equiv \sigma(1, n+M)$, a stationary policy. We may write:

$$V(1;\sigma^N) = \mathbb{E}_1^{\sigma^M} \sum_{n=0}^{\infty} \beta^n h(n) r\left(x_n, a_n\right) = \beta^{-M} \mathbb{E}_1^{\sigma} \sum_{n=M}^{\infty} \beta^n h(n) r\left(x_n, a_n\right) \qquad (4.14)$$

By the optimality principle, $\sigma$is optimal for the criterion on the right hand side, which means that the policy σ^M is optimal for the original criterion. This shows that if this problem has an N-stationary optimal policy, it also has a stationary optimal policy.

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:

[9]Given a randomized N-stationary optimal policy, it is possible to show that there is also a deterministic N-stationary optimal policy.

$$\pi(1,n) = \begin{cases} a_2 & n \bmod 6 = 0 \\ a_1 & n \bmod 6 = 4 \end{cases} \tag{4.15}$$

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

$$V(1;\pi) = \frac{8+3\beta^4}{1-\beta^6} \approx 8.19 \tag{4.16}$$

$$V(1;\sigma_1) = \frac{6+3\beta^2+3\beta^4}{1-\beta^6} \approx 6.79 \tag{4.17}$$

$$V(1;\sigma_2) = \frac{8+4\beta^4+4\beta^8}{1-\beta^{12}} \approx 8.17 \tag{4.18}$$

This shows that for this problem, both stationary policies are suboptimal. Thus the existence of an N-stationary optimal policy is prohibited, since our considerations have shown that it will result in an optimal stationary policy. Since we chose $\beta < 1/2$, $h(n) > \beta h(n+1)$ for every n. The discount function is therefore monotonically decreasing, and for some models does not have an N-optimal policy, as required.

Acknowledgement. I would like to thank Prof. Adam Shwartz for his guidance and his time spent introducing me to MDPs, as well as our many interesting discussions.

References

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

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

[3] F.E. Ritter and L.J. Schooler. The learning curve. In *International encyclopedia of the social and behavioral sciences*. Pergamon, Amsterdam, 2001.

[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.