# Dynamic Programming for Variable Discounted Markov Decision Problems

Eugenio Della Vecchia ♠, Silvia Di Marco ♠, and Fernando Vidal ♣

♠ Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Facultad de Ciencias Exactas, Ingeniera y Agrimensura - Universidad Nacional de Rosario (FCEIA-UNR), Argentina.

♣ FCEIA - UNR, Argentina.

{eugenio,fvidal}@fceia.unr.edu.ar

**Abstract.** We study the existence of optimal strategies and value function of non stationary Markov Decision Processes under variable discounted criteria, when the action space is assumed to be Borel and the action space to be compact.
With this new way of defining the value of a policy, we show existence of Markov deterministic optimal policies in the finite-horizon case, and a recursive method to obtain such ones. For the infinite horizon problem we characterize the value function and show existence of stationary deterministic policies.
The approach presented is based on the use of adequate dynamic programming operators.

**Keywords:** Markov Decision Processes, Variable Discount Factor, Dynamic Programming

*In memoriam Silvia Di Marco*
*1964-2014*

## 1 Introduction

In this work we analyse the existence of optimal strategies and value functions for discrete-time stochastic control models, Markov Decision Processes (**MDP**) under two related discounted criteria, in which the discount factors varies stage to stage.

**MDP** theory have been widely developed during the last sixty years, and the advances and applications have been synthesized in well-known books in the area. In [4, 8], the general state and action spaces case is studied, while in [2, 5, 7] the spaces are assumed to be finite. Besides in [7, 8], several examples, under their respective assumptions, can be found.

A common feature in the treatment presented through these references is the assumption that the discount factor remains constant at the different stages.

This fact significantly simplifies the analysis. In order to obtain a solution, one can directly apply the Banach Contraction Principle, if the immediate return function is bounded.

Criteria with variable discount factors are more realistic from the behavioural viewpoint of the decision makers. Heals [3] provides a complete economic treatment of the subject, analysing different ways of contemplate future reward and costs, providing social and psychological motivations. Examples of control problems giving economic and environmental interests are presented in [1, 9].

The basic idea behind Intertemporal Choice is the assumption that, between present and future consumptions, agents show preferences (or assigns greater ratings) for present ones, and this tendency empirically shows to be decreasing in time. In a similar way, companies assign greater preferences for present rewards in front of feature ones, with a decreasing assessment in time. See, for instance, [6, Chapter 20]

Although the range of results is large, it appears that the literature does not cover the case that is the subject of this paper.

The objective of the present work is to provide a formal framework to variable discounted **MDP**. We assume Borel state and compact action spaces.

This paper is organized as follows. In **Section 2**, we present the **MDP** model, introduce its notation and state the assumptions on the data of the problem. In **Section 3** we present the performance criteria.

The results obtained are presented in **Sections 4** and **5**, for finite-horizon and infinite horizon problems, respectively.

Finally, **Section 6** is devoted to the concluding remarks.

## 2 Preliminaries and Notations

We consider a Markov decision model of the form

$$\mathcal{M} := (\mathcal{S}, \mathcal{A}, \{\mathcal{A}_s : s \in \mathcal{S}\}, \{Q_t\}, \{r_t\}, \{\lambda_t\})$$

where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space. For each $s \in \mathcal{S}$, we define the set $\mathcal{A}_s$ as the set of actions available in state $s$. In such a way $\mathcal{A} = \bigcup_{s \in S} \mathcal{A}_s$.

We put $\mathbb{K} = \{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$. The transition laws $Q_t$ are stochastic kernels on $\mathcal{S}$ given $\mathbb{K}$ and the reward functions $r_t$ are real-valued on $\mathbb{K}$.

We propose a sequence of discount factors $\{\lambda_t\}_t$, to be applied at the different decision epochs. We shall assume no discount factor is applied at the moment when the first decision has to be taken.

If at the time of the $t$-th decision epoch the state of the system is $s_t = s$ and the chosen action is $a_t = a \in \mathcal{A}_s$, an instantaneous reward $r_t^{a_t}(s_t)$ is received, and the system move to a new state $s_{t+1}$ according the probability distribution $Q_t^{a_t}(s_{t+1}|s_t)$.

We will note, for Borel sets $X$ and $Y$, with $\mathbb{P}(X)$ to the family of probability measures on $X$ endowed with the weak topology, and with $\mathbb{P}(X|Y)$ to the family of transition probabilities from $Y$ to $X$.

An admissible history of the process up to the $t$-th decision epoch is a sequence consisting of states $s_k$ and actions $a_k \in \mathcal{A}_{s_k}$, with $k = 0, ..., t-1$, and a final state $s_t$. That is, an element of the form $h_t = (s_0, a_0, ..., s_{t-1}, a_{t-1}, s_t)$. We define the spaces of admissible histories up to stage $t$ by $\mathcal{H}_t$.

A Markov policy is a sequence $\pi = \{\pi_t\}$ of stochastic kernels $\pi_t \in \mathbb{P}(\mathcal{A}|\mathcal{H}_t)$ such that for every $h_t \in \mathcal{H}_t$ and $t \in \mathbb{N}$, $\pi_t(\mathcal{A}_{s_t}|h_t) = \pi_t(\mathcal{A}_{s_t}|s_t)$, where for Borel subsets of $\mathcal{A}_s$, $\pi_t(B|s_t)$ represents the probability of choose an action on $B$, at time $t$ and state $s_t$. We shall say that a distribution $\pi_t$ is deterministic if there exists $a \in \mathcal{A}_{s_t}$, such that $\pi_t(\cdot|s_t) = \delta_a(\cdot)$ (i.e., it assigns probability 1 to action $a$), and we will note $\pi_t = f_t$. A pure Markov policy is a Markov policy $\pi = \{f_t\}$ formed by deterministic distributions.

A Markov policy $\pi = \{\pi_t\}$ is stationary when there exists $f \in \mathbb{P}(\mathcal{A}|\mathcal{S})$ (deterministic or not), such that $f(s) \in \mathbb{P}(\mathcal{A}_s)$ and $\pi_t = f$ for all $s \in \mathcal{S}$ and $t \in \mathbb{N}$. In this case, we identify $\pi$ with $f$, i.e., $\pi = \{f, f, ...\}$. If $f$ is in particular, for any state, concentrated in some action, $f$ is a pure stationary policy.

We denote by $\Pi$ the set of all Markov policies and by $\Pi_{\mathrm{stat}}$ the set of all stationary policies.

For each strategy $\pi \in \Pi$, and any initial state $s$, there exists a unique probability measure $\mathrm{P}_s^\pi$ and stochastic processes $\{S_t\}$ and $\{A_t\}$, where $S_t$ and $A_t$ represent the state and the action at the $t$-th decision epoch. See [4, Appendix C, Proposition C.10]

$\mathbb{E}_s^\pi$ denotes the expectation operator with respect $\mathrm{P}_s^\pi$.

For any given function $h : \mathbb{K} \to \mathbb{R}$ and any $\xi \in \mathbb{P}(\mathcal{A}_s)$ we will write $h^\xi(s)$ instead of $h^{\xi(s)}(s)$, and it will be

$$h^\xi(s) = \int_{\mathcal{A}_s} h^a(s)\xi(da)$$

whenever the integral is well defined.

## 3 Performance Criteria

Through this work, in order to evaluate the performance of policies, we use a total variable discounted criterion.

We assume a discount factor $\lambda_t$ at the $(t-1)$-th decision epoch. More precisely, for $N \geqq 1$, $s \in \mathcal{S}$ and $\pi \in \Pi$, we will evaluate

$$V_N^\pi(s) := \mathbb{E}_s^\pi \left[ r_0^{A_0}(s) + \sum_{t=1}^{N-1} \lambda_{t-1} r_t^{A_t}(S_t) + \lambda_{N-1} r_N(S_N) \right] .$$

In the rest of this work we shall use the convention that variables with negative indices, when multiplying, will be taken as 1. With this,

$$V_N^\pi(s) = \mathbb{E}_s^\pi \left[ \sum_{t=0}^{N-1} \lambda_{t-1} r_t^{A_t}(S_t) + \lambda_{N-1} r_N(S_N) \right] \ .$$

The infinite horizon performance will be analysed through

$$V^\pi(s) := \mathbb{E}_s^\pi \left[ \sum_{t=0}^{\infty} \lambda_{t-1} r_t^{A_t}(S_t) \right] \ .$$

The objective of the controller, in the infinite horizon problem, is to find (when it exists) a policy that solves, given the current state $s$:

$$\pi(s) \in \arg\max_\pi V^\pi(s) \ .$$

Such a strategy $\pi^* \in \Pi$ is said to be optimal, and the function

$$V^*(s) = \sup_{\pi \in \Pi} V^\pi(s)$$

is the optimal value function. Likewise for the finite-horizon problems, noting

$$V_N^*(s) = \sup_{\pi \in \Pi} V_N^\pi(s) \ .$$

## 4 The Finite-Horizon Problem

In the remains of the work we consider the next general assumption.

**Assumption 1**

(a) *The state space $\mathcal{S}$ is a Borel subset of a complete and separable metric space.*
(b) *For each $s \in \mathcal{S}$, the set $\mathcal{A}_s$ is compact.*
(c) *For $s \in \mathcal{S}$, and $t = 0, 1, ..., N-1$, $r_t^{\cdot}(s)$, is upper semicontinuous on $\mathcal{A}_s$.*
(d) *$|r_t^a(s)| \leqq M_t$ and $|r_N(s)| \leqq M_N$, for any $s \in \mathcal{S}$, $a \in \mathcal{A}_s$ and $t = 0, 1, ..., N-1$.*
(e) *For $(s, a) \in \mathbb{K}$ and each bounded measurable function $v$ defined on $\mathcal{S}$, the application $a \mapsto \int v(z) Q_n^a(dz|s)$ is continuous on $\mathcal{A}_s$, for any $n \in \mathbb{N}$.*

For finite state and action spaces, **Assumption 1** trivially holds.

**Remark 1** *Our idea to tackle the problem with variable discount factors is inspired in the dynamic programming approach on constant discounted models, which results a particular case of our, taking $\lambda_\tau = (\alpha)^\tau$.*

*Indeed, the finite-horizon problem can be solved by repeated application of operators of the form (defined where it should be)*

$$(Tv)(s) = \sup_{a \in \mathcal{A}_s} \left\{ r_t^a(s) + \alpha \int_{\mathcal{S}} v(z) Q_t^a(dz|s) \right\} \ ,$$

*and noting that the factor $\alpha$ results in the quotients of two successive discounts in the sequence, (i.e. $\alpha = \frac{\alpha^\tau}{\alpha^{\tau-1}}$), in* **Theorem 1** *we propose dynamic programming operators, actualizing the future, at stage $t$, with the factor $\frac{\lambda_\tau}{\lambda_{\tau-1}}$.*

*Similar modifications will be done in* **Section 5***, when dealing with the infinite horizon problem.*

Although by the economic motivations, it is reasonable to ask the discount factors $\lambda_t$ to be smaller than 1, this assumption is not necessary to assure the well definition of the values in the finite-horizon analysis, neither to proof the next result.

**Theorem 1.** *Let $V_0$, $V_1$,...,$V_N$ be the functions on $\mathcal{S}$ defined by the recursion*

$$V_N(s) = r_N(s) \ ,$$

$$V_n(s) = \sup_{a \in \mathcal{A}_s} \left\{ r_n^a(s) + \frac{\lambda_n}{\lambda_{n-1}} \int_{\mathcal{S}} V_{n+1}(z) Q_n^a(dz|s) \right\} \ , n = N-1, N-2, ..., 0 \ . \quad (1)$$

*Let $f_n^*$ be a function defined on $\mathcal{S}$, where for each $s \in \mathcal{S}$, $f_n^*(s) \in \mathcal{A}_s$ attains the maximum at* (1)*. Then, the functions $f_n^*$ are well defined, the Markov strategy $\pi^* = \{f_0^*, f_1^*, ..., f_{N-1}^*\}$ is optimal, and the value function $V_N^*$ equals $V_0$.*

*Proof.* Let $\pi = \{\pi_t\}$ be an arbitrary policy and let $V_{N,n}^\pi(s)$ the corresponding performance from time $n$ to the terminal time $N$, given the state $s_n = s$ at time $n$, i.e., if $n = 0, 1, \ldots, N-1$. That is, by definition,

$$V_{N,n}^\pi(s) := \mathbb{E}_s^\pi \left[ r_n^{A_n}(S_n) + \frac{1}{\lambda_{n-1}} \left( \sum_{t=n+1}^{N-1} \lambda_{t-1} r_t^{A_t}(S_t) + \lambda_{N-1} r_N(S_N) \right) \right] \ ,$$

$$V_{N,N}^\pi(s) := \mathbb{E}_s^\pi \left[ r_N(S_N) \right] = r_N(s) \ .$$

In particular, note that

$$V_N^\pi(s) = V_{N,0}^\pi(s) \ .$$

To prove the theorem, will shall show that, for all $s \in \mathcal{S}$ and $n = 0, 1, \ldots, N$,

$$V_{N,n}^\pi(s) \leqq V_n(s) \ , \quad (2)$$

with equality if $\pi = \pi^*$:

$$V_{N,n}^{\pi^*}(s) = V_n(s) \ . \quad (3)$$

For $n = 0$, this inequality reads, for any $s \in \mathcal{S}$,

$$V_N^\pi(s) \leqq V_N^*(s) = V_N^{\pi^*}(s) \ ,$$

which yields the desired conclusion.

To obtain the proposed inequalities, we proceed by backward induction.

Observe first that for $n = N$,

$$V^{\pi}_{N,N}(s) = V_N(s) = r_N(s) \ . \tag{4}$$

Let us now assume that for some $n = N-1, \ldots, 0$, and any $s \in \mathcal{S}$,

$$V^{\pi}_{N,n+1}(s) \leqq V_{n+1}(s) \ . \tag{5}$$

Then,

$$\begin{aligned} V^{\pi}_{N,n}(s) = &\ \mathbb{E}^{\pi}_{s}\left[ r^{A_n}_{n}(S_n) + \frac{1}{\lambda_{n-1}}\left(\sum_{t=n+1}^{N-1} \lambda_{t-1} r^{A_t}_{t}(S_t) + \lambda_{N-1} r_N(S_N)\right)\right] \\ = &\ \mathbb{E}^{\pi}_{s}\left[ r^{A_n}_{n}(S_n) + \frac{\lambda_n}{\lambda_{n-1}}\left(\sum_{t=n+1}^{N-1} \frac{\lambda_{t-1}}{\lambda_n} r^{A_t}_{t}(S_t) + \frac{\lambda_{N-1}}{\lambda_n} r_N(S_N)\right)\right] \\ = &\ \int_{\mathcal{A}_s}\left[ r^{a}_{n}(s) + \frac{\lambda_n}{\lambda_{n-1}}\int_{\mathcal{S}} V^{\pi}_{N,n+1}(z)\ Q^{a}_{n}(dz|s)\right]\pi_n(da|s) \ . \end{aligned}$$

Hence, for any $s \in \mathcal{S}$,

$$\begin{aligned} V^{\pi}_{N,n}(s) &\leqq \int_{\mathcal{A}_s}\left[ r^{a}_{n}(s) + \frac{\lambda_n}{\lambda_{n-1}}\int_{\mathcal{S}} V_{n+1}(z)\ Q^{a}_{n}(dz|s)\right]\pi_n(da|s) \\ &\leqq \sup_{a\in\mathcal{A}_s}\left\{ r^{a}_{n}(s) + \frac{\lambda_n}{\lambda_{n-1}}\int_{\mathcal{S}} V_{n+1}(z) Q^{a}_{n}(dz|s)\right\} = V_n(s) \ , \end{aligned} \tag{6}$$

which proves (2).

The existence of the functions $f^*_n$ stated follows by **Assumption 1**. Indeed, by (d), the functions $V_{n+1}$ are bounded, and by (c) and (e), for any $s \in \mathcal{S}$, the application

$$a \mapsto r^{a}_{n}(s) + \frac{\lambda_n}{\lambda_{n-1}}\int_{\mathcal{S}} V_{n+1} Q^{a}_{n}(dz|s)$$

results upper semicontinuous on $\mathcal{A}_s$. Finally, part (b) implies the existence of the maximizing action.

If equality holds in (5), with $\pi = \pi^*$, then (6) becomes equality, for any $t$. Since equality holds in (4), this implies (3) and then $\pi^*$ is the optimal policy.

$\mathcal{Q.E.D.}$

## 5 The Infinite Horizon Problem

In this section we shall work under the next additional assumption, that assure the well definition of the value associated to a policy.

**Assumption 2**

(a) *The functions $r_t$ are uniformly bounded. That is, $|r_t^a(s)| \leqq M$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}_s$ and $t \in \mathbb{N}$.*

(b) *There exists a positive constant $\rho < 1$ such that, for any $t$, $\lambda_t \leqq \ \rho \ \lambda_{t-1}$, for $t \geqq 0$.*

Since **Assumption 2** (b) implies $\lambda_t \leqq \rho^{t+1}$, by easy calculations, it can be verified that for any $\pi \in \Pi$, $V^\pi$ is a bounded function and it holds $||V^\pi||_\infty \leqq \frac{M}{1-\rho}$ . In consequence, there exists the value function $V^*$, bounded by $\frac{M}{1-\rho}$.

The classical approach to characterize value functions and optimal policies in discounted **MDP** consists in the successive application of appropriate dynamic programming operators, which have sense for stationary models. See, for instance, [8, Chapter 5], or [4, Chapter 4].

In order to deal with the infinite horizon non-stationary variable discounted original model, we consider a related stationary one, incorporating the time parameter to the state of the system. A similar construction, in the context of non-stationary finite-horizon Markov games, were proposed in [10, Section 5].

Let us note that, if our aim is to adapt the known constant discount factor proofs to our case, this transformation is even necessary for stationary reward functions and transition probabilities, given the non-stationary character of the discounts.

We shall prove results on this enlarged model, where discounts depends artificially on the state. The relations between policies and values in both models will be pointed in **Remarks 2** and **3**.

Formalizing, let us consider

$$\tilde{\mathcal{M}} := (\tilde{\mathcal{S}}, \tilde{\mathcal{A}}, \{\tilde{\mathcal{A}}_{\tilde{s}} : \tilde{s} \in \tilde{\mathcal{S}}\}, \tilde{Q}, \tilde{r}, \{\tilde{\lambda}_{\tilde{s}}, \tilde{s} \in \tilde{\mathcal{S}}\})$$

where we put

- $\tilde{\mathcal{S}} = \mathcal{S} \times \mathbb{N}_0$
- $\tilde{\mathcal{A}} = \mathcal{A}$
- $\tilde{\mathcal{A}}_{(s,\tau)} = \mathcal{A}_s$ for any $(s,\tau) \in \tilde{\mathcal{S}}$
- $\tilde{r}^a(s,\tau) = r_\tau^a(s)$, for $(s,\tau) \in \tilde{\mathcal{S}}$, $a \in \tilde{\mathcal{A}}_s$.
- $\tilde{Q}^a(z,\tau'|s,\tau) = \begin{cases} Q^a(z|s) & \text{if } \ \tau' = \tau + 1 \\ \quad 0 & \quad \text{otherwise} \end{cases}$ for $(s,\tau),(z,\tau') \in \tilde{\mathcal{S}}$, $a \in \tilde{\mathcal{A}}_s$.
- $\tilde{\lambda}_{(s,\tau)} = \lambda_\tau$, for each $(s,\tau) \in \tilde{\mathcal{S}}$.

**Remark 2** *There is a one to one correspondence between stationary policies in the new model $\tilde{\mathcal{M}}$ and Markov policies in the original $\mathcal{M}$. Moreover, if $\tilde{f}$ is stationary in $\tilde{\mathcal{M}}$, the corresponding Markov policy in $\mathcal{M}$ is given by $\pi = \{f_\tau\}$, where $f_\tau(s) = \tilde{f}(s,\tau)$.*

We shall note $\tilde{\Pi}$ and $\tilde{\Pi}_{\text{stat}}$ to the set of Markov and stationary policies in $\tilde{\mathcal{M}}$.

For $\tilde{\pi} \in \tilde{\Pi}$ and $(s,\tau) \in \tilde{\mathcal{S}}$ there exists $\mathrm{P}^{\tilde{\pi}}_{(s,\tau)}$ a probability measure, with $\mathbb{E}^{\tilde{\pi}}_{(s,\tau)}$ its consequent expectation operator. Let $\{\tilde{S}_t\}$ be the stochastic process on $\tilde{\mathcal{S}}$ of the state system at time $t$, i.e., $\tilde{S}_t = (S_t, t)$.

Since the action spaces are not essentially modified, we skip the *tildes* in the notation related to these parameters.

At this point it is worth alerting the reader about the notation utilized in the rest of the section. We shall write $\tilde{\lambda}_{\tilde{S}_t}$ or $\tilde{\lambda}_{(S_t,t)}$ to the random variables making up the discount stochastic process, and $\tilde{\lambda}_{\tilde{s}_t}$ or $\tilde{\lambda}_{(s_t,t)}$, to its realisations, which, by construction, equals $\lambda_t$.

In this enlarged model we define the value associated to an initial state and a fixed policy, by the performance of the policy in the original problem, from time $\tau$, given the state $s_\tau = s$ at time $\tau$. That is, for $(s,\tau) \in \tilde{\mathcal{S}}$ and $\tilde{\pi} = \{\tilde{f}_0, \tilde{f}_1, ...\} \in \tilde{\Pi}$,

$$\begin{aligned}\tilde{V}^{\tilde{\pi}}(s,\tau) &:= \frac{1}{\tilde{\lambda}_{(s,\tau-1)}} \mathbb{E}^{\tilde{\pi}}_{(s,\tau)} \left[ \tilde{\lambda}_{(s,\tau-1)} \tilde{r}^{A_\tau}(s,\tau) + \sum_{t=\tau+1}^{\infty} \tilde{\lambda}_{\tilde{S}_{t-1}} \tilde{r}^{A_t}(\tilde{S}_t) \right] \\ &= \tilde{r}^{\tilde{f}_\tau}(s,\tau) + \mathbb{E}^{\tilde{\pi}}_{(s,\tau)} \left[ \sum_{t=\tau+1}^{\infty} \frac{\tilde{\lambda}_{\tilde{S}_{t-1}}}{\tilde{\lambda}_{(s,\tau-1)}} \tilde{r}^{A_t}(\tilde{S}_t) \right] .\end{aligned}$$

Here again, the objective will be to find polices $\tilde{\pi}$, such that at state $(s,\tau) \in \tilde{\mathcal{S}}$, solve

$$\tilde{\pi}(s,\tau) \in \arg\max_{\tilde{\pi}} \tilde{V}^{\tilde{\pi}}(s,\tau) .$$

**Remark 3** *The values $\tilde{V}^*(s,\tau)$ represents the optimal expected reward for the infinite horizon variable discounted non-stationary problem $\mathcal{M}$ starting at time $\tau$ in state $s$. In particular, for each $s \in \mathcal{S}$, it is $\tilde{V}^*(s,0) = V^*(s)$.*

We will note with $B(\tilde{\mathcal{S}})$ to the bounded function space defined on $\tilde{\mathcal{S}}$, which results a Banach space (normed and complete) with the sup norm $||\cdot||_\infty$.

Again, inspired as in **Remark 1**, on $B(\tilde{\mathcal{S}})$ we define the new dynamic programming operators, given $\tilde{f} \in \tilde{\Pi}_{\text{stat}}$

$$(T^{\tilde{f}}v)(s,\tau) = \tilde{r}^{\tilde{f}}(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathcal{S}}} v(z,\tau')\tilde{Q}^{\tilde{f}}(dz,\tau'|s,\tau) ,$$

and

$$(Tv)(s,\tau) = \sup_{a \in \mathcal{A}_s} \left\{ \tilde{r}^a(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathcal{S}}} v(z,\tau')\tilde{Q}^a(dz,\tau'|s,\tau) \right\} .$$

Observe that if **Assumption 2** holds, $T$ maps $B(\tilde{\mathcal{S}})$ into itself. Indeed, if $||v||_\infty \leqq C$, then it is verified $||Tv||_\infty \leqq M + \rho\, C$. The same consideration holds for the operators $T^f$.

**Lemma 1** *$T$ and $T^{\tilde{f}}$ are monotone and contractive mappings on $B(\tilde{\mathcal{S}})$, of modulus $\rho$.*

*Proof.* To proof monotonicity, let be $u, w \in B(\tilde{\mathcal{S}})$, with $u \leqq w$. Then, for all $(s,\tau) \in \tilde{\mathcal{S}}$,

$$\begin{aligned}(T^{\tilde{f}}u)(s,\tau) &= \int_{\mathcal{A}_s} \tilde{r}^a(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathcal{S}}} u(z,\tau')\tilde{Q}^a(dz,\tau'|s,\tau)\tilde{f}(da|s,\tau) \\ &\leqq \int_{\mathcal{A}_s} \tilde{r}^a(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathcal{S}}} w(z,\tau')\tilde{Q}^a(dz,\tau'|s,\tau))\tilde{f}(da|s,\tau) \\ &= (T^{\tilde{f}}w)(s,\tau) \ .\end{aligned}$$

Besides, if $\kappa > 0$, since, for all $(s,\tau) \in \tilde{\mathcal{S}}$, $\frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \leqq \rho$,

$$\begin{aligned}(T^{\tilde{f}}(u+\kappa))(s,\tau) &= \int_{\mathcal{A}_s} \tilde{r}^a(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathcal{S}}} [u(z,\tau')+\kappa]\tilde{Q}^a(dz,\tau'|s,\tau)\tilde{f}(da|s,\tau) \\ &= (T^{\tilde{f}}u)(s,\tau) + \kappa\ \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}}\ \leqq\ (T^{\tilde{f}}u)(s,\tau) + \kappa\ \rho \ .\end{aligned}$$

Now, for $u, w \in B(\tilde{\mathcal{S}})$, since $u \leqq w + ||u-w||_\infty$, by the monotonicity property and the previous observation,

$$T^{\tilde{f}}u \ \leqq\ T^{\tilde{f}}w + \rho\ ||u-w||_\infty \ .$$

Interchanging the functions $u$ and $w$,

$$T^{\tilde{f}}w \ \leqq\ T^{\tilde{f}}u + \rho\ ||u-w||_\infty \ ,$$

which implies

$$||T^{\tilde{f}}u - T^{\tilde{f}}w||_\infty \leqq \rho\ ||u-w||_\infty$$

and $T^{\tilde{f}}$ is contractive of modulus $\rho$.

Using similar arguments, it can be shown that $T$ is a monotone and contractive mapping of modulus $\rho$ on $B(\tilde{\mathcal{S}})$.

$\mathcal{Q.E.D.}$

By Banach's Fixed Point Theorem, applied to the contractive operators $T^{\tilde{f}}$ and $T$ defined on $(B(\tilde{\mathcal{S}}), ||\cdot||_\infty)$ (indeed complete), there exist unique bounded functions $v^{\tilde{f}}$ and $v$, satisfying $T^{\tilde{f}}v^{\tilde{f}} = v^{\tilde{f}}$ and $Tv = v$.

**Lemma 2** *The value of a stationary strategy $\tilde{f} \in \tilde{\Pi}_{\text{stat}}$, $\tilde{V}^{\tilde{f}}$ is the unique fixed point of $T^{\tilde{f}}$ on $B(\tilde{\mathcal{S}})$.*

*Proof.* In view of **Lemma 1** and its following observation, it is sufficient to prove that $\tilde{V}^{\tilde{f}}$ is a fixed point of $T^{\tilde{f}}$, which follows with the next calculations, for $(s,\tau) \in \tilde{\mathcal{S}}$, in which we shall use the constant character of the discount on the state $s$. Since $\tilde{\lambda}_{(s,\tau)} = \lambda_\tau$ for all $s \in \mathcal{S}$, the random variables $\tilde{\lambda}_{\tilde{S}_\tau}$ equal the values $\lambda_\tau$.

$$
\begin{aligned}
\tilde{V}^{\tilde{f}}(s,\tau) &= \tilde{r}^{\tilde{f}}(s,\tau) + \mathbb{E}^{\tilde{f}}_{(s,\tau)} \left[ \sum_{t=\tau+1}^{\infty} \frac{\tilde{\lambda}_{\tilde{S}_{t-1}}}{\tilde{\lambda}_{(s,\tau-1)}} \, \tilde{r}^{A_t}(\tilde{S}_t) \right] \\
&= \tilde{r}^{\tilde{f}}(s,\tau) + \mathbb{E}^{\tilde{f}}_{(s,\tau)} \left[ \frac{\tilde{\lambda}_{\tilde{S}_\tau}}{\tilde{\lambda}_{(s,\tau-1)}} \, \tilde{r}^{A_{\tau+1}}(\tilde{S}_{\tau+1}) + \sum_{t=\tau+2}^{\infty} \frac{\tilde{\lambda}_{\tilde{S}_{t-1}}}{\tilde{\lambda}_{(s,\tau-1)}} \, \tilde{r}^{A_t}(\tilde{S}_t) \right] \\
&= \tilde{r}^{\tilde{f}}(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \, \mathbb{E}^{\tilde{f}}_{(s,\tau)} \left[ \tilde{r}^{A_{\tau+1}}(\tilde{S}_{\tau+1}) + \sum_{t=\tau+2}^{\infty} \frac{\lambda_{\tilde{S}_{t-1}}}{\tilde{\lambda}_{\tilde{S}_\tau}} \, \tilde{r}^{A_t}(\tilde{S}_t) \right] \\
&= \tilde{r}^{\tilde{f}}(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \, \mathbb{E}^{\tilde{f}}_{(s,\tau)} \left[ V^{\tilde{f}}(\tilde{S}_{\tau+1}) \right] \; = \; (T^{\tilde{f}} V^{\tilde{f}})(s,\tau)
\end{aligned}
$$

$\mathcal{Q.E.D.}$

**Theorem 2.** *The value function $\tilde{V}^*$ is the unique bounded function on $\tilde{\mathcal{S}}$ satisfying the optimality equation $v = Tv$. That is, for any $(s,t) \in \tilde{\mathcal{S}}$,*

$$
\tilde{V}^*(s,\tau) = \sup_{a \in \mathcal{A}_s} \left\{ \tilde{r}^a(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathcal{S}}} \tilde{V}^*(z,\tau') \tilde{Q}^a(dz,\tau'|s,\tau) \right\} .
$$

*Moreover, there exist stationary policies $\tilde{f}^* \in \tilde{\Pi}_{\text{stat}}$, which at each $(s,\tau) \in \tilde{\mathcal{S}}$ select an action maximizing the r.h.d. of the equation. Any strategy $\tilde{f}^*$ is optimal for the infinite horizon problem, i.e.,*

$$
T^{\tilde{f}^*} \tilde{V}^* \;=\; \tilde{V}^* \; , \quad \textit{and} \quad \tilde{V}^{\tilde{f}^*} = \tilde{V}^* \; .
$$

*Proof.* By **Assumption 1**, parts (c) and (d), in model $\tilde{\mathcal{M}}$, for any $(s,\tau) \in \tilde{\mathcal{S}}$ the application

$$
a \mapsto \tilde{r}^a(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathcal{S}}} \tilde{V}^*(z,\tau') \tilde{Q}^a(dz,\tau'|s,\tau)
$$

is upper semicontinuous on $\mathcal{A}_s$, since $\tilde{V}^*$ is bounded on $\tilde{\mathcal{S}}$.

Besides, by **Assumption 1** (b), $\mathcal{A}_s$ is compact, and there exists an action $a^*_{s,\tau} = \tilde{f}^*(s,\tau) \in \mathcal{A}_s$ which verifies

$$
\begin{aligned}
& r^{a^*_{s,\tau}}(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathcal{S}}} \tilde{V}^*(z,\tau') \tilde{Q}^{a^*_{s,\tau}}(dz,\tau'|s,\tau) \\
&= \sup_{a \in \mathcal{A}_s} \left\{ \tilde{r}^a(s,\tau) + \frac{\tilde{\lambda}_{(s,\tau)}}{\tilde{\lambda}_{(s,\tau-1)}} \int_{\tilde{\mathcal{S}}} \tilde{V}^*(z,\tau') \tilde{Q}^a(dz,\tau'|s,\tau) \right\} .
\end{aligned}
$$

We shall prove that any stationary strategy defined in such a way is optimal, by proving that, for any $\tilde{\pi} \in \tilde{\Pi}$ and $(s, \tau) \in \tilde{\mathcal{S}}$,

$$\tilde{V}^{\tilde{\pi}}(s,\tau) \leqq \tilde{V}^{\tilde{f}^*}(s,\tau) \ .$$

For each $t \geqq 1$, $\tilde{h}_t \in \tilde{\mathcal{H}}_t$, $a \in \mathcal{A}_{s_t}$, and $\tau \geqq t$, since at the $t$-th epoch of decision, given the history $\tilde{h}_t$, the realizations of the processes $\tilde{S}_k$ and $A_k$ were respectively $\tilde{s}_k = (s_k, k)$ and $a_k$, by properties of the conditional expectation we have

$$\begin{aligned}
\mathbb{E}^{\tilde{\pi}}_{(s,\tau)}\left[\tilde{\lambda}_{\tilde{S}_t}\tilde{V}^{\tilde{f}^*}(\tilde{S}_{t+1})\middle|\tilde{h}_t, a_t\right] &= \tilde{\lambda}_{\tilde{s}_t}\ \mathbb{E}^{\tilde{\pi}}_{(s,\tau)}\left[\tilde{V}^{\tilde{f}^*}(\tilde{S}_{t+1})\middle|\tilde{h}_t, a_t\right] \\
&= \tilde{\lambda}_{\tilde{s}_t}\int_{\tilde{\mathcal{S}}}\tilde{V}^{\tilde{f}^*}(z,\tau')\tilde{Q}^{\tilde{f}_t}(dz,\tau'|s_t,t) \\
&= \tilde{\lambda}_{s_{t-1}}\left[\frac{\tilde{\lambda}_{\tilde{s}_t}}{\tilde{\lambda}_{s_{t-1}}}\int_{\tilde{\mathcal{S}}}\tilde{V}^{\tilde{f}^*}(z,\tau')\tilde{Q}^{\tilde{f}_t}(dz,\tau'|s_t,t)\right. \\
&\qquad \left. + \tilde{r}^{\tilde{f}_t}(\tilde{s}_t) - \tilde{r}^{\tilde{f}_t}(\tilde{s}_t)\right] \\
&\leqq \tilde{\lambda}_{s_{t-1}}\left[\tilde{V}^{\tilde{f}^*}(\tilde{s}_t) - \tilde{r}^{\tilde{f}_t}(\tilde{s}_t)\right] \ .
\end{aligned}$$

The above inequality is equivalent to

$$\begin{aligned}
\tilde{\lambda}_{s_{t-1}}\tilde{r}^{\tilde{f}_t}(\tilde{s}_t) \ &\leqq \ \tilde{\lambda}_{s_{t-1}}\tilde{V}^{\tilde{f}^*}(\tilde{s}_t) \\
&- \ \mathbb{E}^{\tilde{\pi}}_{(s,\tau)}\left[\tilde{\lambda}_{\tilde{S}_t}\tilde{V}^{\tilde{f}^*}(\tilde{S}_{t+1})\middle|\tilde{h}_t, a_t\right] \ . \qquad (7)
\end{aligned}$$

On the other hand, for $t = 0$, as consequence of **Lemma 2**, it holds

$$\tilde{r}^{a_0}(s) \leqq \tilde{V}^{\tilde{f}^*}(s) - \mathbb{E}^{\tilde{\pi}}_{(s,0)}\left[\tilde{\lambda}_{(s,0)}\tilde{V}^{\tilde{f}^*}(\tilde{S}_{t+1})\right] \ .$$

Finally, for any $\tau \in \mathbb{N}$, taking expectations under policy $\tilde{\pi}$ and summing for $t = \tau, ..., n$ the preceding inequalities we obtain a telescopic sum which conduces to

$$\begin{aligned}
&\mathbb{E}^{\tilde{\pi}}_{(s,\tau)}\left[\sum_{t=\tau}^{n}\tilde{\lambda}_{\tilde{S}_{t-1}}\tilde{r}^{A_t}(\tilde{S}_t)\right] \\
&\leqq \tilde{\lambda}_{(s,\tau-1)}V^{\tilde{f}^*}(s,\tau) - \mathbb{E}^{\tilde{\pi}}_{(s,\tau)}\left[\tilde{\lambda}_{\tilde{S}_n}\tilde{V}^{\tilde{f}^*}(\tilde{S}_{n+1})\right] \ ,
\end{aligned}$$

which gives

$$\begin{aligned}
&\mathbb{E}^{\tilde{\pi}}_{(s,\tau)}\left[\sum_{t=\tau}^{n}\frac{\tilde{\lambda}_{\tilde{S}_{t-1}}}{\tilde{\lambda}_{(s,\tau-1)}}\tilde{r}^{A_t}(\tilde{S}_t)\right] \\
&\leqq V^{\tilde{f}^*}(s,\tau) - \mathbb{E}^{\tilde{\pi}}_{(s,\tau)}\left[\frac{\tilde{\lambda}_{\tilde{S}_n}}{\tilde{\lambda}_{(s,\tau-1)}}\tilde{V}^{\tilde{f}^*}(\tilde{S}_{n+1})\right] \ .
\end{aligned}$$

Taking now $n \to \infty$, the l.h.s. of the inequality tends to the value $\tilde{V}^{\tilde{\pi}}(s,\tau)$, and the second term in the r.h.s. verifies, since $\rho < 1$,

$$\left| \mathbb{E}^{\tilde{\pi}}_{(s,\tau)} \left[ \frac{\tilde{\lambda}_{\tilde{S}_n}}{\tilde{\lambda}_{(s,\tau-1)}} \tilde{V}^{\tilde{f}^*}(\tilde{S}_{n+1}) \right] \right| \leqq \frac{M\rho^{n-\tau}}{1-\rho} \underset{n\to\infty}{\longrightarrow} 0 \ ,$$

and in consequence,

$$\tilde{V}^{\tilde{\pi}}(s,\tau) \leqq \tilde{V}^{\tilde{f}^*}(s,\tau) \ .$$

The optimality of $\tilde{f}^*$ follows from the arbitrary character of $\tilde{\pi}$.

$\mathcal{Q.E.D.}$

## 6 Concluding Remarks

In this work we have treated the problem of define in a precise way discounted discrete time stochastic control models, were the discounts do not stay constant at each stage, obtaining results in both, the finite and the infinite horizon cases.

For the finite-horizon problem, we have shown the existence of optimal deterministic Markov policies and provide a recursive method that allows to obtain the value function and optimal policies of the model.

For the infinite horizon case we incorporate the time parameter to the state obtaining a stationary model, in which the discount varies with the state. We define appropriate dynamic operators, and characterize the value function as its unique bounded fixed point. We also show in this enlarged model the existence of deterministic stationary optimal strategies. We obtain results for the original problem noting that stationary policies in the stationary model traduces in Markov policies in the non-stationary one.

The practical issue is currently to find algorithms of procedures to evaluate numerically or find good approximations of the value functions for the variable discounted infinite horizon problems, bounding efficiently the errors incurred.

## References

1. Ayong Le Kama A., Schubert K.; *A note on the consequences of an endogenous discounting depending on the environmental quality* Macroeconomic Dynamics 11, 2 (2007) 272-289. `http://halshs.archives-ouvertes.fr/halshs-00206326`
2. Derman C.: Finite State Markovian Decision Processes. Academic Press (1970)
3. Heal G.: Valuing the Future: Economic Theory and Sustainability. Columbia University Press (2000)
4. Hernández-Lerma O., Lasserre J.B.: Discrete-Time Markov Control Processes. Springer-Verlag (1996)
5. Kallenberg L.:Finite state and action MDPS. Handbook of Markov Decision Processes. Methods and applications. Kluwer's international Series (2002)

6. Mas-Colell A., Whinston M., Green J.: Microeconomic Theory. Oxford University Press (1995)
7. Puterman L.: Markov Decision Processes. Wiley and Sons (2005)
8. Ross, S.: Applied Probability Models with Optimization Applications. Holden-Day (1970)
9. Schubert K.: Éléments sur l'actualisation et l'environnement. Rech. économiques de Louvain. 2 (Vol. 72) 157–175 (2006)
10. Tidball M., Altman E.: Approximations in dynamics zero-sum games. SIAM J. Control and Optimization. 34, 1, 311–328 (1996)