6-27-2020

Discrete-Time Control with Non-Constant Discount Factor

Héctor Jasso-Fuentes
CINVESTAV-IPN

José-Luis Menaldi
Wayne State University, menaldi@wayne.edu

Tomás Prieto-Rumeau
National Distance Education University

Recommended Citation

Jasso-Fuentes, H., Menaldi, JL. & Prieto-Rumeau, T. Discrete-time control with non-constant discount factor. *Math Meth Oper Res* **92**, 377–399 (2020). https://doi.org/10.1007/s00186-020-00716-8

Mathematical Methods of Operations Research, Vol. 92, No. 2, 2020, pp. 377-399.
(https://doi.org/10.1007/s00186-020-00716-8)

Discrete-time control with non-constant discount factor

Héctor Jasso-Fuentes[1], José-Luis Menaldi[2], and Tomás Prieto-Rumeau[3]

March 28, 2021

Abstract

This paper deals with discrete-time Markov decision processes (MDPs) with Borel state and action spaces, and total expected discounted cost optimality criterion. We assume that the discount factor is not constant: it may depend on the state and action; moreover, it can even take the extreme values zero or one. We propose sufficient conditions on the data of the model ensuring the existence of optimal control policies and allowing the characterization of the optimal value function as a solution to the dynamic programming equation. As a particular case of these MDPs with varying discount factor, we study MDPs with stopping, as well as the corresponding optimal stopping times and contact set. We show applications to switching MDPs models and, in particular, we study a pollution accumulation problem.

2010 Mathematics Subject Classification: 93E20, 34A38, 60J05.
Keywords and phrases: *Markov decision processes, dynamic programming, optimal stopping problems.*

1 Introduction

The study of Markov decision processes (MDPs) has been very active along the past decades. Many papers and books cover the theory and the applications of this topic: see, among others, the books by Bertsekas [4], Hernández-Lerma and Lasserre [7, 8], Hinderer et al. [10], Puterman [22], and Ross [24], to cite just a few. In most of the literature, two main optimality criteria have prevailed due to a significant number of applications; namely, the discounted and the average cost/reward criteria.

It is worth noting that the discounted criterion has two appealing features: (a) it allows to approximate the total (undiscounted) cost when the discount factor is close to one, (b) it models certain phenomena whose future responses (measured as cost/reward cash flows) need to be measured at the present time; for example, in mathematical finance, the investor is interested in optimizing fair values where the discount factor plays the role of an interest rate.

[1]Departmento de Matemáticas. CINVESTAV-IPN. A. Postal 14-740, Ciudad de México 07000, México. Email: hjasso@math.cinvestav.mx

[2]Department of Mathematics, Wayne State University, Detroit, MI 48202, USA Email: menaldi@wayne.edu

[3]Corresponding author. Department of Statistics and Operations Research, UNED. Calle Senda del Rey 9, 28040 Madrid, Spain. Email: tprieto@ccia.uned.es Supported by grant MTM2016-75497-P from the Spanish Ministerio de Economía y Competitividad

It is also well known that most of the literature on discrete-time discounted MDPs considers *constant* discount factors. There exist, however, a few references dealing with non-constant discount factors, allowing for a dependence on the state-action variables (or even, additionally, a random variable). See, e.g., the references Ilhuicatzi-Roldán [9], Minjárez-Sosa [20], and Wei and Guo [25]. In all these mentioned references, however, it is assumed that the discount factor is *uniformly bounded away from one*. One of the main goals of this paper is precisely to relax this condition and then allow the discount factor to take values close to (or even equal to) one. This will be achieved, among other conditions, by using properties related to *small sets*, in the terminology of Markov chains.

We apply our main results herein to control problems with stopping; i.e., when the dynamic system stops either (a) by some action of the decision-maker or (b) by a natural transition to some given special state. To this end, we use the varying discount factor feature of our MDP as a powerful modeling tool: namely, a discount factor equal to one can model an *instantaneous* transition, whereas a discount factor equal to zero can be interpreted as stopping the dynamics. For more details of MDPs with stopping we can quote the books by Bensoussan [3], Puterman [22], Ross [24], as well as the papers by Rieder [23], Dufour and Piunovskiy [5], and Horiguchi [11, 12], among others. Interestingly, in all these references, instantaneous transitions are not allowed, which is a situation that our model can indeed handle.

The remainder of the paper is organized as follows: in Section 2 we introduce the primitive data of the model and our main assumptions. Next we define the corresponding dynamic programming equation and prove existence of solutions to this equation, which allow to obtain the optimal value of the problem as well as optimal policies. Section 3 is concerned with the study of MDPs with stopping. Within this section we first analyze the simplest case when the system is stopped only by the decision-maker and then we study a more general case when the system can reach an absorbing state by either a natural transition of the system or by means of a decision-maker's action. Finally, in Section 4 we introduce a special case of MDPs with stopping concerned with switching problems and we show an application to a pollution accumulation problem.

Notation and terminology. Throughout this paper:

- We will write $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. By $\mathbb{N} = \{1, 2, 3, \ldots\}$ we will denote the set of natural numbers, and we define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

- Any metric space Z will be endowed with its Borel σ-algebra $\mathcal{B}(Z)$ and measurability (of sets and functions) will be always referred to the corresponding Borel σ-algebras. For product spaces, we will consider the taxicab metric and the corresponding product of Borel σ-algebras.

- A Borel space is a measurable subset of a complete and separable metric space.

- We make the convention that a product of real numbers $\prod_{i \in S} x_i$ over an empty set S equals one, while a sum $\sum_{i \in S} x_i$ over an empty set S equals zero.

- The notation $\delta_x(\cdot)$ stands for the Dirac probability measure concentrated at the point $x \in Z$, while $\mathbb{1}_C(\cdot)$ will denote the indicator function of a set $C \in \mathcal{B}(Z)$.

- Given a transition probability measure $Q(\cdot|\cdot)$ on X given X, we define $Q^1 = Q$ and, recursively for $n > 1$, we let $Q^{n+1}(B|x) = \int_X Q(B|y)Q^n(dy|x)$ for $x \in X$ and $B \in \mathcal{B}(X)$.
- We say that $f : Z \to \overline{\mathbb{R}}$ is lower semicontinuous if the level sets $\{z \in Z : f(z) \leq \beta\}$, for any $\beta \in \mathbb{R}$, are closed. We will write $f \in \overline{\mathbb{L}}(Z)$. If, in addition, f takes values in $\mathbb{R}$, we will simply write $f \in \mathbb{L}(Z)$.
- The family of measurable functions $f : Z \to \overline{\mathbb{R}}$ and $f : Z \to \mathbb{R}$ are respectively denoted by $\overline{\mathbb{M}}(Z)$ and $\mathbb{M}(Z)$.
- By $\mathbb{C}(Z)$ we will denote the family of continuous functions from Z to $\mathbb{R}$.
- In general, for sets of functions, the superscript + stands for functions taking nonnegative values, and the subscript b stands for bounded functions.

2 The general model

The tuple consisting of the elements

$$\mathfrak{M} = (X, A, \mathbb{K}, Q, c, \alpha), \tag{2.1}$$

which are defined next, will be referred to as the Markov decision process (MDP) model.

State-action pairs. The *state space* X and the *action space* A are both Borel spaces. Furthermore, we will consider the family of sets $\{A(x) : x \in X\}$, where $A(x)$ is regarded as the set of feasible actions at state x. For each state $x \in X$, the set $A(x) \subseteq A$ is nonempty and measurable. We will also assume that the graph

$$\mathbb{K} = \{(x, a) \in X \times A : x \in X, a \in A(x)\}$$

is a measurable subset of $X \times A$. The family of measurable functions $f : X \to A$ which satisfy $f(x) \in A(x)$ for all $x \in X$ is supposed to be nonempty. We will denote by $\mathbb{F}$ the class of all such functions.

The dynamic system. The system dynamics is given by the transition kernel

$$Q : \mathcal{B}(X) \times \mathbb{K} \to [0, 1],$$

which satisfies the following conditions: for every $B \in \mathcal{B}(X)$ the function $(x, a) \mapsto Q(B|x, a)$ is measurable on $\mathbb{K}$ and, in addition, for every $(x, a) \in \mathbb{K}$ the mapping $B \mapsto Q(B|x, a)$ is a probability measure on $(X, \mathcal{B}(X))$. Given a measurable function $u \in \mathbb{M}^+(X)$ we define the function $Qu \in \overline{\mathbb{M}}^+(\mathbb{K})$ by means of

$$Qu(x, a) = \int_X u(y)Q(dy|x, a) \quad \text{for } (x, a) \in \mathbb{K}.$$

Control policies. Define $H_0 = X$ and $H_t = \mathbb{K}^t \times X$ for $t \in \mathbb{N}$, and let $H_\infty = \mathbb{K}^\infty$, all endowed with the corresponding product σ-algebras. The history up to time t is

$$h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t) \in H_t \quad \text{for } t \in \mathbb{N}.$$

An admissible policy is a sequence $\nu = \{\nu_t\}_{t\geq 0}$ of transition probability measures on A given H_t such that $\nu_t(A(x_t)|h_t) = 1$ for all $h_t \in H_t$ and $t \in \mathbb{N}_0$. We denote by Π the set of admissible control policies.

If there is some $f \in \mathbb{F}$ such that the policy $\nu \in \Pi$ satisfies $\nu_t(B|h_t) = \delta_{f(x_t)}(B)$ for every $B \in \mathcal{B}(X)$, $h_t \in H_t$, and $t \in \mathbb{N}_0$, then we say that ν is a *deterministic stationary* policy. In what follows, we will identify the set of such policies with $\mathbb{F}$. Hence, we have $\mathbb{F} \subseteq \Pi$.

Construction of the controlled process. By the Ionescu-Tulcea theorem, for any initial state $x \in X$ and any policy $\nu \in \Pi$, there exists a unique probability measure on H_∞, denoted by P_x^ν, which models the controlled dynamic system under ν. Its expectation operator is denoted by E_x^ν. From now on, we will use the notation

$$\omega = (x_0, a_0, \ldots, x_t, a_t, \ldots) \in H_\infty$$

to denote a sample path of the state-action process. Slightly abusing the notation, but without any risk of confusion, we will as well denote by x_t and a_t the projection mappings which associate to each $\omega \in H_\infty$ the corresponding t-th coordinates.

The expected discounted cost optimality criterion. We will consider a measurable *running cost* function $c : \mathbb{K} \to \mathbb{R}$ and a measurable function $\alpha : \mathbb{K} \to [0, 1]$ that will be interpreted as the *discount factor*. Given an initial state $x \in X$ and a control policy $\nu \in \Pi$, we define

$$J(x, \nu) = E_x^\nu\left[\sum_{t=0}^{\infty} c(x_t, a_t) \prod_{j=0}^{t-1} \alpha(x_j, a_j)\right] \tag{2.2}$$

(later, we will provide conditions ensuring that the above expectation is well defined). The optimal discounted cost function is then defined as

$$J^*(x) = \inf_{\nu\in\Pi} J(x, \nu) \quad \text{for } x \in X, \tag{2.3}$$

and we will say that a policy $\nu^* \in \Pi$ is optimal when $J^*(x) = J(x, \nu^*)$ for each $x \in X$.

We impose the following conditions on our control model.

Assumption 2.1. **(i)** The running cost function c is in $\mathbb{L}^+(\mathbb{K})$ and, in addition, there exists a constant $M > 0$ with $\inf_{a\in A(x)} c(x, a) \leq M$ for every $x \in X$.

(ii) The discount factor function α is in $\mathbb{L}^+(\mathbb{K})$ (by definition, α takes values in $[0, 1]$).

(iii) The kernel Q is weakly continuous, i.e, $u \in \mathbb{C}_b(X)$ implies $Qu \in \mathbb{C}_b(\mathbb{K})$.

(iv) The correspondence $x \mapsto A(x)$ is compact-valued and upper semicontinuous.

(v) The optimal discounted cost $J^*(x)$ is finite for every $x \in X$.

Next we make some comments on our assumptions.

Remark 2.2. **(a)** The condition in (i) is equivalent to the existence of some $f \in \mathbb{F}$ such that $\sup_{x\in X} c(x, f(x)) < \infty$. This assumption does not imply that the cost function c is bounded. Also, the cost function c being nonnegative, it is clear that $J(x, \nu)$ in (2.2) is well defined.

(b) Conditions (iii)-(iv) in Assumption 2.1 are standard in the literature. Later on we will provide sufficient conditions for (v) based on the data of the model. □

Given a measurable function $u \in \mathbb{M}^+(X)$, we define the function Tu on X taking values in $[0, \infty]$ as

$$Tu(x) = \inf_{a\in A(x)} \big\{c(x, a) + \alpha(x, a)Qu(x, a)\big\} \quad \text{for each } x \in X,$$

and we say that T admits a measurable selector at u when there exists some $f \in \mathbb{F}$ with

$$Tu(x) = c(x, f(x)) + \alpha(x, f(x))Qu(x, f(x)) \quad \text{for each } x \in X.$$

We also say that $u \in \mathbb{M}^+(X)$ is a solution of the *dynamic programming equation (DPE)* if it satisfies

$$u(x) = Tu(x) \quad \text{for each } x \in X. \tag{2.4}$$

The following lemma ensures the existence of a solution to the DPE (2.4) as well as a regularity property of this solution.

Lemma 2.3. *Under the Assumption 2.1, the following holds:*

(i) *If $u \in \mathbb{L}^+(X)$ then $Tu \in \overline{\mathbb{L}}^+(X)$ and T has a measurable selector at u. If $u \in \mathbb{L}_b^+(X)$ then $Tu \in \mathbb{L}_b^+(X)$.*

(ii) *Let $v_0 = \mathbf{0}$ and define $v_{k+1} = Tv_k$ for $k \geq 0$. Then $\{v_k\}$ converges pointwise and monotonically to some $v^* \in \mathbb{L}^+(X)$ with $v^* \leq J^*$.*

(iii) *The function v^* obtained in (ii) is a solution of the DPE* (2.4).

Proof. (i). Since $u \in \mathbb{L}^+(X)$, there exists a sequence $\{u_k\}$ in $\mathbb{C}_b^+(X)$ such that $u_k \uparrow u$; see Hernández-Lerma and Lasserre [7, Proposition A.2] or Aliprantis and Border [1, Theorem 3.13]. It is easy to check that $(x, a) \mapsto c(x, a) + \alpha(x, a)Qu_k(x, a)$ is in $\mathbb{L}^+(\mathbb{K})$ for each k, since it is obtained by product and sum of nonnegative lower semicontinuous functions. The latter functions increase, as $k \to \infty$, to $(x, a) \mapsto c(x, a) + \alpha(x, a)Qu(x, a)$ which, therefore, is in $\overline{\mathbb{L}}^+(\mathbb{K})$ (see, e.g., Aliprantis and Border [1, Lemma 2.41]). The fact that Tu is lower semicontinuous and that there exists a measurable selector for T at u follows from Proposition D.5 in Hernández-Lerma and Lasserre [7] (although that reference deals with functions taking values in $\mathbb{R}$, it can be easily adapted to functions taking values in $\mathbb{R} \cup \{\infty\}$). Finally, note that if u is bounded then so is Tu because $||Tu|| \leq M + ||u||$.

(ii). The sequence $\{v_k\}$ is in $\mathbb{L}_b^+(X)$ with $||v_k|| \leq kM$. Since $v_0 \leq v_1$ and noting that the operator T is monotone, we can see that $\{v_k\}$ is monotone nondecreasing. Hence, its pointwise limit v^* is a lower semicontinuous function in $\overline{\mathbb{L}}^+(X)$. Let us now prove that $v^* \leq J^*$. For each $k \geq 0$, let $f_k \in \mathbb{F}$ be a measurable selector of T at v_{k+1}, that is,

$$v_{k+1}(x) = c(x, f_k(x)) + \alpha(x, f_k(x))Qv_k(x, f_k(x)) \quad \text{for } x \in X. \tag{2.5}$$

In the sequel, fix some integer $n \geq 0$ and consider the decision times $\{0, \ldots, n\}$. Given $0 \leq t \leq n$, a policy $\nu \in \Pi$, and a history $h_t = (x_0, a_0, \ldots, x_t) \in H_t$, define

$$J_{t,n}(h_t, \nu) = E^{\nu}_{x_0}\Big[\sum_{k=t}^{n} c(x_k, a_k) \prod_{j=t}^{k-1} \alpha(x_j, a_j) \mid h_t\Big].$$

Notice that $J_{t,n}(h_t, \nu)$ depends on ν only through the decision made at times $t, \ldots, n$; that is, on $\{\nu_t, \ldots, \nu_n\}$. Also, let

$$J_{t,n}(h_t) = \inf_{\nu \in \Pi} J_{t,n}(h_t, \nu) \quad \text{for } h_t \in H_t.$$

Define the policy ν_n^* as $\{f_n, \ldots, f_0\}$ on the time horizon $0, \ldots, n$, with f_i as in (2.5), and arbitrarily from time $n+1$ onwards. Our goal now is to show that we have

$$J_{t,n}(h_t) = J_{t,n}(h_t, \nu_n^*) = v_{n+1-t}(x_t) \quad \text{for every } 0 \leq t \leq n \text{ and } h_t \in H_t. \tag{2.6}$$

We will prove it by backwards induction on $t = n, n-1, \ldots, 0$. This equality is obvious for $t = n$ because for every $\nu \in \Pi$ and $h_n \in H_n$ we have

$$J_{n,n}(h_n, \nu) = \int_{A(x_n)} c(x_n, a) \nu_n(da|h_n),$$

and so

$$J_{n,n}(h_n) = \min_{a \in A(x_n)} c(x_n, a) = c(x_n, f_0(x_n)) = J_{n,n}(h_n, \nu_n^*) = v_1(x_n).$$

Suppose that (2.6) holds for some $t+1$ and let us prove it for t. Given arbitrary $\nu \in \Pi$ and $h_t \in H_t$ we have that

$$\begin{aligned}
J_{t,n}(h_t, \nu) &= E^{\nu}_{x_0}\left[E^{\nu}_{x_0}\Big[\sum_{k=t}^{n} c(x_k, a_k) \prod_{j=t}^{k-1} \alpha(x_j, a_j) \mid h_{t+1}\Big] \mid h_t\right] \\
&= E^{\nu}_{x_0}\left[c(x_t, a_t) + \alpha(x_t, a_t) E^{\nu}_{x_0}\Big[\sum_{k=t+1}^{n} c(x_k, a_k) \prod_{j=t+1}^{k-1} \alpha(x_j, a_j) \mid h_{t+1}\Big] \mid h_t\right] \\
&= E^{\nu}_{x_0}\Big[c(x_t, a_t) + \alpha(x_t, a_t) J_{t+1,n}(h_{t+1}, \nu) \mid h_t\Big] \\
&\geq E^{\nu}_{x_0}\Big[c(x_t, a_t) + \alpha(x_t, a_t) v_{n-t}(x_{t+1}) \mid h_t\Big] \\
&= \int_{A(x_t)} [c(x_t, a) + \alpha(x_t, a) Q v_{n-t}(x_t, a)] \nu_t(da|h_t) \\
&\geq v_{n+1-t}(x_t),
\end{aligned}$$

with equality when $\nu = \nu_n^*$. This completes the backward induction argument. Hence, letting $t = 0$ (recall (2.6)) we have thus proved that for every $n \geq 0$ and $x \in X$

$$J_{0,n}(x) = J_{0,n}(x, \nu_n^*) = v_{n+1}(x).$$

Proceeding with the proof, the non negativity of the cost function implies that, for every $x \in X$, $n \geq 0$, and $\nu \in \Pi$,

$$v_{n+1}(x) \leq J_{0,n}(x, \nu) \leq J(x, \nu),$$

and so $v_{n+1}(x) \le J^*(x)$. This shows that the limit function $v^* \le J^*$.

(iii). The operator T being monotone, it is clear that $v_{k+1} = Tv_k \le Tv^*$, and so $v^* \le Tv^*$. To prove the reverse inequality, let $f_k \in \mathbb{F}$, for each $k \ge 0$, be a measurable selector of T at v_k (recall (2.5)); i.e.,

$$v_{k+1}(x) = c(x, f_k(x)) + \alpha(x, f_k(x))Qv_k(x, f_k(x)) \quad \text{for } x \in X. \tag{2.7}$$

For fixed $x \in X$, the sequence $\{f_k(x)\}$ in $A(x)$ has a convergent subsequence $f_{k_n}(x) \to a \in A(x)$. Fix some n' and observe that

$$\liminf_{n\to\infty} Qv_{k_n}(x, f_{k_n}(x)) \ge \liminf_{n\to\infty} Qv_{k_{n'}}(x, f_{k_n}(x)) \ge Qv_{k_{n'}}(x, a),$$

where the first inequality follows by monotonicity of $\{v_k\}$ and the second one by lower semicontinuity of $Qv_{k_{n'}}$ (recall part (i) of this proof). But n' being arbitrary, monotone convergence yields $\liminf_{n\to\infty} Qv_{k_n}(x, f_{k_n}(x)) \ge Qv^*(x, a)$. Take now the $\liminf$ in (2.7) through n' to obtain that $v^*(x) \ge c(x, a) + \alpha(x, a)Qv^*(x, a)$ from which the inequality $v^* \ge Tv^*$ follows. Hence, v^* is indeed a solution of the DPE (2.4). □

The next theorem relates the solution v^* of the DPE introduced in Lemma 2.3 with the optimal cost function J^* in (2.3). Furthermore, it guarantees the existence of optimal control policies.

Theorem 2.4. *Let Assumption 2.1 be satisfied and let $v^* \in \mathbb{L}^+(X)$ be as in Lemma 2.3(ii).*

(i) *The optimal discounted cost J^* equals v^*, and it is the minimal solution in $\mathbb{L}^+(X)$ of the DPE (2.4).*

(ii) *Any measurable selector of T at J^* is an optimal deterministic stationary policy.*

Proof. (i). Given any solution $u \in \mathbb{L}^+(X)$ of the DPE (2.4), let $f \in \mathbb{F}$ be a measurable selector of T at u, that is, with

$$u(x) = c(x, f(x)) + \alpha(x, f(x))Qu(x, f(x)) \quad \text{for each } x \in X.$$

Iteration of this inequality yields

$$\begin{aligned} u(x) &= E_x^f\Big[\sum_{k=0}^{n} c(x_k, f(x_k)) \prod_{t=0}^{k-1} \alpha(x_t, f(x_t))\Big] + E_x^f\Big[u(x_{n+1}) \prod_{t=0}^{n} \alpha(x_t, f(x_t))\Big] \\ &\ge E_x^f\Big[\sum_{k=0}^{n} c(x_k, f(x_k)) \prod_{t=0}^{k-1} \alpha(x_t, f(x_t))\Big] \end{aligned}$$

which, letting $n \to \infty$ and by monotone convergence, implies $u(x) \ge J(x, f) \ge J^*(x)$ for each $x \in X$. Since v^* in Lemma 2.3(ii) is indeed a solution of the DPE, we obtain $v^* \ge J^*$, and so $v^* = J^*$. This implies that J^* is indeed the minimal solution in $\mathbb{L}^+(X)$ of the DPE.

(ii). Use the argument in (i) replacing u with J^* to derive that any measurable selector of T at J^* is an optimal deterministic stationary policy. □

As discussed in Remark 2.2, all the conditions in Assumption 2.1 are stated in terms of the basic primitive data of the control model, except item (v). Next we explore some easily verifiable sufficient conditions for the finiteness of J^*. Our next assumption uses the following terminology; see Section 5.2 in Meyn and Tweedie [19].

Definition 2.5. Given $f \in \mathbb{F}$, we say that a set $C \in \mathcal{B}(X)$ is *small* if there exist $t \in \mathbb{N}$ and a nontrivial measure μ on $(X, \mathcal{B}(X))$ with $P_x^f\{x_t \in B\} \geq \mu(B)$ for all $B \in \mathcal{B}(X)$ and $x \in C$. We will also say that C is μ-small for $f \in \mathbb{F}$ at stage t.

We introduce some further notation. For $f \in \mathbb{F}$ and $0 \leq \beta \leq 1$, let

$$L_{f,\beta} = \{x \in X : \alpha(x, f(x)) \leq \beta\} \quad \text{and} \quad U_{f,\beta} = \{x \in X : \alpha(x, f(x)) > \beta\},$$

be the lower and upper sections of the discount factor function. We then impose the following assumption.

Assumption 2.6. **(i)** There exists $f \in \mathbb{F}$ such that $\sup_{x\in X} c(x, f(x)) = \mathfrak{c} < \infty$ and such that, for some $0 \leq \beta < 1$ and some nontrivial measure μ on $(X, \mathcal{B}(X))$, the set $U_{f,\beta}$ is μ-small for f at some stage n with, in addition, $\mu(L_{f,\beta}) > 0$.

(ii) There exist $0 < \delta < 1$ and $c_0 > 0$ such that

$$\forall (x, a) \in \mathbb{K},\ \alpha(x, a) \geq 1 - \delta \quad \text{implies} \quad c(x, a) \geq c_0.$$

Remark 2.7. **(a)** Under Assumption 2.6(i), the statement $\inf_{a\in A(x)} c(x, a) \leq M$ for $x \in X$ in Assumption 2.1(i) is necessarily satisfied (recall Remark 2.2(a)).

(b) The second part of Assumption 2.6(i) means that there exist $f \in \mathbb{F}$, $0 \leq \beta < 1$, $t \in \mathbb{N}$, and a non-trivial measure μ (which may depend on all these parameters) such that $\mu(L_{f,\beta}) = \mu(X \setminus U_{f,\beta}) > 0$ and

$$P_x^f\{x_t \in B\} \geq \mu(B) \quad \text{for every } B \in \mathcal{B}(X) \text{ and every } x \in U_{f,\beta}. \tag{2.8}$$

(c) The condition (ii) means, roughly speaking, that discount factors close to one imply a positive cost. It is a generalization of previous hypotheses existing in the literature; see, e.g., Assumption 2.4 in Jasso-Fuentes et al. [14].

Proposition 2.8. (i) *If Assumption 2.6(i) holds then* $\sup_{x\in X} J(x, f) < \infty$. *In particular, Assumption 2.1(v) is satisfied.*

(ii) *Under Assumption 2.6(ii), if* $x \in X$ *and* $\nu \in \Pi$ *are such that* $J(x, \nu) < \infty$ *then* $\prod_{i=0}^{k} \alpha(x_i, a_i)$ *converges to* 0 *with* P_x^ν*-probability one as* $k \to \infty$.

Proof. (i). We first prove a preliminary fact. By hypothesis, we can choose $0 < \epsilon < 1$ such that $\mu(L_{f,\beta}) \geq \epsilon$. Suppose that the initial state x is in $U_{f,\beta}$ and let $S = \min\{k > 0 : x_k \in L_{f,\beta}\}$. We have $\{S > n\} \subseteq \{x_n \in U_{f,\beta}\}$ and so, for any $x \in U_{f,\beta}$,

$$P_x^f\{S > n\} \leq P_x^f\{x_n \in U_{f,\beta}\} = 1 - P_x^f\{x_n \in L_{f,\beta}\} \leq 1 - \mu(L_{f,\beta}) \leq 1 - \epsilon.$$

By iteration of this argument, for each $r \geq 1$ we have

$$P_x^f\{S > rn\} \leq (1 - \epsilon)^r, \quad \text{and so} \quad E_x^f[S] \leq n/\epsilon. \tag{2.9}$$

Hence, the exit time from $U_{f,\beta}$ has finite expectation.

Suppose now that the initial state x is in $L_{f,\beta}$ and observe that

$$J(x, f) \leq \mathfrak{c}\ E_x^f\Big[\sum_{k=0}^{\infty} \prod_{j=0}^{k-1} \alpha(x_j, f(x_j))\Big].$$

Define, for each $r \geq 1$,

$$T_r = \min\{k > S_{r-1} : x_k \in U_{f,\beta}\} \quad \text{and} \quad S_r = \min\{k > T_r : x_k \in L_{f,\beta}\}$$

with the convention that $S_0 = 0$. Hence, the state x_k is in $L_{f,\beta}$ for k in the following union of intervals

$$[S_0, T_1) \cup [S_1, T_2) \cup [S_2, T_3) \cup \dots,$$

while the state x_k is in $U_{f,\beta}$ for k in

$$[T_1, S_1) \cup [T_2, S_2) \cup [T_3, S_3) \cup \dots.$$

Observe now the following. Suppose that $r \geq 0$ is given:

- If $j \geq 1$ is such that $S_r < j \leq T_{r+1}$, then $x_{j-1} \in L_{f,\beta}$ and so the current discount factor $\alpha(x_{j-1}, f(x_{j-1}))$ is less than β.
- If $j \geq 1$ is such that $T_r < j \leq S_r$, then $x_{j-1} \in U_{f,\beta}$ and so the current discount factor $\alpha(x_{j-1}, f(x_{j-1}))$ is larger than β, but bounded above by one.

Bearing this in mind, if $j \geq 1$ satisfies $S_r < j \leq T_{r+1}$ for some $r \geq 0$, then the accumulated (multiplicative) discount factor $\prod_{m=0}^{j-1} \alpha(x_m, f(x_m))$ is less than or equal to

$$\prod_{n=0}^{r-1} \Big[\prod_{m=S_n+1}^{T_{n+1}} \alpha(x_{m-1}, f(x_{m-1})) \prod_{m=T_{n+1}+1}^{S_{n+1}} \alpha(x_{m-1}, f(x_{m-1})) \Big] \prod_{m=S_r+1}^{j} \alpha(x_{m-1}, f(x_{m-1}))$$
$$\leq \prod_{n=0}^{r-1} \Big[\beta^{T_{n+1}-S_n} \cdot 1^{S_{n+1}-T_{n+1}} \Big] \cdot \beta^{j-S_r} \leq \beta^{j-S_r+\sum_{n=0}^{r-1}(T_{n+1}-S_n)}. \tag{2.10}$$

Arguing similarly, if $j \geq 1$ satisfies $T_r < j \leq S_r$ for some $r \geq 1$ then the accumulated (multiplicative) discount factor $\prod_{m=0}^{j-1} \alpha(x_m, f(x_m))$ is less than or equal to

$$\prod_{m=1}^{T_r} \alpha(x_{m-1}, f(x_{m-1})) \prod_{m=T_r+1}^{j} \alpha(x_{m-1}, f(x_{m-1})) \leq \beta^{\sum_{n=0}^{r-1}(T_{n+1}-S_n)},$$

where the first product has been bounded using (2.10) and the second product is bounded by the trivial bound 1. Hence, splitting the time horizon $\{1, 2, \dots\}$ in the intervals $(S_r, T_{r+1}]$ and $(T_r, S_r]$ and noting that the discount factor at the initial stage is 1, we can write

$$J(x, f) \leq \mathfrak{c} + \mathfrak{c} E_x^f \Big[\sum_{r=0}^{\infty} \sum_{j=S_r+1}^{T_{r+1}} \beta^{j-S_r+\sum_{n=0}^{r-1}(T_{n+1}-S_n)} \Big] + \mathfrak{c} \sum_{r=1}^{\infty} E_x^f \Big[(S_r - T_r) \beta^{\sum_{n=0}^{r-1}(T_{n+1}-S_n)} \Big].$$

Note that

$$\sum_{r=0}^{\infty} \sum_{j=S_r+1}^{T_{r+1}} \beta^{j-S_r+\sum_{n=0}^{r-1}(T_{n+1}-S_n)} = \sum_{k=1}^{\infty} \beta^k = \frac{\beta}{1-\beta}.$$

Due to the stationarity of the process, together with (2.9) and $\sum_{n=0}^{r-1}(T_{n+1} - S_n) \geq r$, we can deduce

$$E_x^f \Big[(S_r - T_r) \beta^{\sum_{n=0}^{r-1}(T_{n+1}-S_n)} \Big] \leq \beta^r E_x^f \Big[E\big[(S_r - T_r) | h_{T_r} \big] \Big] \leq \beta^r (n/\epsilon).$$

This shows that whenever $x \in L_{f,\beta}$ we have

$$J(x,f) \le \mathfrak{c}\beta \, \frac{1+n/\epsilon}{1-\beta}.$$

Arguing similarly, it can be shown that the same bound holds for an initial state $x \in U_{f,\beta}$. This completes the proof that $\sup_{x\in X} J(x,f)$ is finite. In particular, it follows that J^* is bounded.

(ii). Observe that, for each fixed $\omega \in H_\infty$, the sequence $\prod_{i=0}^{k-1} \alpha(x_i,a_i)$ is monotone non-increasing and, hence, it converges to some $\eta(\omega) \in [0,1]$. Our goal now is to prove that

$$\eta(\omega) > 0 \implies \sum_{k=0}^{\infty} c(x_k,a_k) \prod_{i=0}^{k-1} \alpha(x_i,a_i) = \infty.$$

Taking $0 < \delta < 1$ as in Assumption 2.6(ii), if $\alpha(x_k,a_k) < 1-\delta$ for infinitely many k then the limit $\eta(\omega)$ is necessarily 0. Hence, if $\eta(\omega) > 0$ then there exists some k_0 such that $k \ge k_0$ implies $\alpha(x_k,a_k) \ge 1-\delta$ and, hence, by Assumption 2.6(ii), $c(x_k,a_k) \ge c_0$. So,

$$\sum_{k=0}^{\infty} c(x_k,a_k) \prod_{i=0}^{k-1} \alpha(x_i,a_i) \ge \sum_{k=k_0}^{\infty} c(x_k,a_k) \prod_{i=0}^{k-1} \alpha(x_i,a_i) \ge \sum_{k=k_0}^{\infty} c_0\eta = \infty.$$

Therefore, if

$$J(x,\nu) = E_x^\nu\Big[\sum_{k=0}^{\infty} c(x_k,a_k) \prod_{i=0}^{k-1} \alpha(x_i,a_i)\Big]$$

is finite for some policy $\nu \in \Pi$, then the set of ω for which the limit $\eta(\omega)$ is positive must necessarily have P_x^ν-probability zero (otherwise, $J(x,\nu)$ would be infinite); hence, with P_x^ν-probability one, $\lim_{k\to\infty} \prod_{i=0}^{k-1} \alpha(x_i,a_i) = 0$. □

The next theorem shows an important regularity property of the value function J^* along with uniqueness of the solution of the DPE (2.4).

Theorem 2.9. *If Assumptions 2.1 and 2.6 hold then $J^* \in \mathbb{L}_b^+(X)$. Furthermore it is the unique solution in $\mathbb{L}_b^+(X)$ of the DPE* (2.4)*.*

Proof. We know from Proposition 2.8(i) and Theorem 2.4 that $J^* \in \mathbb{L}_b^+(X)$. Suppose now that u is a nonnegative and bounded solution of the DPE (2.4). By Theorem 2.4(i) we have $u \ge J^*$. For each $(x,a) \in \mathbb{K}$ we have $u(x) \le c(x,a) + \alpha(x,a)Qu(x,a)$ and, in particular, given $\nu \in \Pi$, an initial state $x \in X$, and $n \ge 0$, we have, for any $h_n \in H_n$ and $a_n \in A(x_n)$,

$$\begin{aligned}
&E_x^\nu\Big[u(x_{n+1}) \prod_{i=0}^{n} \alpha(x_i,a_i) \mid h_n, a_n\Big] = Qu(x_n,a_n) \prod_{i=0}^{n} \alpha(x_i,a_i) \\
&= \prod_{i=0}^{n-1} \alpha(x_i,a_i)\Big[c(x_n,a_n) + \alpha(x_n,a_n)Qu(x_n,a_n) - c(x_n,a_n)\Big] \\
&\ge \prod_{i=0}^{n-1} \alpha(x_i,a_i)\Big[u(x_n) - c(x_n,a_n)\Big].
\end{aligned}$$

Taking expectation yields

$$E_x^\nu\Big[c(x_n,a_n)\prod_{i=0}^{n-1}\alpha(x_i,a_i)\Big] \geq E_x^\nu\Big[u(x_n)\prod_{i=0}^{n-1}\alpha(x_i,a_i) - u(x_{n+1})\prod_{i=0}^{n}\alpha(x_i,a_i)\Big].$$

Summing up these inequalities over the indexes $0,\ldots,k$ gives

$$E_x^\nu\Big[\sum_{n=0}^{k} c(x_n,a_n)\prod_{i=0}^{n-1}\alpha(x_i,a_i)\Big] \geq u(x) - E_x^\nu\Big[u(x_{k+1})\prod_{i=0}^{k}\alpha(x_i,a_i)\Big]. \tag{2.11}$$

Suppose now that $\nu = f \in \mathbb{F}$ is a measurable selector of T at J^*, in the DPE $J^* = TJ^*$. We can apply Proposition 2.8(ii) because $J(x,f) = J^*(x) < \infty$, together with dominated convergence, to establish that

$$\lim_{k\to\infty} E_x^f\Big[u(x_{k+1})\prod_{i=0}^{k}\alpha(x_i,a_i)\Big] = 0$$

(here, we also use the fact that u is bounded). It follows from (2.11) by taking the limit as $k\to\infty$ that $J^*(x) = J(x,f) \geq u(x)$. This completes the proof that $u = J^*$. □

3 Markov decision processes with stopping

In this section we describe control models which incorporate some kind of "stopping", either by means of an action that stops (or kills) the process, or by a natural extinction of the process. Such stopping models are relevant in applications (we can quote the references: Bensoussan [3], Puterman [22], Ross [24], as well as the papers: Rieder [23], Dufour and Piunovskiy [5], Horiguchi [11, 12], among others). Interestingly, such stopping models fit in the previously described framework of MDPs with varying discount factor, as we shall see.

A simplified version of a stopping MDP with constant discount factor is as follows. Consider a control model $\mathfrak{M}$ as in (2.1) with constant discount factor $0 < \alpha < 1$. In addition to the control policy $\nu \in \Pi$, the decision-maker is allowed to stop the process. This is modeled by a random variable $\tau : \Omega \to \{0,1,2,\ldots,\infty\}$ such that $\{\tau = t\}$ is $(x_0,a_0,\ldots,x_t,a_t)$-measurable. The corresponding total expected discounted cost is

$$J(x,\nu,\tau) = E_x^\nu\Big[\sum_{t=0}^{\tau-1}\alpha^t c(x_t,a_t) + \alpha^\tau \ell(x_\tau)\Big], \tag{3.1}$$

with ℓ some measurable stopping cost function on X (if $\tau = \infty$ then the righthand term vanishes and the lefthand term is an infinite sum). The objective of the decision-maker is to minimize the performance criterion (3.1) in ν and τ. As we shall see later, the criterion (3.1) is a particular case of the control model introduced in Section 2.

Moreover, since our model allows the discount factor to be equal to one at some stages, we can handle further generalizations of MDPs by considering the larger class of so-named hybrid control models (which allow for instantaneous transitions; see, e.g., Jasso-Fuentes et al. [14, 15]) combined with stopping. To the best of our knowledge, this issue has not been studied in the literature (indeed, the above cited references have studied different versions of stopping problems but none of them allows for instantaneous transitions, i.e., a hybrid dynamics).

In what follows, we shall distinguish two "sources" of stopping: by means of a stopping action or by means of absorption.

Stopping action. We give the following definition of a stopping action in the context of our MDP model $\mathfrak{M}$.

Definition 3.1. Let $\mathfrak{M} = (X, A, \mathbb{K}, Q, c, \alpha)$ be the general control model described in (2.1). Given $x \in X$, we say that $a \in A(x)$ is a *stopping action* (at the state x) when $\alpha(x, a) = 0$.

With this definition in mind, it is clear that, for $x \in X$ and $\nu \in \Pi$, the total expected discounted reward can be written

$$J(x, \nu) = E_x^\nu\Big[\sum_{t=0}^{\infty} c(x_t, a_t) \prod_{j=0}^{t-1} \alpha(x_j, a_j)\Big] = E_x^\nu\Big[\sum_{t=0}^{\tau} c(x_t, a_t) \prod_{j=0}^{t-1} \alpha(x_j, a_j)\Big]$$

where we define τ on H_∞ as the time a stopping action is taken, that is,

$$\tau(\omega) = \min\{t \geq 0 : \alpha(x_t, a_t) = 0\}.$$

Note that, in the context of our control model $\mathfrak{M}$, although the policy ν chosen by the decision-maker does not formally kill the process (in fact, the state process continues its evolution), all the costs from time $\tau + 1$ onwards are irrelevant: the discount factor that vanishes has indeed produced the effect of stopping the process.

In case that $\alpha(x, a) > 0$ for all $(x, a) \in \mathbb{K}$ we can artificially add a stopping action $\mathfrak{a}$ to the action set, so that the augmented set of available actions at $x \in X$ are $A(x) \cup \{\mathfrak{a}\}$ with corresponding discount factor $\alpha(x, \mathfrak{a}) = 0$ and cost function given by the stopping cost ℓ, that is, $c(x, \mathfrak{a}) = \ell(x)$ for each $x \in X$. In this case, the decision-maker can stop the process at any point in time by taking the stopping action $\mathfrak{a}$. Letting $\tau(\omega) = \min\{t \geq 0 : a_t = \mathfrak{a}\}$ we have

$$J(x, \nu) = E_x^\nu\Big[\sum_{t=0}^{\tau-1} c(x_t, a_t) \prod_{j=0}^{t-1} \alpha(x_j, a_j) + \ell(x_\tau) \prod_{j=0}^{\tau-1} \alpha(x_j, a_j)\Big].$$

In this case, the DPE equation (2.4) $J^* = TJ^*$ becomes

$$J^*(x) = \ell(x) \wedge \min_{a \in A(x)} \Big\{c(x, a) + \alpha(x, a) \int_X J^*(y) Q(dy|x, a)\Big\} \quad \text{for } x \in X.$$

It is clear now that the criterion in (3.1) is indeed a particular case of our control model with varying discount factor.

We define the *contact set* as

$$D^* = \{x \in X : \ell(x) = J^*(x)\}.$$

It is the subset of the state space X on which it is optimal for the decision-maker to stop the dynamics; that is, when $x_t \in D^*$ (the process hits D^*) then the optimal action is $\mathfrak{a}$ and the process is stopped; otherwise, when $x_t \notin D^*$, the optimal action is an action in $A(x)$ for which the natural dynamics of the process continues. The optimal stopping time is then $\tau^* = \min\{t \geq 0 : x_t \in D^*\}$.

Absorption state and absorption action. Since the previous "stopping action" is not possible when the discount factor is kept constant, we explore a different source of stopping, which is produced when the process reaches some special states, but not necessarily when the decision-maker stops the process, as seen before.

Definition 3.2. Consider the general control model $\mathfrak{M} = (X, A, \mathbb{K}, Q, c, \alpha)$ as in (2.1).

(i) We say that $\partial \in X$ is an *absorption state* when $Q(\cdot|\partial, a) = \delta_\partial(\cdot)$, for any $a \in A(\partial)$.

(ii) If ∂ is an absorption state and $x \in X$ is not, then we say that $a \in A(x)$ is a *possible absorption action* when $Q(\{\partial\}|x, a) > 0$, and a *full absorption action* (or simply *absorption action*) when $Q(\{\partial\}|x, a) = 1$, and in this case, the absorption action a is denoted by a_∂, without ambiguity. Moreover, this absorption action a_∂ is called *absorption-stopping action* whenever $c(\partial, a) = 0$, for any $a \in A(\partial)$.

We make some remarks on this definition.

Remark 3.3. (a) Note that absorption states and (possible) absorption actions can be part of the model $\mathfrak{M}$ itself or, alternatively, they can be artificially added for modeling purposes.

(b) The notion of an absorption action is closely related to a stopping action in the previous paragraph. Indeed, for any absorption state ∂ it may be convenient (but not necessary) to add the condition $c(\partial, a) = 0$ for every $a \in A(\partial)$, so that "absorption action" and "absorption-stopping action" are the same concepts; otherwise, the cost may become infinite. In this case, the absorption(-stopping) action a_∂ effectively produces a stopping action, i.e., transitions are stopped (in the sense that the state remains constant) and costs cease (in the sense that costs vanish) immediately after applying a_∂. In this case, for any policy $\nu \in \Pi$, the stopping time $\tau = \inf\{t \geq 0 : a_t = a_\partial\}$ plays the role of τ in (3.1). Also note that absorption states have to do with all the transitions thereafter, while (possible) absorption actions have to do only with the immediate transition following the current state.

For instance, assuming that there exists only one absorption state ∂, we can define the random variable τ on H_∞ as the lifetime of the process, namely,

$$\tau_\partial = \inf\{t \geq 0 : x_t = \partial\}.$$

Hence, if the decision-maker uses a policy $\nu \in \Pi$ and the process reaches state ∂, the process gets trapped thereafter in ∂ with no subsequent cost, assuming $c(\partial, a) = 0$ for every $a \in A(\partial)$. In this context, stopping of the process occurs as a natural phenomenon, which is not forced by the decision-maker. Indeed, when taking an absorption action there is a possibility of killing the process, but it is not necessarily killed with certainty. In this case, the corresponding total expected discounted cost becomes

$$J(x, \nu) = E_x^\nu\Big[\sum_{t=0}^{\infty} c(x_t, a_t) \prod_{j=0}^{t-1} \alpha(x_j, a_j)\Big] = E_x^\nu\Big[\sum_{t=0}^{\tau_\partial - 1} c(x_t, a_t) \prod_{j=0}^{t-1} \alpha(x_j, a_j)\Big],$$

and the corresponding optimality equation reads

$$J^*(x) = \min_{a \in A(x)} \Big\{c(x, a) + \alpha(x, a) \int_X J^*(y) Q(dy|x, a)\Big\} \quad \text{for } x \in X \setminus \{\partial\},$$

and $J^*(\partial) = 0$. Once again, we see that our general model $\mathfrak{M}$ can handle situations when natural absorption (or killing) of a controlled process occurs.

4 Applications and examples

In this section we describe how a switching MDP with stopping can be described as a particular case of our model in this paper. We also make a detailed analysis of a practical example: a stock pollution control problem.

Switching MDPs with stopping. We consider here a switching control model with stopping, and we show how it fits into the framework of our MDPs with varying discount factor studied so far.

In switching control models with stopping, the dynamic system is subject to *changes of modes*, or *configurations*. The decision-maker can choose the times when the dynamics switches from one mode to another, while between switching times the dynamics behaves as a usual Markov decision process. Such models have been studied for different type of dynamics (mostly in continuous-time) when control is only applied to switching between modes; see, for instance, Bensoussan and Lions [6], and Menaldi and Blankenship [18] among others.

In our context, there are N controlled Markov chains, labeled $i = 1, \ldots, N$, all taking values in the state space Y and with common action space V; both Y and V are assumed to be Borel spaces. The corresponding transition kernels are written $Q_i(B \mid y, a)$, for $B \in \mathcal{B}(Y)$ and $(y, a) \in Y \times A$, for each $i = 1, \ldots, N$.

The controlled switching model with stopping is then defined by the following elements. The state space is

$$X = Y \times \{1, \ldots, N\} \cup \{\partial\},$$

where ∂ is an isolated absorption state. In the pair $(y, k) \in X$, we interpret $y \in Y$ as the fast variable, representing the state of the system and $k \in \{1, \ldots, N\}$ as the slow variable indicating the current mode of the system. The action space consists of $V \cup \{a_\partial\} \cup \{1, \ldots, N\}$, the latter being $N + 1$ isolated points. The set of available actions are

$$A(y, k) = V \cup \{a_\partial\} \cup \{1, \ldots, k-1, k+1, \ldots, N\}$$

for a state $(y, k) \in Y \times \{1, \ldots, N\}$ and $A(\partial) = \{a_\partial\}$, i.e., a_∂ is an absorption(-stopping) action. The interpretation is that, in state $(y, k) \in X$, the decision-maker can take either a usual action $a \in V$, or switch to a mode $1 \leq a \leq N$ different from the current mode k, or stop the process through the action $a = a_\partial$.

The dynamics of the model is defined as follows. Starting from a state $(y, k) \in X$ we have

$$Q(dz \times dm \mid (y, k), a) = \begin{cases} Q_k(dz \times dm \mid y, a) & \text{if } a \in V \\ \delta_{(y,a)}(dz \times dm) & \text{if } 1 \leq a \leq N \text{ with } a \neq k. \\ \delta_\partial(dz) & \text{if } a = a_\partial, \end{cases}$$

and note that $Q(\{\partial\}|(y, k), a) = 0$ for any $(y, k) \in Y \times \{1, \ldots, N\}$ and $a \in A(y, k)$, while $Q(\{\partial\}|\partial, a_\partial) = 1$. The dynamics can be as well stated in terms of an explicit transition function: starting from the state (y_t, k_t) at time $t \geq 0$ and taking an action a_t yields:

$$\underbrace{(y_{t+1}, k_{t+1}) = F(y_t, k_t, a_t, w_t)}_{\textit{standard sub-dynamic}} \quad \text{when } a_t \in V, \text{ or}$$

$$\underbrace{(y_{t+1}, k_{t+1}) = (y_t, a_t)}_{\textit{switching sub-dynamic}} \quad \text{when } 1 \leq a_t \leq N \text{ with } a \neq k_t,$$

and $y_{t+1} = \partial$ whenever $a_t = a_\partial$; here, $\{w_t\}$ is a sequence of i.i.d. random variables. This characterization of the dynamics of an MDP can be proved in a general setting, as for instance in Proposition 8.6 in Kallenberg [17]. A typical or common situation is when modes remain constant except when a switching is applied, namely, when

$$Q(dz \times dm \mid (y,k),a) = Q_k(dz \mid y,a)\,\delta_{\{k\}}(dm) \quad \text{if } a \in V$$

or equivalently

$$\underbrace{(y_{t+1},k_{t+1}) = \big(F(y_t,k_t,a_t,w_t),k_t\big)}_{\textit{standard sub-dynamic}} \quad \text{when } a_t \in V$$

are satisfied.

We consider the following running cost functions: the cost functions $\ell_i : Y \times V \to \mathbb{R}^+$ are the cost functions under each of the N controlled Markov chains, for $i = 1,\dots,N$. There is a cost for switching given by $l : \{1,\dots,N\} \times \{1,\dots,N\} \mapsto \mathbb{R}^+$ and, finally, a terminal cost function $\ell_0 : Y \to \mathbb{R}^+$. Each of these functions satisfies Assumptions 2.1(i) in their respective domains. The cost function c for a state $(y,k) \in Y \times \{1,\dots,N\}$ is given by

$$c(y,k,a) = \ell_k(y,a)\mathbf{1}_V(a) + l(k,a)\mathbf{1}_{\{1,\dots,N\}}(a) + \ell_0(y)\mathbf{1}_{\{a_\partial\}}(a),$$

and $c(\partial, a_\partial) = 0$. On the other hand, the discount factor is given by

$$\alpha(y,k,a) = \varrho_k(y)\mathbf{1}_V(a) + \mathbf{1}_{\{1,\dots,N\}}(a)$$

for $(y,k) \in Y \times \{1,\dots,N\}$ and $\alpha(\cdot, a_\partial) = 0$, where $\varrho_k : Y \to (0,1)$, for $1 \le k \le N$, are the discount functions of each of the N controlled Markov chains. They are supposed to satisfy Assumption 2.1(ii).

Finally, the set of admissible control policies is the same as Π (see Section 2), but using $A(x)$ introduced in previous paragraphs.

Remark 4.1. Observe that, once the state ∂ is reached, the process remains in this state thereafter, without any further cost. Essentially, the process has been stopped. Note also that the discount factor of an action in $\{1,\dots,N\}$ equals one: this last fact can be interpreted as an undiscounted instantaneous transition.

The function $u : Y \times \{1,\dots,N\} \to \mathbb{R}^+$ is a solution of the dynamic programming equation if for $(y,k) \in Y \times \{1,\dots,N\}$ we have

$$u(y,k) = \min\big\{\mathcal{M}u(y,k), \mathcal{H}u(y,k), \ell_0(y)\big\}$$

where the operators $\mathcal{H}$ and $\mathcal{M}$ are defined as

$$\mathcal{H}u(y,k) := \inf_{a\in V}\Big\{\ell_k(y,a) + \varrho_k(y)\int_X u(z,k)Q_k(dz|y,k,a)\Big\},$$
$$\mathcal{M}u(y,k) := \min_{a\in\{1,\dots,N\}\setminus\{k\}}\big\{l(k,a) + u(y,a)\big\}.$$

We put $u(\partial) = 0$.

We conclude this section by illustrating how switching MDPs can be applied to a control problem of practical interest.

A pollution accumulation problem. This problem has been studied for both discrete-time and continuous-time models; see, for instance, Morimoto [21], Jasso-Fuentes and Yin [16], Jasso-Fuentes and López-Barrientos [13], among others. With our varying discount factor control models, we can consider an additional feature in such problems, namely, the level of environmental contingency (implemented by the government) can be seen as an action taken by the decision-maker. Such an approach has not been considered so far in the literature.

Suppose that an economy consumes a specific good and that, as a byproduct of this consumption, it generates pollution. The pollution stock at time $t \geq 0$ is y_t, while k_t, taking values in $\{1, 2, \ldots, N\}$, denotes the level or mode of environmental contingency decided by the government at time t. They obey the following recurrent equation: for a fixed initial condition (y_0, k_0) we have

$$y_{t+1} = y_t - g(k_t)y_t + p(\mathsf{c}_t) + \zeta_t \quad \text{for } t \geq 0$$

where (y_t, k_t) is the state, and k_t is also the switching control with c_t being the usual control. The interpretation is that the stock of pollution $y_{t+1} \geq 0$ is obtained from:

- The stock $y_t \in \mathbb{R}^+$ of pollution at the previous time.
- The decay rate of pollution associated to the mode $i \in \{1, 2, \ldots, N\}$ is given by $0 \leq g(i) < 1$. Such a rate represents, e.g., any governmental actions leading to a decrease of pollution (related to the level of environmental contingency), natural cleaning of pollution (winds, rains), etc. At time $t \geq 0$, the corresponding rate is given by $g(k_t)$.
- The quantity $\mathsf{c}_t \geq 0$ denotes the consumption rate at time t, with range in $[0, \gamma(k_t)]$, where the upper bound of the interval is given by a constant $\gamma(i)$, for $1 \leq i \leq N$, depending on the contingency level. Usually, such bound is imposed by national or worldwide protocols. Finally, $p(\mathsf{c}_t)$ is the amount of pollution derived from a consumption equal to c_t, for some function $p : \mathbb{R}^+ \to \mathbb{R}^+$.
- The random variables ζ_t, assumed to be nonnegative and i.i.d., are random disturbances modeling external or unpredictable events. Let $F_\zeta : \mathbb{R}^+ \to [0, 1]$ be the corresponding distribution function.

A pollution stock of y in mode k, combined with a level of consumption equal to c, produces a disutility of $\ell_k(y, \mathsf{c})$. It is natural, hence, to assume that the functions $\ell_k : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$, for $1 \leq k \leq N$, are increasing in y (for fixed c) and decreasing in c (for fixed y). The government is allowed to switch the modes of environmental contingency: such changes are instantaneous in time and they produce a cost equal to $l(j, k) > 0$ when switching from mode j to mode k. Concerning the discounting, each mode $k \in \{1, \ldots, N\}$ is associated to a discount factor $0 < \rho_k < 1$. The fact that discounting varies with k can be interpreted as follows: a strict environmental protocol may have a positive future impact, modeled by a small discount factor ρ_k, meaning that future disutility will be smaller; on the contrary, a permissive environmental protocol can have a negative future impact, thus corresponding to a discount factor ρ_k closer to 1, meaning a lesser diminution of future disutility.

Formally, the control model is given by the following elements. The state space is $X = \mathbb{R}^+ \times \{1, \ldots, N\}$ and the action space is $A = \mathbb{R}^+ \cup \{1, \ldots, N\}$. Here, we should interpret $1, \ldots, N$ as (isolated) labels, rather than real numbers. The admissible actions at state $(y, k) \in X$ are

$$A(y, k) = \big[0, \gamma(k)\big] \cup \{1, \ldots, k-1, k+1, \ldots, N\},$$

modeling either a consumption level in the interval $[0, \gamma(k)]$ or a change in the environmental protocol from k to a different one.

The dynamic equations are, starting from the state (y_t, k_t) at time $t \geq 0$,

$$\underbrace{\big(y_{t+1}, k_{t+1}\big) = \big(y_t - g(k_t)y_t + p(a_t) + \zeta_t, k_t\big)}_{\textit{standard sub-dynamic}} \quad \text{when } a_t \in [0, \gamma(k_t)]$$

$$\underbrace{\big(y_{t+1}, k_{t+1}\big) = \big(y_t, a_t\big)}_{\textit{switching sub-dynamic}} \quad \text{when } a_t \in \{1, \dots, N\} \text{ with } a_t \neq k_t.$$

The cost function is given by

$$c(y, k, a) = \ell_k(y, a)\mathbf{1}_{\mathbb{R}^+}(a) + l(k, a)\mathbf{1}_{\{1,\dots,N\}}(a)$$

and the discount factor function is

$$\alpha(y, k, a) = \rho_k \mathbf{1}_{\mathbb{R}^+}(a) + \mathbf{1}_{\{1,\dots,N\}}(a)$$

for any $(y, k) \in X$ and $a \in A(y, k)$.

Given a control policy $\nu \in \Pi$, its total expected discounted disutility is (recall (2.2))

$$J((y, k), \nu) = E^{\nu}_{(y,k)}\bigg[\sum_{t=0}^{\infty} c(y_t, k_t, a_t) \prod_{j=0}^{t-1} \alpha(x_j, k_j, a_j)\bigg]$$

for an initial state $(y, k) \in X$. The objective is to find an optimal $\nu \in \Pi$ prescribing a consumption-switching policy yielding the minimal total expected discounted disutility:

$$J^*(y, k) = \inf_{\nu \in \Pi} J((y, k), \nu) \quad \text{for } (y, k) \in X.$$

The dynamic programming equation takes the form

$$J(y,k) = \min_{0 \leq a \leq \gamma(k)} \Big\{\ell_k(y,a) + \rho_k \int_0^{\infty} J\big(y - g(k)y + p(a) + z, k\big) F_\zeta(dz)\Big\} \wedge \min_{j \neq k} \Big\{l(k,j) + J(y,j)\Big\}.$$

We suppose that the pollution accumulation model satisfies the following conditions.

Assumption 4.2. **(a)** The mappings p and ℓ_k, for $1 \leq k \leq N$, are continuous on their respective domains.

(b) The functions $y \mapsto \ell_k(y, 0)$ have sub-linear growth, i.e., for some constants $A, B > 0$

$$\ell_k(y, 0) \leq Ay + B \quad \text{for each } y \geq 0 \text{ and } 1 \leq k \leq N,$$

and the random variables ζ_t have finite first moment: $\int z F_\zeta(dz) < \infty$.

(c) For each $k \in \{1, \dots, N\}$ the function $y \mapsto \ell_k(y, \gamma(k))$ is bounded on $\mathbb{R}^+$.

We have the next result.

Proposition 4.3. (i) *Under Assumptions 4.2(a) and (b), the optimal discounted cost function J^* is the minimal solution in $\mathbb{L}^+(X)$ of the dynamic programming equation.*

(ii) *Under Assumptions 4.2(a) and (c), the optimal discounted cost function J^* is the unique solution in $\mathbb{L}_b^+(X)$ of the dynamic programming equation.*

In either case, there exists an optimal deterministic stationary policy in $\mathbb{F}$ and it can be obtained as a measurable selector of the DP equation at J^*.

Proof. (i). We make the proof under Assumptions 4.2(a)-(b). Clearly, Assumptions 2.1(i)–(iv) hold (in particular, a policy that switches mode at any time, i.e., takes actions in $\{1,\dots,N\}$, satisfies the condition in 2.1(i)). Also, note that item (iii) in Assumption 2.1 directly follows from the dominated convergence theorem. It remains to verify (v) in this same assumption. Let $f_0 \in \mathbb{F}$ be the policy that prescribes a consumption equal to 0 at each stage. Starting from the initial state (y_0, k_0), it is clear by an induction argument that

$$\begin{aligned} y_{t+1} &= y_t - g(k_0)y_t + p(0) + \zeta_t \\ &\le y_t + p(0) + \zeta_t \\ &\le y_0 + (t+1)p(0) + \sum_{j=0}^{t} \zeta_j. \end{aligned}$$

Hence, for each $t \ge 0$ we have

$$E^{f_0}_{(y_0,k_0)}[y_t] \le y_0 + t\big(p(0) + E[\zeta]\big),$$

where $E[\zeta]$ represents the finite expectation of the random disturbances. Using the sub-linear growth condition on $y \mapsto \ell_{k_0}(y,0)$ we obtain

$$E^{f_0}_{(y_0,k_0)}[\ell_{k_0}(y_t,0)] \le E^{f_0}_{(y_0,k_0)}[Ay_t + B] \le Ay_0 + B + At\big(p(0) + E[\zeta]\big).$$

The discount factor being equal to ρ_{k_0} at each stage $t \ge 0$, it follows that

$$J((y_0,k_0),f_0) = \sum_{t=0}^{\infty} \rho_{k_0}^t E^{f_0}_{(y_0,k_0)}[\ell_{k_0}(y_t,0)] \le \frac{Ay_0 + B}{1-\rho_{k_0}} + \frac{A\rho_{k_0}\big(p(0)+E[\zeta]\big)}{(1-\rho_{k_0})^2}.$$

Hence, for any initial state $(y_0,k_0) \in X$ we have that $J^*(y_0,k_0)$ is indeed finite. The stated result is now a consequence of Theorem 2.4.

(ii). We now make the proof under Assumptions 4.2(a) and (c). Clearly, Assumptions 2.1(i)–(iv) hold, and we are going to show that Assumption 2.6 is verified. Consider the following policy $f^* \in \mathbb{F}$: if the state (y,k) is such that:

- $k = N$ or $y > 0$ then let $f^*(y,k) = \gamma(k)$ (i.e., the maximal consumption);
- if $y = 0$ and $1 \le k < N$ then $f^*(0,k) = N$ (i.e, switch to mode N).

Note that when $k = N$ or $y > 0$ we have $c((y,k), f^*(y,k)) = \ell_k(y,\gamma(k))$ which is bounded by hypothesis, and thus $\sup_{x\in X} c(x, f^*(x))$ is indeed finite. Now, let

$$\beta = \max_{1\le k\le N} \rho_k < 1 \quad \text{so that} \quad U_{f^*,\beta} = \{0\} \times \{1,\dots,N-1\}.$$

If the initial state of the system is $(y_0,k_0) = (0,k_0) \in U_{f^*,\beta}$ then, after one transition, the state of the system is $(y_1,k_1) = (0,N)$. Consider the measure μ on X given by $\mu = \delta_{(0,N)}$. It

is clear that μ is the probability distribution of (y_1, k_1) under $P^{f^*}_{(y_0,k_0)}$. So, it is indeed true that for every $(y_0, k_0) \in U_{f^*,\beta}$ and every $B \in \mathcal{B}(X)$ we have that (cf. (2.8))

$$P^{f^*}_{(y_0,k_0)}\{(y_1, k_1) \in B\} \geq \mu(B),$$

that is, the set $U_{f^*,\beta}$ is μ-small for f^* at stage $t = 1$. In addition, we also have that $\mu(L_{f^*,\beta}) = 1$ because $(0, N) \in L_{f^*,\beta}$. Hence, the condition in Assumption 2.6(i) holds, while it is straightforward to check Assumption 2.6(ii). The stated result is now a direct consequence of Theorem 2.9. □

It is worth noting that our hypotheses in Assumption 4.2 are not restrictive, in general, and they allow for unbounded cost functions c. Under Assumption 4.2(b), the optimal cost J^* is not necessarily bounded, but we know that it is, at most, of linear growth: for some constants $\mathbf{A}, \mathbf{B} > 0$ we have

$$J^*(y, k) \leq \mathbf{A}y + \mathbf{B}$$

for every $y \geq 0$ and $1 \leq k \leq N$. An example of an unbounded cost function satifying our conditions would be

$$\ell_k(y, \mathsf{c}) = \epsilon_k + D_k \log(1 + y) F_k(\mathsf{c})$$

for some constants $\epsilon_k, D_k > 0$ and some positive decreasing function F_k of c.

On the other hand, Assumption 4.2(c) imposes that the disutility for the maximal consumption is bounded when the level of pollution varies. In that case, we can prove that the value function J^* is bounded. An example of an unbounded cost function for which our conditions hold would be

$$\ell_k(y, \mathsf{c}) = \epsilon_k + D_k \log(1 + y)(\gamma(k) - \mathsf{c})$$

for some constants $\epsilon_k, D_k > 0$. The above cost functions, in accordance with the discussion at the beginning of the example, are increasing in y for fixed c and decreasing in c for fixed y.

References

[1] Aliprantis, C. D., Border, K.C.: *Infinite dimensional analysis.* Springer-Verlag, New York, 2006.

[2] Bäuerle, N., Rieder, U.: *Markov decision processes with application to finance.* Springer. Berlin Heidelberg, 2011.

[3] Bensoussan, A.: *Dynamic programming and inventory control.* IOS Press, Amsterdam, 2011.

[4] Bertsekas, D.P.: *Dynamic programming and stochastic control.* Academic Press, New York, 1976.

[5] Dufour,F., Piunovskiy, A.: (2010) Multiobjective stopping problem for discrete-time Markov processes: conves analytic approach. *J. Appl. Prob.* **47**, pp. 947–966.

[6] Bensoussan A. and Lions, J.L.: *Applications of Variational Inequalities in Stochastic Control*, North-Holland, Amsterdam, 1982.

[7] Hernández-Lerma, O., Lasserre, J.B.: *Discrete-time Markov control processes: Basic optimality criteria.* Springer-Verlag, New York, 1996.

[8] Hernández-Lerma, O., Lasserre, J.B.: *Further topics on discrete-time Markov control processes.* Springer-Verlag, New York, 1999.

[9] Ilhuicatzi-Roldán, R., Cruz-Suárez, H., Chávez-Rodríguez, S.: (2017). Markov decision processes with time-varying discount factors and random horizon. *Kybernetica* **53**, pp. 82–98.

[10] Hinderer, K., Rieder, U., Stieglitz, M.: *Dynamic optimization.* Springer-Verlag, New York, 2010.

[11] Horiguchi, M.: (2001). Stopped Markov decision processes with a stopping time constraint. *Math. Meth. Oper. Res.* **53**, pp. 279–295.

[12] Horiguchi, M.: (2001). Stopped Markov decision processes with multiple constraints. *Math. Meth. Oper. Res.* **54**, pp. 455–469.

[13] Jasso-Fuentes, H., López-Barrientos, J.D.: (2015) On the use of stochastic differential games against nature to ergodic control problems with unknown parameters. *Internat. J. Control* **88**, pp. 897-909.

[14] Jasso-Fuentes, H., Menaldi, J.L., Prieto-Rumeau, T.: (2020) Discrete-time hybrid control in Borel spaces. *Appl. Math. Optim.* **81**, pp. 409–441.

[15] Jasso-Fuentes, H., Menaldi, J.L., Prieto-Rumeau, T., Robin, M.: (2018) Discrete-time hybrid control in Borel spaces: average cost optimality criterion. *J. Math. Anal. Appl.* **462**, pp. 1695-1713.

[16] Jasso-Fuentes, H., Yin, G.: *Advanced criteria for controlled Markov-modulated diffusions in an infinite horizon: overtaking, bias, and Blackwell optimality.* Science Press, Beijing China, 2013.

[17] Kallenberg, O: *Foundations of Modern Probability* Springer-Verlag, New York, 2002.

[18] Menaldi J.L.; Blankenship, G.L.: (1984) Optimal stochastic scheduling of power generation systems with scheduling delays and large cost differentials, *SIAM J.Control Optim.*, **22**, pp. 121–132.

[19] Meyn, S., Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability.* Cambridge University Press.

[20] Minjárez-Sosa, A.: (2015) Markov control models with unknown random state–action-dependent discount factors *TOP* **23**, pp. 743–772.

[21] Morimoto, H.: *Stochastic control and mathematical modeling.* Encyclopedia of Mathematics and its Applications, Cambridge University Press, New York , 2010.

[22] Puterman, M.L.: *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons Inc, New York, 1994.

[23] Rieder, U.: (1975) On stopped decision processes with discrete time parameter. *Stoch. Process Appl.* **3**, pp. 365–383.

[24] Ross, S.M.: *Introduction to stochastic dynamic programming.* Academic Press, New York, 1983.

[25] Wei, Q; Guo, X.P.: (2011) Markov decision processes with state-dependent discount factors and unbounded rewards/costs *Oper. Res. Lett.* **39**, pp. 369–374.