International Journal of Pure and Applied Mathematics
Volume 119 No. 2 2018, 343-356
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: 10.12732/ijpam.v119i2.7

MARKOV DECISION PROCESSES ON BOREL SPACES WITH RANDOM HORIZON AND NONZERO TERMINAL COST

Ilhuicatzi-Roldán Rocio[1] §, Flores-Hernández Rosa M.[2]
[1]Facultad de Ciencias Básicas, Ingeniería y Tecnología
Universidad Autónoma de Tlaxcala
Av. Ángel Solana s/n, Apizaco, 90300, MÉXICO
[2]Facultad de Ciencias Básicas, Ingeniería y Tecnología
Universidad Autónoma de Tlaxcala
Calz. Apizaquito s/n, Apizaco, 90300, MÉXICO

Abstract: In this paper, an optimal control problem with the expected total cost as performance criterion is considered. In this case, a random horizon and a nonzero terminal cost in the performance criterion is supposed. The terminal cost depends on the state of the system at the random occurrence time of termination of the process. Then, under the assumption that the random horizon is independent of the stochastic control process and its probability distribution has a finite support, the dynamic programming equation is obtained, which solves the problem proposed. Also, two examples are included, one is a linear quadratic control problem and other is an inventory control problem, both with a random horizon and a nonzero terminal cost.

AMS Subject Classification: 93E20, 90C40, 90C39
Key Words: Markov decision process, total cost, random horizon, nonzero terminal cost

1. Introduction

This paper is related with the theory of Markov Decision Processes (see [2], [4], [6]), which considers stochastic control problems in discrete time. In each

Received: June 21, 2017
Revised: March 21, 2018
Published: July 4, 2018

§Correspondence author

instant of time it is allowed to perform a control of the process, generating a sequence of actions, also called controls or decisions, which is known as a control policy. To assess the quality of each policy, it has a performance criterion (or objective function).

A classic performance criterion is the expected total cost with a finite horizon and a terminal cost which is depending on the final state of the process. However, in the literature there exist other variants on the horizon, such as an infinite horizon (see [2], [4]) or a random horizon (see [3], [5], [6]). For the problems with a random horizon is considered that the horizon can be dependent on the state, it is dependent on the state and the action, or it has an arbitrary probability distribution independent of the stochastic control process (see [6], p.127). Note that, when a random horizon is considered is possible incur in a terminal cost in the just randomized moment when the control process is finished.

This paper considers the optimal control problem studied in [3], in which a random horizon with an arbitrary probability distribution, independent of the control process, is assumed. The problem under the independence assumption is interesting since can be found on real applications, in this form, external factors which can finish the process are considered. In [3], firstly, through dynamic programming approach, the problem with a finite support for the random horizon is solved. This result is fundamental for the analysis in the case of random horizon with infinite support.

The contribution of this paper is to study the problem with a random horizon adding a nonzero terminal cost in the model, in the case when the probability distribution of the random horizon has a finite support. Then, the equation of dynamic programming is obtained. It can be seen that both, the random horizon and the nonzero terminal cost allow to model real situations. The random horizon, in the context of this article, permit model unexpected events such as the bankruptcy in economic models, the discontinuance of an article in an inventory system, a failure of a system in engineering, the extinction of some natural resource in biology, etc. On the other hand, the nonzero terminal cost can be justified in a real way, for example, in an inventory control problem would be a cost for the leftover inventory.

This article includes two examples, both with a random horizon and a nonzero terminal cost. The first is a linear quadratic control problem, which is a problem with Borel spaces and unbounded cost per stage function; the solution is obtained in analytic form. The second example is an inventory control problem with discrete spaces, here the solution is implemented in a computational program and numerical results are presented.

The work is organized as follows: in the second section, the basic theory of the Markov decision or control processes is presented. Later, in the third section, the control problem with a random horizon and nonzero terminal cost is detailed. Then, in the fourth section, the corresponding control problem is analyzed and solved via dynamic programming. Finally, in the fifth section, the theory developed is exemplified with two problems.

2. Markov Decision Processes

Let $(X, A, \{A(x) : x \in X\}, Q, c)$ be a stationary Markov decision model at discrete time (see [2], [4] and [6]), which consists of the state space X, the action, control or decision set A, a family $\{A(x) : x \in X\}$ of nonempty measurable subsets $A(x)$ of A, whose elements are the feasible actions when the system is in the state $x \in X$; the set $\mathbb{K} := \{(x, a) : x \in X,\ a \in A(x)\}$ of the feasible state-action pairs is assumed to be a measurable subset of $X \times A$. The following component is the transition law Q, which is a stochastic kernel on X given $\mathbb{K}$. Finally, $c : \mathbb{K} \to \mathbb{R}$ is a measurable function called the cost per stage function.

In this document, X and A are assumed to be Borel spaces (i.e. Borel subsets of a separable complete metric space) with Borel $\sigma-$ algebras $\mathcal{B}(X)$ and $\mathcal{B}(A)$, respectively.

A Markov Decision Process (MDP) evolves as follows: at the initial decision epoch, the system occupies the state $x_0 = x \in X$ and a decision maker (or controller) chooses an action $a_0 = a \in A(x)$. Then, a cost $c(x_0, a_0)$ is incurred and the system jumps to a state x_1, according to the transition law $Q(\cdot \mid x, a)$, and the process is repeated. Thus, for each $n \geq 1$ an admissible history $h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n)$, with $(x_k, a_k) \in \mathbb{K}$, for $k = 0, 1, ..., n-1$ and $x_n \in X$, of an MDP up to the n-th transition is obtained. Let $\mathbb{H}_n, n = 0, 1, ...,$ denotes the set of all admissible histories of the system up to the n-th transition. Then, a *control policy* $\pi = \{\pi_n\}$ is a sequence of stochastic kernels π_n on A given $\mathbb{H}_n$, satisfying the constraint $\pi_n(A(x_n) \,|h_n) = 1$, for each $h_n \in \mathbb{H}_n$ and $n = 0, 1....$ The collection of all policies is denoted by Π.

$\mathbb{F}$ denotes the set of measurable functions $f : X \to A$ such that $f(x) \in A(x)$, for all $x \in X$. A deterministic Markov policy is a sequence $\pi = \{f_t\}$ such that $f_t \in \mathbb{F}$, for $t = 0, 1, \ldots$. A deterministic Markov policy $\pi = \{f_t\}$ is said to be stationary if f_t is independent of t, i.e., $f_t = f \in \mathbb{F}$, for all $t = 0, 1, \ldots$; in this case π is denoted by f and $\mathbb{F}$ is refered as the *set of stationary policies*.

In many cases, the evolution of a MDP is specified by a discrete time or

difference equation of the form

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t = 0, 1, \ldots$$

with x_0 given, where $\{\xi_t\}$ is a sequence of independent and identically distributed random variables with values in a Borel space S and a common distribution μ, independent of the initial state x_0. In this case, the transition law Q is given by

$$Q(B|x,a) = \int_S I_B\big(F(x,a,s)\big)\mu(ds),$$

$B \in \mathcal{B}(X)$ and $(x,a) \in \mathbb{K}$, where $I_B(\cdot)$ denotes the indicator function of the set B.

Let $(\Omega, \mathcal{F})$ be the measurable space consisting of the canonical sample space $\Omega = \mathbb{H}_\infty := (X \times A)^\infty$ and let $\mathcal{F}$ be the corresponding product σ-algebra. The elements of Ω are sequences of the form $\omega = (x_0, a_0, x_1, a_1, \ldots)$ with $x_t \in X$ and $a_t \in A$, for all $t = 0, 1, \ldots$. The projections x_t and a_t from Ω to the sets X and A are called state and action variables, respectively.

Let $\pi = \{\pi_t\}$ be an arbitrary policy and let μ be an arbitrary probability measure on X called the initial distribution. Then, by the theorem of C. Ionescu-Tulcea (see [4]), there is an unique probability measure P_μ^π on $(\Omega, \mathcal{F})$ which is supported on $\mathbb{H}_\infty$, i.e., $P_\mu^\pi(\mathbb{H}_\infty) = 1$. The stochastic process $(\Omega, \mathcal{F}, P_\mu^\pi, \{x_t\})$ is called a *Markov decision process* or a *Markov control process, at discrete time.*

The expectation operator with respect to P_μ^π is denoted by E_μ^π. If μ is concentrated at the initial state $x \in X$, then P_μ^π and E_μ^π are written as P_x^π and E_x^π, respectively.

3. Statement of the Problem

Let $(\Omega', \mathcal{F}', P)$ be a probability space and let $(X, A, \{A(x) : x \in X\}, Q, c)$ be a stationary Markov decision model with a discrete time planning horizon τ, where τ is considered as a random variable on $(\Omega', \mathcal{F}')$ with the probability distribution $\rho_t := P(\tau = t)$, $t = 0, 1, \ldots, T$, for T a positive integer. For $\pi \in \Pi$ and $x \in X$, define the performance criterion as

$$j^\tau(\pi, x) := E\left[\sum_{t=0}^{\tau} c(x_t, a_t) + c(x_{\tau+1})\right], \tag{1}$$

where E denotes the expected value with respect to the joint distribution of the process $\{(x_t, a_t)\}$ and τ, and $c(x_{\tau+1})$ is a nonzero terminal cost, which is a given measurable function on X. In this document, the performance criterion (1) will be called *expected total cost with a random horizon and a nonzero terminal cost*. Then, the optimal value function is defined as

$$J^{\tau}(x) := \inf_{\pi \in \Pi} j^{\tau}(\pi, x),$$

$x \in X$. The optimal decision problem with a random horizon is to find a policy $\pi^* \in \Pi$ such that $j^{\tau}(\pi^*, x) = J^{\tau}(x)$, $x \in X$, in this case, π^* is said to be *optimal*.

Assumption 1. For each $x \in X$ and $\pi \in \Pi$ the induced process $\{(x_t, a_t) : t = 0, 1, 2, \ldots\}$ is independent of τ.

Observe that under Assumption 1, $\pi \in \Pi$, $x \in X$, and $P_k = \sum_{n=k}^{T} \rho_n = P(\tau \geq k)$, $k = 0, 1, \ldots, T$,

$$\begin{aligned}
& E\left[\sum_{t=0}^{\tau} c(x_t, a_t) + c(x_{\tau+1})\right] \\
= \; & E\left[E\left[\sum_{t=0}^{\tau} c(x_t, a_t) + c(x_{\tau+1}) \mid \tau\right]\right] \\
= \; & \sum_{n=0}^{T} E_x^{\pi}\left[\sum_{t=0}^{n} c(x_t, a_t) + c(x_{n+1})\right] \rho_n \\
= \; & E_x^{\pi}\left[\sum_{n=0}^{T} \rho_n \sum_{t=0}^{n} c(x_t, a_t)\right] + E_x^{\pi}\left[\sum_{n=0}^{T} \rho_n c(x_{n+1})\right] \\
= \; & E_x^{\pi}\left[\sum_{t=0}^{T} c(x_t, a_t) \sum_{n=t}^{T} \rho_n\right] + E_x^{\pi}\left[\sum_{n=0}^{T} \rho_n c(x_{n+1})\right] \\
= \; & E_x^{\pi}\left[\sum_{t=0}^{T} P_t \; c(x_t, a_t)\right] + E_x^{\pi}\left[\sum_{n=0}^{T} \rho_n c(x_{n+1})\right] \\
= \; & E_x^{\pi}\left[\sum_{t=0}^{T} P_t \; c(x_t, a_t)\right] + E_x^{\pi}\left[\sum_{n=1}^{T+1} \rho_{n-1} c(x_n)\right] \\
= \; & E_x^{\pi}\left[c(x_0, a_0) + \sum_{t=1}^{T}\left[P_t \; c(x_t, a_t) + \rho_{t-1} c(x_t)\right] + \rho_T c(x_{T+1})\right],
\end{aligned}$$

given that $P_0 = \sum_{n=0}^{T} \rho_n = P(\tau \geq 0) = 1$. Thus, the optimal decision problem with a random horizon τ and a nonzero terminal cost $c(x_{\tau+1})$ is equivalent to

the optimal decision problem with a planning horizon $T+1$, a nonhomogeneous cost and a nonzero terminal cost $\rho_T c(x_{T+1})$.

Remark 2. It is not difficult to observe that if the terminal cost $c(x_{\tau+1}) = 0$, then $E\big[\sum_{t=0}^{\tau} c(x_t, a_t)\big] = E_x^{\pi}\big[\sum_{t=0}^{T} P_t\ c(x_t, a_t)\big]$ (see Remark 3.1 in [3]).

4. Characterization of the Optimal Solution using Dynamic Programming Approach

Let X be a metric space and $v : X \to \mathbb{R} \cup \{+\infty\}$ a function with $v(x) < \infty$ for at least one point $x \in X$. v is said lower semicontinuous (l.s.c.) at x if

$$\liminf_{n\to\infty} v(x_n) \geq v(x)$$

for any sequence $\{x_n\}$ in X that converges to x. The function v is called l.s.c. on X if it is l.s.c. at every $x \in X$.

Consider the following assumption on the elements of Markov decision model:

Assumption 3. (a) The one-stage cost $c(\cdot, \cdot)$ and the terminal cost $c(\cdot)$ are l.s.c. and nonnegative. Also, $c(\cdot, \cdot)$ is inf-compact on $\mathbb{K}$, that is, the set $\{a \in A(x) : c(x, a) \leq \lambda\}$ is a compact set for every $x \in X$ and $\lambda \in \mathbb{R}$.

(b) Q is either strongly continuous or weakly continuous. The transition law Q is weakly continuous if the function

$$v'(x, a) := \int_X v(y)Q(dy|x, a)$$

is continuous and bounded on $\mathbb{K}$ for every continuous bounded function v on X. The transition law Q is strongly continuous if the function $v'(x, a)$ is continuous and bounded on $\mathbb{K}$ for every measurable bounded function v on X.

Theorem 4. *Let $J_0, J_1, ..., J_{T+1}$ be the functions on X defined by*

$$J_{T+1}(x) := \rho_T c(x),$$

for $t = T, T-1, ..., 1$,

$$J_t(x) := \min_{a\in A(x)} \left[P_t c(x, a) + \rho_{t-1} c(x) + \int_X J_{t+1}(y)Q(dy \mid x, a)\right], \qquad (2)$$

and

$$J_0(x) := \min_{a \in A(x)} \left[c(x,a) + \int_X J_1(y) Q(dy \mid x, a) \right], \quad x \in X. \tag{3}$$

Under Assumption 3 these functions are measurable and for each $t = 0, 1, ..., T$, *there exist* $f_t \in \mathbb{F}$ *such that for all* $x \in X$, *(2) attains the minimum in* $f_t(x) \in A(x)$, $t = 1, 2, ..., T$, *and (3) attains the minimum in* $f_0(x) \in A(x)$. *This implies that for* $x \in X$

$$J_t(x) = P_t c(x, f_t(x)) + \rho_{t-1} c(x) + \int_X J_{t+1}(y) Q(dy \mid x, f_t(x)),$$

$t = 1, 2, ..., T$, *and*

$$J_0(x) = c(x, f_0(x)) + \int_X J_1(y) Q(dy \mid x, f_0(x)).$$

Then, the deterministic Markov policy $\pi^* = \{f_0, ..., f_T\}$ *is optimal and the optimal value function is given by* $J^\tau(x) = \jmath^\tau(\pi^*, x) = J_0(x)$, $x \in X$.

The proof of previous theorem is similar to the proof of Theorem 3.2.1 in [4].

Remark 5. There exist other conditions for which Theorem 4 is valid, for example see [4], pp. 27-28, or adequate versions of the conditions given in [6], p. 154.

5. Examples with a Random Horizon and a Nonzero Terminal Cost

5.1. Linear Quadratic Control Problem

The linear quadratic control model with a random horizon and a nonzero terminal cost is a modified version of the linear quadratic model presented in [4], p. 34, and is defined as follow: Let $X = A = A(x) = \mathbb{R}$. The cost per stage function is given by

$$c(x, a) = x^2 + a^2,$$

$(x, a) \in \mathbb{K}$. The transition law is induced by the following difference equation:

$$x_{t+1} = x_t + a_t + \xi_t,$$

$t = 0, 1, \ldots, \tau$, with x_0 known. In this case, $\{\xi_t\}$ is a sequence of independent and identically distributed random variables taking values in $S = \mathbb{R}$ with a continuous bounded density function Δ, such that $E[\xi_0] = 0$ and $E[\xi_0^2] = \sigma^2 < +\infty$,

where ξ_0 is a generic element of the sequence $\{\xi_t\}$. Moreover, for the random horizon τ it is consider that $\rho_t := P(\tau = t)$, $t = 0, 1, \ldots, T$ with $T < \infty$, $P_t := P(\tau \geq t)$ and a terminal cost $c(x) = x^2$.

Lemma 6. *The model of Example 5.1 satisfies Assumption 3 and has the optimal solution given by* $\pi^* = \{f_t(x)\}$, *with*

$$\begin{aligned} f_0(x) &= \frac{-C_1 x}{1 + C_1}, \\ f_t(x) &= \frac{-C_{t+1} x}{P_t + C_{t+1}}, \quad t = 1, 2, \ldots, T, \end{aligned}$$

and

$$J^{\tau}(x) = C_0 x^2 + D_0,$$

where the constants C_k, $k = 0, 1, \ldots, T+1$ *and* D_k, $k = 0, 1, \ldots, T$ *are obtained through the following recurrence relations:*

$$\begin{aligned} C_{T+1} &= \rho_T, \\ C_t &= \frac{P_t P_{t-1} + C_{t+1}(P_t + P_{t-1})}{P_t + C_{t+1}}, \quad t = T, T-1, \ldots, 1, \\ C_0 &= \frac{1 + 2C_1}{1 + C_1}, \end{aligned}$$

and

$$\begin{aligned} D_T &= C_{T+1}\sigma^2, \\ D_t &= C_{t+1}\sigma^2 + D_{t+1}, \quad t = T-1, T-2, \ldots, 0. \end{aligned}$$

Proof. First, it will verify that Example 5.1 satisfies Assumption 3. For it, observe the following: the cost per stage and the terminal cost are nonnegative and continuous functions. Next, let $A_\lambda(x) := \{a \in A(x) : c(x, a) \leq \lambda\}$, $\lambda \in \mathbb{R}$. Then,

$$A_\lambda(x) = \begin{cases} \emptyset & \text{if } \lambda < x^2 \\ \{0\} & \text{if } \lambda = x^2 \\ \left[-\sqrt{\lambda - x^2}, \sqrt{\lambda - x^2}\right] & \text{if } \lambda > x^2. \end{cases}$$

Since $A_\lambda(x)$ is compact for each $x \in \mathbb{R}$, then c is an inf-compact function on $\mathbb{K}$. Now, it is verified that the transition law is strongly continuous. For it, let $v : X \to \mathbb{R}$ a measurable bounded function and observe that

$$v'(x, a) = \int_X v(y) Q(dy \mid x, a) = \int_{-\infty}^{\infty} v(x + a + s)\Delta(s) ds.$$

Making the change of variable $u = x + a + s$, it is obtained that

$$v'(x,a) = \int_{-\infty}^{\infty} v(u)\Delta(u - x - a)du.$$

Let $\{(x_k, a_k)\}$ a sequence such that $\lim_{k\to\infty}(x_k, a_k) = (x', a')$. Then, by Dominated Convergence Theorem and the continuity of Δ,

$$\lim_{k\to\infty} v'(x_k, a_k) = v'(x', a'),$$

hence v' is continuous and bounded on $\mathbb{K}$. Thus, the transition law for the linear quadratic Model is strongly continuous. In this form, the Assumption 3 is verified.

Second, by Theorem 4, the dynamic programming equation for the model of Example 5.1 is as follow:

$$\begin{aligned} J_{T+1}(x) &= \rho_T x^2, \\ J_t(x) &= \min_{a\in\mathbb{R}} \left[P_t(x^2 + a^2) + \rho_{t-1}x^2 + E[J_{t+1}(x + a + \xi)]\right], \\ &\quad t = T, T-1, ..., 1, \\ J_0(x) &= \min_{a\in\mathbb{R}} \left[x^2 + a^2 + E[J_1(x + a + \xi)]\right], \end{aligned}$$

$x \in X$.
More specifically, for $t = T + 1$, $J_{T+1}(x) = C_{T+1}x^2$, with $C_{T+1} = \rho_T$.
For $t = T$,

$$J_T(x) = \min_{a\in\mathbb{R}} \left[P_T(x^2 + a^2) + \rho_{T-1}x^2 + E[C_{T+1}(x + a + \xi)^2]\right]$$

$$= \min_{a\in\mathbb{R}} \left[P_T(x^2 + a^2) + \rho_{T-1}x^2 + C_{T+1}(x^2 + a^2 + \sigma^2 + 2xa)\right]$$

$$= \min_{a\in\mathbb{R}} \left[(P_T + C_{T+1})a^2 + 2C_{T+1}xa + (P_{T-1} + C_{T+1})x^2 + C_{T+1}\sigma^2\right], \tag{4}$$

and the minimization process yields the selector:

$$f_T(x) = \frac{-C_{T+1}x}{P_T + C_{T+1}}.$$

Replacing a by f_T in the last expression into brackets in (4), is obtained that

$$J_T(x) = \frac{-C_{T+1}^2 x^2}{P_T + C_{T+1}} + (P_{T-1} + C_{T+1})x^2 + C_{T+1}\sigma^2$$

$$\begin{aligned} &= \frac{P_T P_{T-1} + C_{T+1}(P_T + P_{T-1})}{P_T + C_{T+1}} x^2 + C_{T+1}\sigma^2 \\ &= C_T x^2 + D_T, \end{aligned}$$

where $C_T = \frac{P_T P_{T-1} + C_{T+1}(P_T + P_{T-1})}{P_T + C_{T+1}}$ and $D_T = C_{T+1}\sigma^2$.
In a similar form, for $t = T - 1$,

$$J_{T-1}(x) = \min_{a \in \mathbb{R}} \left[P_{T-1}(x^2 + a^2) + \rho_{T-2} x^2 + E[C_T(x + a + \xi)^2 + D_T] \right]$$

$$\begin{aligned} &= \min_{a \in \mathbb{R}} \left[P_{T-1}(x^2 + a^2) + \rho_{T-2} x^2 + C_T(x^2 + a^2 + \sigma^2 + 2xa) + D_T \right] \\ &= \min_{a \in \mathbb{R}} \left[(P_{T-1} + C_T)a^2 + 2C_T xa + (P_{T-2} + C_T)x^2 + C_T\sigma^2 + D_T \right], \end{aligned}$$

obtaining the selector

$$f_{T-1}(x) = \frac{-C_T x}{P_{T-1} + C_T}.$$

Thus,

$$\begin{aligned} J_{T-1}(x) &= \frac{P_{T-1}P_{T-2} + C_T(P_{T-1} + P_{T-2})}{P_{T-1} + C_T} x^2 + C_T\sigma^2 + D_T \\ &= C_{T-1}x^2 + D_{T-1}, \end{aligned}$$

where $C_{T-1} = \frac{P_{T-1}P_{T-2} + C_T(P_{T-1} + P_{T-2})}{P_{T-1} + C_T}$ and $D_{T-1} = C_T\sigma^2 + D_T$.
Continuing with this process, finally for $t = 0$,

$$\begin{aligned} J_0(x) &= \min_{a \in \mathbb{R}} \left[x^2 + a^2 + E[C_1(x + a + \xi)^2 + D_1] \right] \\ &= \min_{a \in \mathbb{R}} \left[x^2 + a^2 + C_1(x^2 + a^2 + \sigma^2 + 2xa) + D_1 \right] \\ &= \min_{a \in \mathbb{R}} \left[(1 + C_1)a^2 + 2C_1 xa + (1 + C_1)x^2 + C_1\sigma^2 + D_1 \right], \end{aligned}$$

and the selector obtained is

$$f_0(x) = \frac{-C_1 x}{1 + C_1}.$$

Therefore,

$$\begin{aligned} J_0(x) &= \frac{1 + 2C_1}{1 + C_1} x^2 + C_1\sigma^2 + D_1 \\ &= C_0 x^2 + D_0 \\ &= J^\tau(x). \end{aligned}$$

In this form, the proof is concluded. □

5.2. Inventory Control Problem

Consider a problem of ordering a quantity of a certain item at the beginning of each of τ periods, where τ is a random variable such that $\rho_t := P(\tau = t)$, $t = 0, 1, \ldots, T$, with $T < \infty$, for meeting a stochastic demand. Let us denote by

x_t : stock available at the beginning of the tth period,

a_t : stock ordered (and immediately delivered) at the beginning of the tth period,

ξ_t : demand during the tth period with probability distribution known,

where x_t, a_t and ξ_t are supposed nonnegative integer variables. We assume that $\xi_0, \xi_1, \ldots, \xi_T$ are independent and identically distributed random variables such that the probability distribution is given by $\varrho_k := P(\xi_t = k)$, $k = 0, 1, \ldots, C$, where $C < \infty$ is considered as an upper bound on the stock $x_t + a_t$ that can be stored. Thus, $X = A = \{0, 1, \ldots, C\}$ and $A(x) = \{0, 1, \ldots, C - x\}$.
Also, if it is assumed that the excess of demand $\xi_t - x_t - a_t$ is lost, then the stock evolves according to the discrete time equation

$$x_{t+1} = \max\{0, x_t + a_t - \xi_t\}.$$

Finally, the cost per stage is given by

$$c(x_t, a_t, \xi_t) = da_t + h \max\{0, x_t + a_t - \xi_t\} + p \max\{0, \xi_t - x_t - a_t\},$$

where

d is the ordering cost per unit,

h is the holding cost in inventory per unit,

p is the shortage cost for each unfilled demand,

and it is considered a nonzero terminal cost $c(x)$, $x = 0, 1, 2, \ldots, C$. A simpler version of this inventory control problem is presented in [1], p. 18. In this case, the existence of an optimal policy in Theorem 4 is assured by the assumptions in [6], p. 154.

The dynamic programming equation for the inventory control problem, presented above, is given as follow (see Theorem 4 and [4], p. 32):

$$J_{T+1}(x) \quad = \quad \rho_T c(x),$$

$$J_t(x) = \min_{a \in A(x)} E_\xi \left[P_t c(x,a) + \rho_{t-1} c(x) + J_{t+1}(\max\{0, x + a - \xi\}) \right],$$

$$t = T, T-1, T-2, \ldots, 1,$$

$$J_0(x) = \min_{a \in A(x)} E_\xi \left[c(x,a) + J_1(\max\{0, x + a - \xi\}) \right],$$

$x \in X$.

This equation has been programmed in Maple to solve numerical problems as the following:

Consider an inventory system with a finite capacity $C = 8$ and costs per unit by ordering, holding in inventory and unfilled demand given by $d = 2$, $h = 1$ and $p = 3$, respectively. The probability distribution of the demand is given by $\varrho_k = P(\xi = k)$, $k = 0, 1, \ldots, C$, where

$$\varrho = [0.03, 0.04, 0.18, 0.19, 0.2, 0.09, 0.07, 0.12, 0.08].$$

The probability distribution of the horizon is

$$\rho = [0, 0.03, 0.03, 0.05, 0.07, 0.09, 0.13, 0.13, 0.13, 0.09, 0.09, 0.08, 0.08],$$

where $\rho_t = P(\tau = t)$, $t = 0, 1, \ldots, T$, with $T = 12$. The terminal cost in this case is proposed as follow:

$$c(x) = x, \quad x = 0, 1, \ldots, C;$$

c is an increasing function because it is associated with the number of items remained in the inventory when the process is finished.

In Table 1, the optimal policy is showed, that is, for each stage t and each level of the inventory x is indicated the number of items to be ordered.

In Table 2, the optimal value is presented for each initial state, and in Table 3, the optimal policy is showed when a null terminal cost is considered.

Stage State	0	1	2	3	4	5	6	7	8	9	10	11	12
0	3	4	4	4	3	3	3	3	3	3	3	2	0
1	2	3	3	3	2	2	2	2	2	2	2	1	0
2	1	2	3	2	1	1	1	1	1	1	1	0	0
3	0	1	1	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1: Optimal policy

Initial state x	0	1	2	3	4
Optimal value	74.1327	72.1327	70.1327	68.1327	66.4791
Initial state x	5	6	7	8	
Optimal value	65.4477	64.7114	64.2653	64.3073	

Table 2: Optimal value

6. Conclusions

In this paper, the dynamic programming equation for the optimal control problem with a random horizon and nonzero terminal cost has been established. This equation has been applied for solving two problems. In this case, the support of the probability distribution of the random horizon has been assumed finite. This result can be necessary for approaching the problem when a random horizon assumes an infinite support for its probability distribution, considering also a nonzero terminal cost.

Stage / State	0	1	2	3	4	5	6	7	8	9	10	11	12
0	4	4	4	4	4	3	3	3	3	3	3	2	0
1	3	3	3	3	3	2	2	2	2	2	2	1	0
2	2	2	2	2	2	1	1	1	1	1	1	0	0
3	1	1	1	1	1	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3: Optimal policy with a null terminal cost

References

[1] D.P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, USA, 1987.

[2] D.P. Bertsekas, S.E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, Massachusetts, 1978.

[3] H. Cruz-Suárez, R. Ilhuicatzi-Roldán, R. Montes-de-Oca, Markov decision processes on Borel spaces with total cost and random horizon, *Journal of Optimization Theory and Applications*, **162** (2013), 329-346.

[4] O. Hernández-Lerma, J.B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[5] T. Iida, M. Mori, Markov decision processes with random horizon, *Journal of the Operations Research*, **39** (1996), 592-603.

[6] M.L. Puterman, *Markov Decision Process: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, New York, 1994.