Two person zero-sum semi-Markov games with unknown holding times distribution on one side: discounted payoff criterion[*]

J. Adolfo Minjárez-Sosa[†]and Fernando Luque-Vásquez
Departamento de Matemáticas, Universidad de Sonora
Rosales s/n, Centro, 83000 Hermosillo, Sonora, MEXICO

Abstract

This paper deals with two person zero-sum semi-Markov games with possibly unbounded payoff function, under a discounted payoff criterion. Assuming that the distribution of the holding times H is unknown for one of the players, we combine suitable methods of statistical estimation of H with control procedures to construct an asymptotically discount optimal pair of strategies.
AMS 2000 subject classifications: 91A15, 91A25, 90C40.

Key Words: Zero-sum semi Markov games, discounted payoff, asymptotic optimality, Shapley equation.

Abbreviated title: Zero-sum semi-Markov games with unknown holding time distribution.

1 Introduction

This paper concerns two-person zero-sum semi-Markov games (SMGs) in Borel spaces, with possibly unbounded payoff function, under a discounted

[*] Work supported partially by Consejo Nacional de Ciencia y Tecnologia (CONACyT) under Grant 37239E.

[†] Corresponding author (e-mail: aminjare@gauss.mat.uson.mx)

payoff criterion. The game can be formulated as follows: there are two players with opposite objectives. If at the nth decision epoch, the game is in the state $x_n = x$, then the players independently of each other choose actions $a_n = a$ and $b_n = b$, and the following happens: the game remains in the state x during a nonnegative random time δ_{n+1} with distribution H, and a payoff r is generated which represents a reward for player 1 and a cost for player 2; moreover, the game jumps to a new state $x_{n+1} = y$ according to some transition law. Once the transition to the state y occurs, the process is repeated. Payoff accumulates throughout the evolution of the game and, the goal of each player is to optimize the total discounted payoff.

The class of zero-sum SMGs we are interested in is when the distribution H of the holding (or sojourn) times is known by player 1 but unknown by player 2. In addition, as usual, we suppose that the payoff r is the sum of an immediate payoff imposed at the moment when the players choose their decisions, plus a payoff rate imposed until the transition to a new state of the game occurs. In this context, at the time of the nth decision epoch T_n, when the game is in state $x_n = x$, player 1 may choose the action $a_n = a$ in a standard way, whereas player 2, before choosing the action b_n, must implement a statistical estimation method to obtain an estimate H_n of H, and then selects an action $b = b_n(H_n)$.

The actions applied by players at the decision epochs, are selected according to rules known as strategies. Hence, our main contribution in this paper is the following. Assuming that the game model satisfies sufficient conditions for the existence of the value of the game and for the existence of a solution to the *Shapley equation*, a suitable estimation method of H is used by player 2 to construct a discounted optimal pair of strategies (π^1_*, π^2_*) for players 1 and 2. However, since the discounted payoff criterion depends heavily on the decisions selected at the first stages (precisely when the information about the distribution H is deficient), we cannot ensure, in general, optimality of the pair (π^1_*, π^2_*). Therefore, the optimality will be analyzed in an asymptotic sense motivated by the paper of Schäl [14] (see also [2]) for Markov control processes.

The study of zero-sum stochastic Markov games was started by L. Shapley [15], and several extensions of that work have been proposed. In particular, related papers on semi-Markov games are [6], [7], [9], [11], [12] and [17], in which the distribution H is supposed to be known for both players. To the best our knowledge, there are no works dealing with semi-Markov games in the context of our paper.

The remainder of the paper is organized as follows: in Section 2, we introduce the semi-Markov game model we will be dealing with, and in Section 3 we introduce the performance criterion. The main result is stated in Section 4 and the proof is given in Section 5. Finally, an example of a storage system satisfying all the hypotheses of the paper is described in Section 6.

Notation. Given a Borel space X (that is, a Borel subset of a complete and separable metric space) its Borel sigma-algebra is denoted by $\mathcal{B}(X)$, and "measurable", for either sets or functions, means "Borel measurable". Given a Borel space X, we denote by $\mathbb{P}(X)$ the family of probability measures on X, endowed with the weak topology. Let X and Y be Borel spaces. Then a stochastic kernel $\gamma(dx \mid y)$ on X given Y is a function such that $\gamma(\cdot \mid y)$ is a probability measure on X for each fixed $y \in Y$, and $\gamma(B \mid \cdot)$ is a measurable function on Y for each fixed $B \in \mathcal{B}(X)$. In addition, we denote by $\mathbb{P}(X \mid Y)$ the family of stochastic kernels on X given Y.

2 Semi-Markov game model

We consider a two-person semi-Markov game model of the form

$$\mathcal{GM} := (X, A, B, \mathbb{K}_A, \mathbb{K}_B, Q, H, D, d)\,, \tag{1}$$

where X is the state space, A and B are the action spaces for players 1 and 2, respectively. The sets X, A and B are assumed to be Borel spaces and $\mathbb{K}_A \in \mathcal{B}(X \times A)$ and $\mathbb{K}_B \in \mathcal{B}(X \times B)$ are the constraint sets. For every $x \in X$, we define the sets $A(x) := \{a \in A : (x,a) \in \mathbb{K}_A\}$ and $B(x) := \{b \in B : (x,a) \in \mathbb{K}_B\}$, whose elements are the available actions for player 1 and player 2 in state x, respectively. The set $\mathbb{K} = \{(x,a,b) : x \in X,\ a \in A(x),\ b \in B(x)\}$ of admissible state-actions triplets is assumed to be a Borel subset of the Cartesian product $X \times A \times B$. The transition law $Q(\cdot \mid \cdot)$, is a stochastic kernel on X given $\mathbb{K}$, and $H(\cdot \mid x,a,b)$ is the distribution function of the holding time at state $x \in X$ when the actions $a \in A(x)$ and $b \in B(x)$ are chosen, which is known by player 1 but unknown by player 2. Finally, the payoff functions D and d are possibly unbounded and measurable real-valued functions on $\mathbb{K}$.

The game is played as follows: If at time of the nth decision epoch, the state of the game is $x_n = x$, and the actions chosen by player 1 and 2 are $a_n = a \in A(x)$ and $b = b_n(H_n) \in B(x)$, then the game remains in the state x during a nonnegative random time δ_{n+1} with distribution H, and the following happen: 1) player 1 receives an immediate reward $D(x, a, b)$ while player 2 incurs an immediate cost $D(x, a, b)$; 2) the game jumps to a new state $x_{n+1} = y$ according to the transition law $Q(\cdot \mid x, a, b)$; and 3) a reward rate (cost rate) $d(x, a, b)$ for player 1 (player 2) is imposed until the transition occurs. Once the transition to state y occurs, the process is repeated. Thus, the goal of player 1 is to maximize his/her reward, whereas that of player 2 is to minimize his/her cost.

Observe that the decision epochs are $T_n := T_{n-1} + \delta_n$ for $n \in\mathbb{N}$, and $T_0 = 0$. The random variable $\delta_{n+1} = T_{n+1} - T_n$ is called the sojourn or holding time at state x_n.

Remark 2.1 *a) We shall assume that the payoffs are continuously discounted. That is, for a given discount factor $\alpha > 0$, a payoff R incurred at time t is equivalent to a payoff $R\exp(-\alpha t)$ at time 0. In this sense, the one-stage reward for player 1 and cost for player 2 takes the form:*

$$r(x, a, b) := D(x, a, b) + d(x, a, b) \int_0^\infty \int_0^t \exp(-\alpha s) ds H(dt \mid x, a, b), \quad (x, a, b) \in \mathbb{K}. \tag{2}$$

Hence, the function r is also unknown for player 2 (since r depends on H which is unknown for player 2).

b) In addition, we will suppose that the distribution H is independent of the admissible state-actions triplets $(x, a, b) \in \mathbb{K}$ and it has a density ρ. That is, there exists a distribution function G (unknown) with a density ρ such that

$$H(t \mid x, a, b) = G(t) = \int_0^t \rho(s) ds \quad \forall (x, a, b) \in \mathbb{K}, \ \ t \geq 0.$$

Now, defining

$$\Delta_\alpha := \int_0^\infty \exp(-\alpha s)\rho(s)ds \tag{3}$$

and

$$\tau_\alpha := \frac{1-\Delta_\alpha}{\alpha}, \tag{4}$$

it follows that the payoff function (2) takes the form

$$r(x,a,b) = D(x,a,b) + \tau_\alpha d(x,a,b), \quad (x,a,b) \in \mathbb{K}. \tag{5}$$

Assumption 2.2 *There exist $q \in (1,2)$ and a measurable function $\bar{\rho} : [0,\infty) \to [0,\infty)$ such that $\rho \in L_q([0,\infty))$, $\rho(s) \leq \bar{\rho}(s)$ almost everywhere with respect to the Lebesgue measure, and*

$$\int_0^\infty (\bar{\rho}(s))^{2-q}\, ds < \infty.$$

For example, if $\bar{\rho}(s) := M' \min\{1, 1/s^{1+r}\}$, $s \in [0,\infty)$, for some $r > 0$, then there are plenty of densities that satisfy Assumption 2.2.

We define the spaces of admissible histories of the game up to the nth decision epoch by $\mathbb{H}_0 := X$, and $\mathbb{H}_n := (\mathbb{K} \times \Re_+)^n \times X$ for $n \in \mathbb{N} := \{1,2,...\}$. A typical element of $\mathbb{H}_n$ is written as $h_n = (x_0, a_0, b_0, \delta_1, ..., x_{n-1}, a_{n-1}, b_{n-1}, \delta_n, x_n)$. A *strategy* for player 1 is a sequence $\pi^1 = \{\pi_n^1\}$ of stochastic kernels $\pi_n^1 \in \mathbb{P}(A \,|\mathbb{H}_n)$ such that $\pi_n^1(A(x_n) \mid h_n) = 1$ for all $h_n \in \mathbb{H}_n$ and $n \in \mathbb{N}$. We denote by Π^1 the set of all strategies for player 1. A strategy $\pi^1 = \{\pi_n^1\}$ for player 1 is called *stationary* if there exists $f \in \mathbb{P}(A \mid X)$ such that $f(x) \in \mathbb{P}(A(x))$ and $\pi_n^1 = f$ for all $x \in X$ and $n \in \mathbb{N}$. In this case, we identify π^1 with f, i.e., $\pi^1 = f = \{f, f, ...\}$. We denote by Π_S^1 the set of all stationary strategies for player 1.

The sets Π^2 and Π_S^2 of all strategies and all stationary strategies, respectively, for player 2, are defined similarly.

Let $(\Omega, \mathfrak{A})$ be the canonical measurable space that consist of the sample space $\Omega = (\mathbb{K} \times \Re_+)^\infty$ and its product $\sigma-$algebra $\mathfrak{A}$. Then for each pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ and each initial state $x \in X$, there exist a probability measure $P_x^{\pi^1,\pi^2}$ and a stochastic process $\{(x_n, a_n, b_n, \delta_{n+1})\}$, $n = 0, 1, ...$, where x_n, a_n, b_n represent the state and the actions for player 1 and 2, respectively, at the nth decision epoch, whereas δ_{n+1} represents the time between the nth and $(n+1)$th decision epoch. $E_x^{\pi^1,\pi^2}$ denotes the expectation operator with respect $P_x^{\pi^1,\pi^2}$. We note that by Remark 2.1(b), the distribution of δ_n $(n = 1, 2, ...)$ is independent of the strategies π^1 and π^2 and

$$P_x^{\pi^1,\pi^2}[\delta_n \leq t] =: P[\delta_n \leq t] = \int_0^t \rho(s) ds.$$

Assumption 2.3 *There exist $\varepsilon > 0$ and $\theta > 0$ such that*

$$\int_0^\theta \rho(s) ds \leq 1 - \varepsilon.$$

Assumption 2.3 ensures that in a bounded time interval there are at most a finite number of transitions of the process. On the other hand, following similar ideas as in [16] for semi-Markov control processes, we have that

$$\Delta_\alpha < 1, \tag{6}$$

which in turn yields

$$\tau_\alpha < 1/\alpha. \tag{7}$$

Let γ be a real number such that $\Delta_\alpha \leq \gamma < 1$.

Assumption 2.4 *a) For each $x \in X$ the sets $A(x)$ and $B(x)$ are compact.*
b) For each $(x, a, b) \in \mathbb{K}$, $r(x, \cdot, b)$ is upper semi-continuous (u.s.c.) on $A(x)$, and $r(x, a, \cdot)$ is lower semi-continuous (l.s.c.) on $B(x)$.
c) There exist a measurable function $W_0 : X \to [1, \infty)$ and positive constants $\bar{c}_0$, $p > 1$, $d_0 < \infty$, and, $\beta_0 < 1$ such that

$$\max\{|D(x, a, b)|, |d(x, a, b)|\} \leq \bar{c}_0 W_0(x),$$

and

$$\int_X W_0^p(y)Q(dy \mid x,a,b) \leq \beta_0 W_0^p(x) + d_0, \tag{8}$$

for all $(x,a,b) \in \mathbb{K}$.
d) For each $(x,a,b) \in \mathbb{K}$ *and each bounded measurable function* v *on* X, *the functions*

$$a \longmapsto \int_X v(y)Q(dy \mid x,a,b) \quad \textit{and} \quad b \longmapsto \int_X v(y)Q(dy \mid x,a,b) \tag{9}$$

are continuous on $A(x)$ *and* $B(x)$ *respectively. In addition, (9) holds when* v *is replaced with* W_0.

Remark 2.5 *a) Applying Jensen's inequality to (8) yields*

$$\int_X W_0(y)Q(dy \mid x,a,b) \leq \beta' W_0(x) + d, \ \textit{for all} \ \ (x,a,b) \in \mathbb{K}, \tag{10}$$

where $\beta' = \beta_0^{1/p}$ *and* $d = d_0^{1/p}$. *Moreover, a consequence of both inequalities (8) and (10) is (see [1, 3]):*

$$\sup_{n\geq 0} E_x^{\pi^1,\pi^2}[W_0^p(x_n)] < \infty \quad \textit{and} \quad \sup_{n\geq 0} E_x^{\pi^1,\pi^2}[W_0(x_n)] < \infty,$$

for each pair $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ *and* $x \in X$.
b) Using similar arguments to those used in the proof of Proposition 8.3.4 and Remark 8.3.5(a) in [3] we can prove that Assumption 2.4 implies the existence of a measurable function $W : X \to [1,\infty)$ *and positive constants* k, $\bar{c}_1$ *and* β, *such that* $\beta\gamma < 1$ *and for all* $(x,a,b) \in \mathbb{K}$,

(i) $W(x) \leq kW_0(x)$;

(ii) $\max\{|D(x,a,b)|, |d(x,a,b)|\} \leq \bar{c}_1 W(x)$;

(iii) $\int_X W(y)Q(dy \mid x,a,b) \leq \beta W(x)$.

Thus, by (i) we have

$$\sup_{n\geq 0} E_x^{\pi^1,\pi^2}[W^p(x_n)] < \infty \quad \textit{and} \quad \sup_{n\geq 0} E_x^{\pi^1,\pi^2}[W(x_n)] < \infty, \tag{11}$$

c) From (5), (ii) and (7),

$$|r(x,a,b)| \leq \bar{c}W(x) \quad \textit{for all } (x,a,b) \in \mathbb{K}, \tag{12}$$

where $\bar{c} := \bar{c}_1 \left(1 + \frac{1}{\alpha}\right)$.
d) For any probability measures $\mu \in \mathbb{P}(A(x))$ *and* $\lambda \in \mathbb{P}(B(x))$*, and any function* $u : X \to \Re$ *we write*

$$r(x,\mu,\lambda) := \int\limits_{B(x)} \int\limits_{A(x)} r(x,a,b)\mu(da)\lambda(db)$$

and

$$\int\limits_X u(y)\, Q(dy|x,\mu,\lambda) := \int\limits_{B(x)} \int\limits_{A(x)} \int\limits_X u(y)\, Q(dy|x,a,b)\mu(da)\lambda(db).$$

In particular

$$Q(D|x,\mu,\lambda) := \int\limits_{B(x)} \int\limits_{A(x)} Q(D|x,a,b)\mu(da)\lambda(db).$$

We denote by $\mathbb{B}_W^\infty$ the normed linear space of all measurable functions $u : X \to \Re$ with the finite norm $\|u\|_W$ defined as

$$\|u\|_W := \sup_{x \in X} \frac{|u(x)|}{W(x)}.$$

3 Discounted optimality criterion

For each pair of strategies (π^1, π^2) and initial state $x_0 = x \in X$, we define the total expected $\alpha-$discounted payoff as

$$V(x,\pi^1,\pi^2) := E_x^{\pi^1,\pi^2}\left[\sum_{n=0}^{\infty} \exp(-\alpha T_n) r(x_n,a_n,b_n)\right], \tag{13}$$

We define the lower and the upper value functions as:

$$L(x) := \sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} V(x,\pi^1,\pi^2), \quad x \in X, \tag{14}$$

and

$$U(x) := \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} V(x, \pi^1, \pi^2), \quad x \in X. \tag{15}$$

A pair (π^1_*, π^2_*) is said to be an *optimal pair of strategies* if for all $x \in X$,

$$U(x) := \inf_{\pi^2 \in \Pi^2} V(x, \pi^1_*, \pi^2) \text{ and } L(x) := \sup_{\pi^1 \in \Pi^1} V(x, \pi^1, \pi^2_*). \tag{16}$$

If such an optimal pair exists, then $U(x) = L(x)$ for all $x \in X$, and the common function is called the value of the game and is denoted by $V(x)$. Observe that in this case $V(x) = V(x, \pi^1_*, \pi^2_*)$.

Assumptions 2.3 and 2.4 ensure the existence of a value of the game. More precisely, from [9] we have:

Proposition 3.1 *Suppose that Assumptions 2.3 and 2.4 hold. Then*
a) The game has a value $V \in \mathbb{B}^\infty_W$, that is, $L(x) = U(x) = V(x)$ for all $x \in X$. Moreover, there exists a constant $M < \infty$ such that

$$\|V\|_W \leq M/(1 - \Delta_\alpha). \tag{17}$$

b) The value of the game V satisfies, for all $x \in X$,

$$\begin{aligned} V(x) &= \sup_{\mu \in \mathbb{P}(A(x))} \inf_{\lambda \in \mathbb{P}(B(x))} \left\{ r(x, \mu, \lambda) + \Delta_\alpha \int_X V(y) \, Q(dy|x, \mu, \lambda) \right\} \\ &= \inf_{\lambda \in \mathbb{P}(B(x))} \sup_{\mu \in \mathbb{P}(A(x))} \left\{ r(x, \mu, \lambda) + \Delta_\alpha \int_X V(y) \, Q(dy|x, \mu, \lambda) \right\}. \end{aligned} \tag{18}$$

c) There exist $f^* \in \mathbb{P}(A(x))$ and $g^* \in \mathbb{P}(B(x))$ such that, for all $x \in X$,

$$V(x) = \inf_{\lambda \in \mathbb{P}(B(x))} \left\{ r(x, f^*, \lambda) + \Delta_\alpha \int_X V(y) \, Q(dy|x, f^*, \lambda) \right\} \tag{19}$$

$$= \sup_{\mu \in \mathbb{P}(A(x))} \left\{ r(x, \mu, g^*) + \Delta_\alpha \int_X V(y) \, Q(dy|x, \mu, g^*) \right\}$$

$$= r(x, f^*, g^*) + \Delta_\alpha \int_X V(y) \, Q(dy|x, f^*, g^*). \tag{20}$$

In addition, (f^*, g^*) is an optimal pair of strategies.

Remark 3.2 *Observe that (18) is equivalent to*

$$\sup_{\mu\in I\!\!P(A(x))}\inf_{\lambda\in I\!\!P(B(x))}\Phi(x,\mu,\lambda)=\inf_{\lambda\in I\!\!P(B(x))}\sup_{\mu\in I\!\!P(A(x))}\Phi(x,\mu,\lambda)=0,$$

where

$$\Phi(x,\mu,\lambda)=r(x,\mu,\lambda)+\Delta_\alpha\int\limits_X V\,(y)\,Q(dy|x,\mu,\lambda)-V(x), \qquad (21)$$

for $x\in X$, $\mu\in I\!\!P(A(x))$, $\lambda\in I\!\!P(B(x))$. The optimal pair (f^*,g^*) (see (20)), satisfies $\Phi(x,f^*,g^*)=0$. Furthermore, observe that for all $x\in X$

$$\Phi(x,f^*,\lambda)\geq 0 \qquad \forall\lambda\in I\!\!P(B(x)) \qquad (22)$$

and

$$\Phi(x,\ \mu,g^*)\leq 0 \qquad \forall\mu\in I\!\!P(A(x)). \qquad (23)$$

These facts motivate the following definition.

Definition 3.3 *A pair of strategies $(\pi_*^1,\pi_*^2)\in\Pi^1\times\Pi^2$ is said to be asymptotically discount optimal if, for each $x\in X$,*

$$\lim_{n\to\infty}E_x^{\pi_*^1,\pi^2}\Phi(x_n,a_n,b_n)\geq 0 \qquad \forall\pi^2\in\Pi^2$$

and

$$\lim_{n\to\infty}E_x^{\pi^1,\pi_*^2}\Phi(x_n,a_n,b_n)\leq 0 \qquad \forall\pi^1\in\Pi^1.$$

Observe that if (π_*^1,π_*^2) is an asymptotically discount optimal pair of strategies, then, for each $x\in X$,

$$E_x^{\pi_*^1,\pi_*^2}\Phi(x_n,a_n,b_n)\to 0 \text{ as } n\to\infty.$$

4 Construction of strategies

Since all the components of the game model are known to player 1, he/she may construct his/her strategies in a standard way. In contrast, player 2 must combine suitable statistical density estimation methods of ρ with control procedures in order to construct his/her strategies.

Let $f^* \in \mathbb{P}(A(x))$ be a maximizer satisfying (19). We define the strategy $\pi^1_* \in \Pi^1_S$ for player 1 as $\pi^1_* = \{f^*\}$.

4.1 Construction of strategies for player 2

Density estimation. Let $\delta_1, \delta_2, ..., \delta_n$ be independent realizations (observed by player 2 up to the moment of the nth decision epoch) of r.v.'s with the unknown density ρ, and let $\hat{\rho}_n(s) := \hat{\rho}_n(s; \delta_1, \delta_2, ..., \delta_n),\ s \in \Re_+$, be an estimator of ρ such that

$$E\left\|\rho - \hat{\rho}_n\right\|_q^{qp'/2} \to 0 \ \text{ as } \ n \to \infty, \tag{24}$$

where q and p are as in Assumption 2.2 and Assumption 2.4(c), respectively, and $1/p + 1/p' = 1$. Examples of estimators satisfying (24) are given, for instance, in [4].

To construct strategies for player 2, we estimate ρ by the projection ρ_n of $\hat{\rho}_n$ on the set of densities D in $L_q([0,\infty))$ defined as follows:

$$D := \left\{ \zeta : \zeta \text{ is a density on } L_q([0,\infty)),\ \int_0^\infty \exp(-\alpha s)\zeta(s)ds \le \gamma, \right.$$
$$\left. \int_0^\theta \zeta(s)ds < 1 - \varepsilon,\ \zeta(s) \le \bar{\rho}(s) \text{ a.e.} \right\}. \tag{25}$$

See Assumption 2.3 and 2.4 for the constants θ, ε, and γ, and observe that $\rho \in D$. The existence (and uniqueness) of the estimator ρ_n is guaranteed because the set D is convex and closed in $L_q([0,\infty))$, which can be easily proved following the ideas in [1, 5, 10]. In fact, $\rho_n \in D$ is the "best approximation" of the estimator $\hat{\rho}_n$ on the set D. That is, for each $n \in \mathbb{N}$,

$$\left\|\rho_n - \hat{\rho}_n\right\|_q = \inf_{\zeta \in D} \left\|\zeta - \hat{\rho}_n\right\|_q. \tag{26}$$

In addition, denoting

$$\eta_n := \int\limits_0^\infty |\rho(s) - \rho_n(s)| \, ds, \quad n \in \mathbb{N},$$

and letting p' as in (24), we have

$$E\left[\eta_n^{p'}\right] \to 0 \;\text{ as }\; n \to \infty. \tag{27}$$

Indeed, let $M' < \infty$ such that $\int_0^\infty (\bar{\rho}(s))^{2-q} \, ds \leq M'$ (see Assumption 2.2). Then, for each $n \in\mathbb{N}$, applying Holder's inequality, we have,

$$\begin{aligned}
\eta_n &= \int\limits_0^\infty |\rho(s) - \rho_n(s)|^{\frac{2-q}{2}} \, |\rho(s) - \rho_n(s)|^{\frac{q}{2}} \, ds \\
&\leq \left(\int\limits_0^\infty |\rho(s) - \rho_n(s)|^{2-q} \, ds \right)^{1/2} \left(\int\limits_0^\infty |\rho(s) - \rho_n(s)|^q \, ds \right)^{1/2} \\
&\leq \left(\int\limits_0^\infty (2\bar{\rho}(s))^{2-q} \, ds \right)^{1/2} \left(\int\limits_0^\infty |\rho(s) - \rho_n(s)|^q \, ds \right)^{1/2} \\
&\leq 2^{\frac{2-q}{2}} M' \, \|\rho - \rho_n\|_q^{q/2} \\
&\leq 2^{\frac{2-q}{2}} M' 2^{q/2} \, \|\rho - \hat{\rho}_n\|_q^{q/2},
\end{aligned}$$

where the last inequality follows from (26). Hence, (27) follows from (24).

On the other hand, for $n \in\mathbb{N}$, let (as in (3) and (4))

$$\Delta_n := \int\limits_0^\infty \exp(-\alpha s) \rho_n(s) ds \tag{28}$$

and

$$\tau_n := \frac{1 - \Delta_n}{\alpha}. \tag{29}$$

Observe that $\Delta_n < 1$ which in turn implies that $\tau_n < 1/\alpha$ (see (6) and (7)). Furthermore, for each $n \in \mathbb{N}$,

$$|\Delta_\alpha - \Delta_n| \leq \eta_n \tag{30}$$

and

$$|\tau_\alpha - \tau_n| \leq \frac{\eta_n}{\alpha}. \tag{31}$$

Construction of strategies. We define the sequence $\{L_n\}$ of functions in $\mathbb{B}_W^\infty$ as:

$$L_0(x) = 0;$$

$$L_n(x) = \inf_{\lambda \in \mathbb{P}(B(x))} \sup_{\mu \in \mathbb{P}(A(x))} \left\{ r_n(x,\mu,\lambda) + \Delta_n \int_X L_{n-1}(y)\, Q(dy|x,\mu,\lambda) \right\}, \tag{32}$$

for $n \in \mathbb{N}, \quad x \in X$, where r_n is the approximate payoff function (see (5)):

$$r_n(x,a,b) = D(x,a,b) + \tau_n d(x,a,b), \quad (x,a,b) \in \mathbb{K}. \tag{33}$$

Observe that

$$|r(x,a,b) - r_n(x,a,b)| \leq \frac{\bar{c}\eta_n}{\alpha} W(x), \quad (x,a,b) \in \mathbb{K}, \quad n \in \mathbb{N}, \tag{34}$$

and (see (12))

$$|r_n(x,a,b)| \leq \bar{c}W(x), \quad (x,a,b) \in \mathbb{K}, \quad n \in \mathbb{N}.$$

Thus, a straightforward calculation shows that, for some constant C_2,

$$|L_n(x)| \leq C_2 W(x) \quad \forall n \in \mathbb{N}, \quad x \in X. \tag{35}$$

On the other hand, it is easy to prove that (following similar ideas to prove the interchange of inf and sup in (18)) for $n \in \mathbb{N}, \quad x \in X$,

$$L_n(x) = \sup_{\mu \in \mathbb{P}(A(x))} \inf_{\lambda \in \mathbb{P}(B(x))} \left\{ r_n(x,\mu,\lambda) + \Delta_n \int_X L_{n-1}(y)\, Q(dy|x,\mu,\lambda) \right\}. \tag{36}$$

Now, applying standard arguments on the existence of minimizers (see, e.g., [3, 8, 13]), under Assumptions 2.3 and 2.4, we have that for each $n \in \mathbb{N}$, there exists $g_n = g_n^{\rho_n} \in \mathbb{P}(B(x))$ such that

$$L_n(x) = \sup_{\mu \in \mathbb{P}(A(x))} \left\{ r_n(x, \mu, g_n) + \Delta_n \int_X L_{n-1}(y)\, Q(dy|x, \mu, g_n) \right\}, \quad x \in X, \tag{37}$$

where the minimization is done for every $\omega \in \Omega$.

We define the strategy $\hat{\pi}^2 = \{\hat{\pi}_n^2\}$ for player 2 by $\hat{\pi}_n^2 := g_n$ for all $n \in \mathbb{N}$, and $\hat{\pi}_0^2$ is any fixed action.

We can now state our main result as follows.

Theorem 4.1 *Under Assumptions 2.2-2.4, $(\pi_*^1, \hat{\pi}^2)$ is an asymptotically discount optimal pair of strategies.*

5 Proof of Theorem 4.1

Throughout the proof, we will repeatedly use the following inequalities. For any $u \in \mathbb{B}_W(X)$,

$$|u(x)| \leq \|u\|_W\, W(x) \tag{38}$$

and

$$\int_X u(y) Q(dy \mid x, a, b) \leq \beta \|u\|_W\, W(x), \tag{39}$$

for all $(x, a, b) \in \mathbb{K}$. The inequality (38) is a consequence of the definition of $\|\cdot\|_W$, whereas (39) follows from (38) and (iii) in Remark 2.5(b).

Lemma 5.1 *Suppose that Assumptions 2.2-2.4 hold. Then*

$$\lim_{n\to\infty} E_x^{\pi^1,\pi^2} \|V - L_n\|_W^{p'} = 0,$$

for every $x \in X$ and $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$.

Proof. Let us define the operators

$$Tu(x) := \inf_{\lambda\in\mathbb{P}(B(x))} \sup_{\mu\in\mathbb{P}(A(x))} \left\{ r(x,\mu,\lambda) + \Delta_\alpha \int_X u(y)Q(dy|x,\mu,\lambda) \right\},$$

$$T_m u(x) := \inf_{\lambda\in\mathbb{P}(B(x))} \sup_{\mu\in\mathbb{P}(A(x))} \left\{ r_m(x,\mu,\lambda) + \Delta_m \int_X u(y)Q(dy|x,\mu,\lambda) \right\},$$

for, $m \in\mathbb{N}$, $x \in X$, $u \in\mathbb{B}_W(X)$. By Assumption 2.4(c), T and T_m map $\mathbb{B}_W(X)$ into itself. In [9] has been proved that T is a contraction operator with modulus $\beta\Delta_\alpha$. It can also be proved that for each $m \in\mathbb{N}$, T_m is a contraction operator with modulus $\beta\Delta_m$. Thus

$$\|Tu - Tv\|_W \leq \beta\Delta_\alpha \|u - v\|_W \tag{40}$$

and

$$\|T_m u - T_m v\|_W \leq \beta\Delta_m \|u - v\|_W ,$$

for all $u, v \in\mathbb{B}_W(X)$, $m \in\mathbb{N}$. Now (see (28)) since $\Delta_m \leq \gamma < 1$, we have for all $u, v \in\mathbb{B}_W(X)$, $m \in\mathbb{N}$,

$$\|T_m u - T_m v\|_W \leq \beta\gamma \|u - v\|_W . \tag{41}$$

Note that from Assumption 2.4 (see Remark 2.5(a)), $\beta\gamma < 1$.

From (18) and (32),

$$TV = V \quad \text{and} \quad T_n L_{n-1} = L_n, \quad n \in \mathbb{N}.$$

Therefore, from (41), for each $n \in\mathbb{N}$,

$$\|V - L_n\|_W \leq \|TV - T_n V\|_W + \beta\gamma \|V - L_{n-1}\|_W . \tag{42}$$

On the other hand, from (17) and (34)

$$|TV(x) - T_nV(x)| \leq \sup_{\lambda\in\mathbb{P}(B(x))} \sup_{\mu\in\mathbb{P}(A(x))} \left\{ |r(x,\mu,\lambda) - r_n(x,\mu,\lambda)| \right.$$

$$\left. + |\Delta_\alpha - \Delta_n| \int_X V(y)Q(dy|x,\mu,\lambda) \right\}$$

$$\leq \left(\frac{\bar{c}}{\alpha} + \frac{M\beta}{1-\Delta_\alpha} \right) \eta_n W(x), \qquad x \in X, \quad n \in \mathbb{N}, \tag{43}$$

which implies

$$\|TV - T_nV\|_W \leq M_1\eta_n, \quad n \in \mathbb{N}, \tag{44}$$

where $M_1 := \frac{\bar{c}_2}{\alpha} + \frac{M\beta}{1-\Delta_\alpha}$.

Combining (42) and (44) we obtain, for each $n \in\mathbb{N}$,

$$E_x^{\pi^1,\pi^2} \|V - L_n\|_W^{p'} \leq M_1^{p'} E_x^{\pi^1,\pi^2} \left[\eta_n^{p'} \right] + (\beta\gamma)^{p'} E_x^{\pi^1,\pi^2} \|V - L_{n-1}\|_W^{p'}. \tag{45}$$

Now, note that from (17) and (35), $l := \limsup_{n\to\infty} E_x^{\pi^1,\pi^2} \|V - L_n\|_W^{p'} < \infty$. Hence, since $\beta\gamma < 1$, taking lim sup as $n \to \infty$ in both sides of (45), we obtain,

$$l \leq \frac{M_1^{p'}}{1-(\beta\gamma)^{p'}} \lim_{n\to\infty} E_x^{\pi^1,\pi^2} \left[\eta_n^{p'} \right].$$

Finally, observing that $E_x^{\pi^1,\pi^2}[\eta_n] = E[\eta_n]$ (since ρ_n does not depend on $x \in X$ and $(\pi^1,\pi^2) \in \Pi^1 \times \Pi^2$), (27) yields the desired result.■

Proof of Theorem 4.1.

For each $n \in\mathbb{N}$, we define the function Φ_n as (see Remark 3.2)

$$\Phi_n(x,\mu,\lambda) := r_n(x,\mu,\lambda) + \Delta_n \int_X L_{n-1}(y)Q(dy|x,\mu,\lambda) - L_n(x). \tag{46}$$

Let $\pi^1 \in \Pi^1$ be an arbitrary strategy for player 1, and let $\{(x_n, a_n, g_n)\}$ be a sequence of state-actions triplets corresponding to application of $(\pi^1, \hat{\pi}^2)$. By the definition of the strategy $\hat{\pi}^2$ (see (37)) we have, for each $n \in \mathbb{N}$,

$$\Phi_n(x_n, a_n, g_n) \leq \sup_{\mu \in \mathbb{P}(A(x_n))} \left\{ r_n(x_n, \mu, g_n) + \Delta_n \int_X L_{n-1}(y) Q(dy|x_n, \mu, g_n) \right\} - L_n(x_n) = 0.$$

Thus, for each $n \in \mathbb{N}$,

$$\begin{aligned} \Phi(x_n, a_n, g_n) &\leq \Phi(x_n, a_n, g_n) - \Phi_n(x_n, a_n, g_n) \\ &\leq \sup_{\lambda \in \mathbb{P}(B(x_n))} \sup_{\mu \in \mathbb{P}(A(x_n))} |\Phi(x_n, \mu, \lambda) - \Phi_n(x_n, \mu, \lambda)| \\ &\leq W(x_n) \sup_{x \in X} [W(x)]^{-1} \sup_{\lambda \in \mathbb{P}(B(x))} \sup_{\mu \in \mathbb{P}(A(x))} |\Phi(x, \mu, \lambda) - \Phi_n(x, \mu, \lambda)| . \end{aligned} \tag{47}$$

On the other hand, from (21) and (46), (adding and subtracting the term $\Delta_n \int V(y) Q(dy \mid x, \mu, \lambda)$) and using (38), (39), (17) and (34), we get (see (43) and (44)), for each $x \in X$, $n \in \mathbb{N}$, $\mu \in \mathbb{P}(A(x))$ and $\lambda \in \mathbb{P}(B(x))$,

$$\begin{aligned} |\Phi(x, \mu, \lambda) - \Phi_n(x, \mu, \lambda)| &\leq |r(x, \mu, \lambda) - r_n(x, \mu, \lambda)| + |V(x) - L_n(x)| \\ &\quad + |\Delta_\alpha - \Delta_n| \int_X V(y) Q(dy \,|x, \mu, \lambda) \\ &\quad + \Delta_n \int_X |V(y) - L_{n-1}(y)| \, Q(dy \,|x, \mu, \lambda) \\ &\leq M_1 \eta_n W(x) + \|V - L_n\|_W W(x) \\ &\quad + \gamma \|V - L_{n-1}\|_W W(x). \end{aligned}$$

Hence, for each $n \in \mathbb{N}$,

$$\begin{aligned} &\sup_{x \in X} [W(x)]^{-1} \sup_{\lambda \in \mathbb{P}(B(x))} \sup_{\mu \in \mathbb{P}(A(x))} |\Phi(x, \mu, \lambda) - \Phi_n(x, \mu, \lambda)| \\ &\leq M_1 \eta_n + \|V - L_n\|_W + \gamma \|V - L_{n-1}\|_W , \end{aligned}$$

which combined with (47) yields

$$\begin{aligned} \Phi(x_n, a_n, g_n) &\leq M_1 \eta_n W(x_n) + \|V - L_n\|_W W(x_n) \\ &\quad + \gamma \|V - L_{n-1}\|_W W(x_n). \end{aligned} \tag{48}$$

Letting $M_2 := \sup_n \left(E_x^{\pi^1,\hat{\pi}^2} W^p(x_n)\right)^{1/p} < \infty$ (see (11)) and applying Holder's inequality in (48), we obtain,

$$E_x^{\pi^1,\hat{\pi}^2}\Phi(x_n,a_n,b_n) \leq M_2M_1\left(E_x^{\pi^1,\hat{\pi}^2}\eta_n^{p'}\right)^{1/p'} + M_2\left(E_x^{\pi^1,\hat{\pi}^2}\|V-L_n\|_W^{p'}\right)^{1/p'}$$
$$+ M_2\gamma\left(E_x^{\pi^1,\hat{\pi}^2}\|V-L_{n-1}\|_W^{p'}\right)^{1/p'}. \qquad (49)$$

To conclude, taking limit as $n \to \infty$ in (49) and observing that $E_x^{\pi^1,\hat{\pi}^2}[\eta_n] = E[\eta_n]$, Lemma 5.1 and (27) yield

$$\lim_{n\to\infty} E_x^{\pi^1,\hat{\pi}^2}\Phi(x_n,a_n,b_n) \leq 0 \quad \forall \pi^1 \in \Pi^1.$$

In addition, from the relation (22) and definition of the strategy π_*^1, we get

$$\lim_{n\to\infty} E_x^{\pi_*^1,\pi^2}\Phi(x_n,a_n,b_n) \geq 0 \quad \forall \pi^2 \in \Pi^2.$$

Thus, $(\pi_*^1, \hat{\pi}^2)$ is an asymptotically discount optimal pair of strategies.■

6 Example

We consider a storage system whose inputs are controlled in the following manner: at the time when an amount of product $M > 0$ accumulates for admission in the system, player 1 chooses a decision $a \in [a_*, 1] =: A$ $(0 < a_* < 1)$, that represents the portion of M to be admitted. On the other hand, there is a continuous consumption of the admitted product, controlled by the player 2. That is, at the time of each decision epoch, player 2 chooses a number $b \in [b_*, b^*] =: B$ $(0 < b_* < b^*)$ which represents the consumption rate per unit time. Thus, if $x_n \in X := [0,\infty)$ represents the stock level, a_n and b_n are the decisions of players 1 and 2, respectively, at the time of the nth decision epoch T_n, then the game evolves according to the equation

$$x_{n+1} = (x_n + a_nM - b_n\delta_{n+1})^+$$

with $\delta_{n+1} := T_{n+1} - T_n$ $(n = 0, 1, 2, ...)$. It is clear that the distribution of the holding time δ_{n+1} is independent of (x_n, a_n, b_n), and we assume that δ_n $(n = 1, 2, ...)$ has a density ρ that satisfies Assumptions 2.2 and 2.3. Moreover, the payoff function is given by

$$r(x,a,b) := \bar{d}b\tau_\alpha - D_1x - D_2a \qquad (50)$$

with $\bar{d}$, D_1, D_2 positive constants, and τ_α as in (4). We assume that the following is satisfied

Assumption 6.1 $E\delta > M/b_*$.

Let Ψ be the moment generating function of the random variable $M - b\delta$, that is:

$$\Psi(t) = E[\exp(t(M - b_*\delta))].$$

Then, Assumption 6.1 implies $\Psi'(0) < 0$, and since $\Psi(0) = 1$, there exists $\lambda > 0$ such that $\Psi(\lambda) < 1$. In addition, by the continuity of Ψ, we can choose $p > 1$ such that

$$\beta_0 := \Psi(p\lambda) = E[\exp(\lambda p(M - b_*\delta))] < 1.$$

Note that by the description of the system and (50), Assumption 2.4 (a), (b) are satisfied. Now, let $\bar{M}$ be a positive constant such that for each $x \in X$,

$$\max\{\bar{d}b^*, D_1 x + D_2\} \leq \bar{M}e^{\lambda x},$$

and define $W_0(x) := \bar{M}e^{\lambda x}$. Then, for $(x, a, b) \in \mathbb{K}$,

$$\begin{aligned} \int \bar{M}^p e^{\lambda p y} Q(dy \mid x, a, b) &= \int_0^\infty \bar{M}^p e^{\lambda p (x + aM - bs)^+} \rho(s) ds \\ &\leq \bar{M}^p P[x + aM - bs \leq 0] + \bar{M}^p e^{\lambda p x} \int_0^\infty e^{\lambda p (M - bs)} \rho(s) ds \\ &\leq \bar{M}^p + W_0^p(x) E[e^{\lambda p (M - b\delta)}] \leq \bar{M}^p + W_0^p(x) E[e^{\lambda p (M - b_*\delta)}] \\ &\leq \beta_0 W_0^p(x) + \bar{M}^p. \end{aligned}$$

Thus, Assumption 2.4 (c) is satisfied. To verify Assumption 2.4 (d), let v be a bounded measurable function on X, and for every $a \in A$ and $b \in B$, let $\rho_{(a,b)}$ be the density of $aM - b\delta$. Observe that

$$\rho_{(a,b)}(y) = \frac{1}{b}\rho(\frac{aM - y}{b}), \quad -\infty < y \leq aM.$$

In addition, for each $y \in \mathbf{R}$, $(a,b) \longmapsto \rho_{(a,b)}(y)$ is continuous on $A \times B$. Then,

$$\begin{aligned}\int_X v(y)Q(dy \mid x,a,b) &= \int_0^{\infty} v[(x+y)^+]\rho_{(a,b)}(y)dy \\ &= v(0)\int_{-\infty}^{-x} \rho_{(a,b)}(y)dy + \int_{-x}^{\infty} v(x+y)\rho_{(a,b)}(y)dy \\ &= v(0)\int_{-\infty}^{-x} \rho_{(a,b)}(y)dy + \int_{0}^{\infty} v(y)\rho_{(a,b)}(y-x)dy.\end{aligned}$$

Thus by Scheffé's Theorem,

$$(a,b) \longmapsto \int_X v(y)Q(dy \mid x,a,b)$$

defines a continuous function on $A \times B$. Finally, replacing $v(\cdot)$ by the function $W_0(\cdot)$ and using similar arguments, we obtain that Assumption 2.4 (d) holds.

References

[1] E.I. Gordienko, J.A. Minjárez-Sosa, *Adaptive control for discrete-time Markov processes with unbounded costs: discounted criterion,* Kybernetika 34 (1998), pp. 217–234.

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

[3] O. Hernández-Lerma, J.B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes,* Springer-Verlag, New York, 1999.

[4] R. Hasminskii, I. Ibragimov, *On density estimation in the view of Kolmogorov's ideas in approximation theory,* Ann. Statist., 18 (1990), 999-1010.

[5] N. Hilgert, J.A. Minjárez-Sosa, *Adaptive policies for time-varying stochastic systems under discounted criterion,* Math. Methods Oper. Res., 54 (2001), pp. 491-505.

[6] A. Jaskiewicz, *Zero-sum semi-Markov games,* SIAM J. Control Optim. 41 (2002), 723-739.

[7] A.K. Lal, S. Sinha, *Zero-sum two person semi-Markov games,* J. Appl. Prob. 29 (1992), 56-72.

[8] F. Luque-Vásquez, M.T. Robles-Alcaraz, *Controlled semi-Markov models with discounted unbounded costs,* Bol. Soc. Mat. Mexicana 39 (1994), 51-68.

[9] F. Luque-Vásquez, *Zero-sum semi-Markov games in Borel spaces: discounted and average payoff,* Bol. Soc. Mat. Mexicana 8 (2002), 227-241.

[10] F. Luque-Vásquez, J.A. Minjárez-Sosa, *Semi-Markov control processes with unknown holding times distribution under a discounted criterion,* To appear in Math. Methods Oper. Res.

[11] A.S. Nowak, *Some remarks on equilibria in semi-Markov games,* Appl. Math. (Warsaw) 27-4 (2000), 385-394.

[12] W. Polowczuk, *Nonzero semi-Markov games with countable state spaces,* Appl. Math. (Warsaw) 27-4 (2000), 395-402.

[13] U. Rieder, *Measurable selection theorems for optimization problems,* Manuscripta Math. 24 (1978), 115–131.

[14] M. Schäl, *Estimation and control in discounted stochastic dynamic programming,* Stochastics 20 (1987), 51-131.

[15] L Shapley, *Stochastic games,* Proc. Natl. Acad. Sci. U.S.A. 39 (1953), 1095-1100.

[16] O. Vega-Amaya, *Average optimality in semi-Markov control models on Borel spaces: unbounded costs and controls,* Bol. Soc. Mat. Mexicana 38 (1993), 47-60.

[17] O. Vega-Amaya, *Zero-sum semi-Markov games: Fixed point solutions of the Shapley equation,* SIAM J. Control Optim. 42-5 (2003), 1876-1894.