Rolling horizon procedures in Semi-Markov Games: The Discounted Case

Eugenio Della Vecchia, Silvia C. Di Marco, Alain Jean-Marie

Rolling horizon procedures in Semi-Markov Games: The Discounted Case

Eugenio Della Vecchia, Silvia Di Marco, Alain Jean-Marie

RESEARCH REPORT
N° 8019
July 2012
Project-Team Maestro

ISSN 0249-6399   ISRN INRIA/RR--8019--FR+ENG

Rolling horizon procedures in Semi-Markov Games: The Discounted Case

Eugenio Della Vecchia*, Silvia Di Marco†, Alain Jean-Marie‡

Project-Team Maestro

Research Report n° 8019 — version 2 — initial version July 2012 — revised version April 2014 — 23 pages

Abstract: We study the properties of the rolling horizon and the approximate rolling horizon procedures for the case of two-person zero-sum discounted semi-Markov games with infinite horizon, when the state space is a borelian set and the action spaces are considered compact. Under suitable conditions, we prove that the equilibrium is the unique solution of a dynamic programming equation, and we prove bounds which imply the convergence of the procedures when the horizon length tends to infinity. The approach is based on the formalism for Semi-Markov games developed by Luque-Vásquez in [11], together with extensions of the results of Hernández-Lerma and Lasserre [4] for Markov Decision Processes and Chang and Marcus [2] for Markov Games, both in discrete time. In this way we generalize the results on the rolling horizon and approximate rolling horizon procedures previously obtained for discrete-time problems.

Key-words: Semi-Markov games, Rolling horizon procedures

* CONICET - UNR, Argentina

† CONICET - UNR, Argentina

‡ INRIA and LIRMM, CNRS/Université Montpellier 2, 161 Rue Ada, F-34392 Montpellier, Alain.Jean-Marie@inria.fr.

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Procédures d'horizon roulant dans les jeux semi-Markoviens: le cas actualisé

Résumé : Nous étudions les propriétés de la procédure de décision à horizon roulant et une approximation de cette procédure, pour le cas de jeux semi-Markoviens à somme nulle avec horizon infini et actualisation, quand l'espace d'états est un ensemble borélien et les espaces d'actions sont compacts. Sous des hypothèses appropriées, nous montrons que l'équilibre est l'unique solution de l'équation de programmation dynamique associée au jeu, puis nous prouvons des bornes d'erreur impliquant la convergence des procédures quand l'horizon de programmation tend vers l'infini. Notre approche est basée sur le formalisme pour les jeux semi-Markoviens développé par Luque-Vásquez [11], joint à des extensions des résultats de Hernández-Lerma et Lasserre [4] pour les processus de décision Markoviens et Chang et Marcus [2] pour les jeux Markoviens, ces deux derniers travaux étant en temps discret. De cette façon, nous généralisons les résultats sur la procédure à horizon roulant obtenus pour les problèmes en temps discret.

Mots-clés : Jeux semi-Markoviens, procédure à horizon roulant

In memoriam Silvia Di Marco
1964–2014

1 Introduction

In this work we analyze two approximation procedures applied to zero-sum semi-Markov games with the expected total discounted reward as the performance criterion. Specifically, we work with methods derived from the Rolling Horizon procedure.

Semi-Markov Games (**SMG**) generalize Markov Games (**MG**) by allowing the decision maker to choose actions whenever the system state changes, modeling the system evolution in continuous-time and allowing the time spent in a particular state to follow an arbitrary probability distribution. The system state may change several times between decision epochs but only the state at a decision epoch is relevant to the decision maker. **SMG** with discounted reward are analyzed in [11] and later, in [9, 12]. All these papers deal with the characterization of the value function as the fixed point of certain dynamical programming operators.

On the other hand, Rolling Horizon (**RH**) control is an usual procedure for making decisions in many infinite stage decision problems. It is based on choosing the best most immediate action based on the knowledge of the information of the problem just for a certain number of periods in the future. One design issue of the controller will be then to determine how many periods in the future must be taken into account, in order to make the optimal immediate decision [16]. **RH** strategies are largely used in several areas: we can mention here production control problems, stabilization of control systems, and macro-planning problems. The study of this and other applications can be found in [10].

The objective of the present work is to extend the analysis of the accuracy of the **RH** method and of an Approximate Rolling Horizon (**ARH**) method to **SMG** with the total discounted reward criterion, when the state space is assumed to be Borel, and the action spaces compact. As mentioned previously, the **SMG** structure was studied from the theoretical viewpoint by [11] and we borrow some of their results. On the one hand, [4, 5, 2] can be seen as particular cases of our results shown here because the two first investigate the accuracy of the **RH** procedure for discrete-time Markov Decision Processes (**MDP**) (just one decision-maker) with bounded and unbounded rewards functions, respectively, and the last one deals with discrete-time zero-sum Markov games with finite state spaces. The present work also extends our previous work [3] where we study the **RH** procedure applied to semi-Markov Decision Processes (**SMDP**) (continuous temporal evolution of the problem where just one decision maker acts). Also as a particular case, all the results obtained here apply to continuous-time **MG** where, up to our knowledge, **RH** and **ARH** have not been applied so far.

This paper is organized as follows. In Section 2, we present the **SMG** model, introduce its notation and state the assumptions on the data of the problem. In Section 3 we present the performance criterion and the dynamic programming operator for this case, mentioning the results on the associated optimality equation and recursion scheme. We prove characterization and optimality results under assumptions that are, up to our knowledge, more general than those used in the literature.

Section 4 contains our contributions about the convergence of the values of **RH** and **ARH** policies to the values of the game. The approach in this section is based on [5] where the case of discrete-time **MDP** is treated, and on [2] for the case of finite state spaces. As compared with this last work, we have not only proved results of the **ARH** for more general state spaces and possibly unbounded cost functions, but we also have improved significantly the error bounds. Finally, Section 5 is devoted to the concluding remarks.

2 Preliminaries and Notations

We consider a semi-Markov game of the form

$$M := (\mathcal{S}, \mathcal{A}, \mathcal{B}, \{\mathcal{A}_s : s \in \mathcal{S}\}, \{\mathcal{B}_s : s \in \mathcal{S}\}, Q, F, \ell, \alpha)$$

where

- $\mathcal{S}$ is the state space.
- $\mathcal{A}$ and $\mathcal{B}$ are the action spaces for players 1 and 2 respectively.
- For every $s \in \mathcal{S}$, we define the sets $\mathcal{A}_s$ and $\mathcal{B}_s$ as the sets of actions available in state s for respective players, and, therefore
$$\mathcal{A} = \bigcup_{s \in \mathcal{S}} \mathcal{A}_s, \quad \mathcal{B} = \bigcup_{s \in \mathcal{S}} \mathcal{B}_s.$$
- We set $\mathbb{K} = \{(s, a, b) : s \in \mathcal{S}, a \in \mathcal{A}_s, b \in \mathcal{B}_s\}$.
- The transition law $Q(\cdot|\cdot)$, is a stochastic kernel on $\mathcal{S}$ given $\mathbb{K}$, and $F(\cdot|s, a, b)$ is the distribution function of the holding time in state $s \in \mathcal{S}$ when actions $a \in \mathcal{A}_s$ and $b \in \mathcal{B}_s$ are chosen.
- The reward function is $\ell : \mathbb{K} \mapsto \mathbb{R}$ and α is a discount factor.

If at time of the n-th decision epoch, the state of the system is $s_n = s$, and the chosen actions are $a_n \in \mathcal{A}_s$ and $b_n \in \mathcal{B}_s$, then the system remains in the state s during a nonnegative random time δ_{n+1} with distribution $F(\cdot|s_n, a_n, b_n)$ and a continuously discounted, stationary reward $\ell(s_n, a_n, b_n)$ is received by player 1 and paid by player 2. The decision epochs are therefore $T_n := T_{n-1} + \delta_n$ for $n \geqq 1$, and $T_0 = 0$. The random variable $\delta_{n+1} = T_{n+1} - T_n$ is called the sojourn or holding time at stage n.

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

The space $\mathcal{H}_n := (\mathbb{K} \times \mathbb{R}^+)^n \times \mathcal{S}$ of admissible histories of the process at the n-th decision epoch, consists of sequences of states, decisions and sojourn times up to that epoch. At the initial epoch T_0, the history consists of the initial state $s_0 \in \mathcal{S}$. At the first decision epoch T_1, the two initial actions chosen by the players, the holding time at initial state and the new state are added to the initial state, and so on. A typical element of $\mathcal{H}_n$ is therefore written as

$$h_n = (s_0, a_0, b_0, \delta_1, s_1, a_1, b_1, \delta_2, \ldots, s_{n-1}, a_{n-1}, b_{n-1}, \delta_n, s_n)\ .$$

A strategy for player 1 is a sequence $\pi = \{\pi_0, \pi_1, \pi_2, \ldots\}$ where π_n is a probability measure on its action set $\mathcal{A}_{s_n}$, given the whole history h_n up to time n. A *Markov* strategy (again for player 1) is a strategy of the form $\pi = \{f_0, f_1, f_2, \ldots\}$, where each measure f_t is allowed to depend only on the current state s_t and the stage parameter t. In particular, when each f_t is also independent on t, the strategy it is said *stationary.*

Formalizing the previous concepts, a strategy for player 1 (resp. player 2) is a sequence $\pi = \{\pi_n\}$ (resp. $\gamma = \{\gamma_n\}$) of stochastic kernels $\pi_n \in \mathbb{P}(\mathcal{A}|\mathcal{H}_n)$ (resp. $\gamma_n \in \mathbb{P}(\mathcal{B}|\mathcal{H}_n)$), such that for every $h_n \in \mathcal{H}_n$ and $n \in \mathbb{N}$, $\pi_n(\mathcal{A}_{s_n}|h_n) = 1$ (resp. $\gamma_n(\mathcal{B}_{s_n}|h_n) = 1$).

A strategy π for player 1 is a Markov strategy (or Markov policy) if there exists a sequence $\{f_n\}$ of stochastic kernels $f_n \in \mathbb{P}(\mathcal{A}|\mathcal{S})$ such that for every $n \in \mathbb{N}$, $h_n \in \mathcal{H}_n$ and $\tilde{\mathcal{A}} \subset \mathcal{A}_{s_n}$, $\pi_n(\tilde{\mathcal{A}}|h_n) = f_n(\tilde{\mathcal{A}}|s_n)$. Likewise for player 2: there exists $\{g_n\}$ such that $\gamma_n(\tilde{\mathcal{B}}|h_n) = g_n(\tilde{\mathcal{B}}|s_n)$, for all $\tilde{\mathcal{B}} \subset \mathcal{B}_{s_n}$. Here, $f_n(\tilde{\mathcal{A}}|s_n)$ represents the probability that player 1 chooses actions in $\tilde{\mathcal{A}}$, at stage n in state s_n; likewise for player 2 with $g_n(\tilde{\mathcal{B}}|s_n)$. We denote with Π (resp. Γ) the set of all Markov strategies of player 1 (resp. 2) and with some abuse of notation, we write $\pi = \{f_n\}$ and $\gamma = \{g_n\}$.

For player 1, a Markov strategy $\pi = \{f_n\}$ is stationary if there exists $f \in \mathbb{P}(\mathcal{A}|\mathcal{S})$ such that $f(s) \in \mathbb{P}(\mathcal{A}_s)$ and $f_n = f$ for all $s \in \mathcal{S}$ and $n \in \mathbb{N}$. In this case, we identify π with f, i.e., $\pi = f = \{f, f, ...\}$. We denote by Π_{stat} the set of all stationary strategies for player 1. Similarly, for player 2, a Markov strategy γ is stationary if there exists $g \in \mathbb{P}(\mathcal{B}|\mathcal{S})$ such that $g(s) \in \mathbb{P}(\mathcal{B}_s)$ and $g_n = g$ for all $s \in \mathcal{S}$ and $n \in \mathbb{N}$. For player 2, we denote Γ_{stat} the set of all stationary strategies.

For each pair of strategies $(\pi, \gamma) \in \Pi \times \Gamma$, and any initial state s there exist a unique probability measure $\mathrm{P}_s^{\pi,\gamma}$ and stochastic processes $\{S_t\}$, $\{A_t\}$, $\{B_t\}$ and $\{\delta_t\}$. S_t, A_t and B_t represent the state and the actions at the t-th decision epoch. $\mathbb{E}_s^{\pi,\gamma}$ denotes the expectation operator with respect to $\mathrm{P}_s^{\pi,\gamma}$.

We note

$$\beta(s, a, b) := \int_0^\infty e^{-\alpha t} F(\mathrm{d}t|s, a, b) \tag{1}$$

the Laplace transform of δ, conditioned on (s, a, b) and evaluated at α, and

$$\vartheta(s, a, b) = \frac{1 - \beta(s, a, b)}{\alpha} \ . \tag{2}$$

From here on, we make the following abuse of notation: for each $s \in \mathcal{S}$ and given a pair of probability distributions ξ and ζ on $\mathcal{A}_s$ and $\mathcal{B}_s$ respectively, $\int_{\mathcal{A}_s} \int_{\mathcal{B}_s} h(s, a, b)\zeta(\mathrm{d}b)\xi(\mathrm{d}a)$ whenever the integral is well defined, will be denoted by $h(s, \xi, \zeta)$. Also, for a function ϕ defined on $\mathbb{K}$, we note $\phi(s, f, g)$ instead of $\phi(s, f(s), g(s))$, for given stationary strategies $(f, g) \in \Pi_{\text{stat}} \times \Gamma_{\text{stat}}$.

In order to evaluate the performance of policies, we use a total discounted criterion. We assume that the rewards are continuously discounted, with a discount factor α. More precisely let, for $n \geqq 1$, $s \in \mathcal{S}$, $\pi \in \Pi$ and $\gamma \in \Gamma$, the expected n-stage α-discounted reward be defined as follows:

$$\begin{aligned} V_n^{\pi,\gamma}(s) \quad &:= \quad \mathbb{E}_s^{\pi,\gamma} \sum_{k=0}^{n-1} \int_{T_k}^{T_{k+1}} e^{-\alpha t} \ell(S_k, A_k, B_k) \mathrm{d}t \\ &= \quad \mathbb{E}_s^{\pi,\gamma} \sum_{k=0}^{n-1} e^{-\alpha T_k} \frac{1 - e^{-\alpha \delta_{k+1}}}{\alpha} \ \ell(S_k, A_k, B_k) \ . \end{aligned}$$

The infinite-horizon total expected α-discounted payoff is

$$V^{\pi,\gamma}(s) \ := \ \mathbb{E}_s^{\pi,\gamma} \sum_{k=0}^{\infty} e^{-\alpha T_k} \frac{1 - e^{-\alpha \delta_{k+1}}}{\alpha} \ \ell(S_k, A_k, B_k) \ . \tag{3}$$

Alternatively, given a pair of policies π and γ, we can write its reward using the variables β and ϑ and obtain for the finite stage horizon and for the infinite horizon respectively,

$$V^{\pi,\gamma}(s) = \mathbb{E}_s \left[\sum_{t=0}^{\infty} \left(\prod_{k=0}^{t-1} \beta(S_k, A_k, B_k) \right) \vartheta(S_t, A_t, B_t) \ell(S_t, A_t, B_t) \right]$$

$$V_n^{\pi,\gamma}(s) = \mathbb{E}_s\left[\sum_{t=0}^{n-1}\left(\prod_{k=0}^{t-1}\beta(S_k,A_k,B_k)\right)\vartheta(S_t,A_t,B_t)\ell(S_t,A_t,B_t)\right] ,$$

where we adopt the usual conventions that $\prod_{k=0}^{-1} X_k = 1$ and $\sum_{t=0}^{-1} Y_t = 0$. Briefly, the first expression above comes from the following considerations.

$$\begin{aligned} V^{\pi,\gamma}(s) &= \mathbb{E}_s^{\pi,\gamma}\left[\sum_{t=0}^{\infty}\int_{T_t}^{T_t+\delta_{t+1}} e^{-\alpha u}\mathrm{d}u\ \ell(S_t,A_t,B_t)\right] \\ &= \mathbb{E}_s^{\pi,\gamma}\left[\sum_{t=0}^{\infty} e^{-\alpha T_t}\frac{1-e^{-\alpha\delta_{t+1}}}{\alpha}\ \ell(S_t,A_t,B_t)\right] \\ &= \mathbb{E}_s^{\pi,\gamma}\left[\sum_{t=0}^{\infty} e^{-\alpha(\sum_{k=0}^{t}\delta_k)}\frac{1-e^{-\alpha\delta_{t+1}}}{\alpha}\ \ell(S_t,A_t,B_t)\right] \\ &= \mathbb{E}_s^{\pi,\gamma}\left[\sum_{t=0}^{\infty}\left(\prod_{k=0}^{t-1}\beta(S_k,A_k,B_k)\right)\vartheta(S_t,A_t,B_t)\ell(S_t,A_t,B_t)\right] . \end{aligned}$$

Similarly, the second one can be justified.

At this point, observe that we can work with an instantaneous one-step reward functions r: $\mathbb{K}\to\mathbb{R}$ defined by $r(s,a,b) = \vartheta(s,a,b)\ell(s,a,b)$. We obtain the new expressions

$$V^{\pi,\gamma}(s) = \mathbb{E}_s^{\pi,\gamma}\left[\sum_{t=0}^{\infty}\left(\prod_{k=0}^{t-1}\beta(S_k,A_k,B_k)\right)r(S_t,A_t,B_t)\right] , \tag{4}$$

$$V_n^{\pi,\gamma}(s) = \mathbb{E}_s^{\pi,\gamma}\left[\sum_{t=0}^{n-1}\left(\prod_{k=0}^{t-1}\beta(S_k,A_k,B_k)\right)r(S_t,A_t,B_t)\right] . \tag{5}$$

We now state our working assumptions. A first set of "general" assumptions bears on the distribution probabilities of the holding time, the instantaneous reward function and the transition kernels.

Assumption 1.

(a) The state space $\mathcal{S}$ is a Borel subset of a complete and separable metric space.

(b) For each $s\in\mathcal{S}$, the sets $\mathcal{A}_s$ and $\mathcal{B}_s$ are compact.

(c) For each $s\in\mathcal{S}$, and $b\in\mathcal{B}_s$, $\ell(s,\cdot,b)$ is upper semi-continuous on $\mathcal{A}_s$.

(d) For each $s\in\mathcal{S}$, and $a\in\mathcal{A}_s$, $\ell(s,a,\cdot)$ is lower semi-continuous on $\mathcal{B}_s$.

(e) For each $s\in\mathcal{S}$ and each bounded measurable function v on $\mathcal{S}$, the function $(a,b)\mapsto \int v(y)Q(dy|s,a,b)$ is continuous on $\mathcal{A}_s\times\mathcal{B}_s$.

(f) The function $(s,a,b)\mapsto\int_0^\infty tF(dt|s,a,b)$ is continuous on $\mathbb{K}$.

Remark 2.1. In view of Equations (1) and (2), **Assumption 1** (c)-(d) implies

(c') For each $s\in\mathcal{S}$, and $b\in\mathcal{B}_s$, $r(s,\cdot,b)$ is upper semi-continuous on $\mathcal{A}_s$.

(d') For each $s\in\mathcal{S}$, and $a\in\mathcal{A}_s$, $r(s,a,\cdot)$ is lower semi-continuous on $\mathcal{B}_s$.

It also implies the measurability of these functions, as required in the definition of $V_n^{\pi,\gamma}$ and $V^{\pi,\gamma}$. ■

Assumption 2. $\rho := \sup_{(s,a,b)\in\mathbb{K}} \beta(s,a,b) < 1$.

The following result has been extracted from [11, Lemma 4.1].

Proposition 2.1. *If there exists a pair of positive numbers θ and ϵ such that*

$$F(\theta|s,a,b) \leqq 1-\epsilon$$

*for all $(s,a,b)\in\mathbb{K}$, then **Assumption 2** holds with $\rho = 1-\epsilon+\epsilon e^{\alpha\theta}$.*

Remark 2.2. The theory of **SMDP** can be found in [1] for the case of Borel state spaces, and in [15] and [14] for the case of discrete and finite state spaces, respectively. All these references work with **Assumption 1** and with the assumption stated in Proposition 2.1, restricted to only one player.

Items (a)-(e) in **Assumption 1** are those considered in [11]. Item (f), adopted from [8, 9], replaces the more restrictive assumption presented in the first paper, which asks, for all t, for the continuity of $F(t|s,a,b)$ on $\mathbb{K}$. This stronger assumption, restricted to one player, is also made in [1] and [3] for the **SMDP** case. ■

Let us denote with $\mathcal{M}(\mathcal{S})$ the space of measurable functions on $\mathcal{S}$, and $\mathcal{M}_+(\mathcal{S})$ the subspace of nonnegative functions of $\mathcal{M}(\mathcal{S})$. On $\mathcal{M}_+(\mathcal{S})$, we define the operator L which maps v to Lv in the following way: for $s\in\mathcal{S}$,

$$(Lv)(s) := \sup_{a\in\mathcal{A}_s, b\in\mathcal{B}_s} \int_{\mathcal{S}} v(z)Q(\mathrm{d}z|s,a,b).$$

Let $L^n v = L(L^{n-1}v)$ for $n\in\mathbb{N}$, with $L^0 v = v$. Clearly, $L(\lambda v) = \lambda L(v)$ for every positive scalar λ.

We now state assumptions bearing on the cost function and the way it interacts with transition kernels.

Assumption 3. *Let $R:\mathcal{S}\mapsto\mathbb{R}$ be defined by*

$$R(s) := \sum_{t=0}^{\infty} \rho^t (L^t r_0)(s),$$

where $r_0(s) = \sup_{a\in\mathcal{A}_s, b\in\mathcal{B}_s} |r(s,a,b)|$. Assume $R(s)<\infty, \forall s\in\mathcal{S}$.

If $\mu\in\mathcal{M}_+(\mathcal{S})$ is strictly positive, for $v\in\mathcal{M}(\mathcal{S})$ we define the μ-weighted norm $||v||_\mu = \sup_{s\in\mathcal{S}} |v(s)|/\mu(s)$.

Assumption 4. *There exist a measurable function $\mu: S\to[1,\infty)$ and a positive constant m such that for all $(s,a,b)\in\mathbb{K}$,*

(a) $|r(s,a,b)| \leqq m\,\mu(s)$,

(b) $\int \mu(z)Q(dz|s,a,b) \leqq \mu(s)$.

Assumption 5. *There exists $M>0$ such that $|r(s,a,b)|\leqq M$ for all $(s,a,b)\in\mathbb{K}$.*

Remark 2.3. An assumption similar to **Assumption 3** appears in [1] for the **SMDP** case and in [5] for the **MDP** case, both for a single player, while **Assumption 4** could be associated to similar ones in [14, Chapter 6] and in [6, Chapter 2], both of them for **MDP**. ■

Remark 2.4. As a consequence of **Assumption 4** (b), we have, for all $s \in \mathcal{S}$ and all $v \in \mathcal{M}(\mathcal{S})$,

$$\begin{aligned} \left| \int_{\mathcal{S}} v(z) Q(\mathrm{d}z|s,a,b) \right| &\leqq \int_{\mathcal{S}} \frac{|v(z)|}{\mu(z)} \, \mu(z) Q(\mathrm{d}z|s,a,b) \leqq ||v||_\mu \int_{\mathcal{S}} \mu(z) Q(\mathrm{d}z|s,a,b) \\ &\leqq ||v||_\mu \, \mu(s) \ . \end{aligned} \tag{6}$$

With the definition of r_0 introduced in **Assumption 3**, **Assumption 4 (a)** can be written $||r_0||_\mu \leqq m$. **Assumption 4 (b)** can be written $L\mu \leqq \mu$ equivalently. If r verifies **Assumption 5**, then by setting $\mu \equiv 1$ and $m = M$, **Assumption 4** is satisfied.

In [5] and in [7, Proposition 4.3.1., p. 53], it is shown that **Assumption 4** implies **Assumption 3** for the case of **MDP**. A similar argument is valid for **SMG**s. Indeed, with (6),

$$\begin{aligned} |(L^t r_0)(s)| &= \left| \sup_{a \in \mathcal{A}_s, b \in \mathcal{B}_s} \int_{\mathcal{S}} (L^{t-1} r_0)(z) Q(\mathrm{d}z|s,a,b) \right| \\ &\leqq \sup_{a \in \mathcal{A}_s, b \in \mathcal{B}_s} \left| \int_{\mathcal{S}} (L^{t-1} r_0)(z) Q(\mathrm{d}z|s,a,b) \right| \\ &\leqq ||L^{t-1} r_0||_\mu \, \mu(s) \ , \end{aligned}$$

for all $t \geqq 1$. By induction, for any t,

$$||L^t r_0||_\mu \leqq ||L^{t-1} r_0||_\mu \leqq \cdots \leqq ||r_0||_\mu \leqq m$$

and then $||R||_\mu \leqq \frac{m}{1-\rho}$, or $R \leqq \frac{m}{1-\rho}\mu$, implying the finiteness of R. ■

Remark 2.5. This semi-Markov environment covers two important special cases:

1. Discrete-time models. In this case $F(\cdot|s,a,b) = \delta_1(\cdot)$ for all $(s,a,b) \in \mathbb{K}$. This corresponds to the theory of **MG**s. Such a function $F(\cdot)$ satisfies **Assumption 1 (f)**.

2. Continuous-time Markov models. This arises if the holding time distributions are exponential: $F(\mathrm{d}u|s,a,b) = \delta(s,a,b)e^{-\delta(s,a,b)u}\mathrm{d}u$, where $\delta(s,a,b)$ is a continuous function from $\mathbb{K}$ into $(0,\infty)$ (continuity is required for satisfying **Assumption 1 (f)**). The process $\{S_t\}$ turns out to be a Markov process when (π,γ) is a pair of Markov policies, and a time-homogeneous Markov process when (π,γ) is a pair of stationary policies.

■

3 Performance Criterion and Related Results

Under suitable hypotheses, in this section we characterize the value function for the finite and infinite-horizon games and the strategies which produce the equilibria through operators defined in (9) and (10) below. We also analyze the convergence of the values for the finite-horizon games to that of the infinite-horizon one. These results will be used to prove the convergence of **RH** and **ARH** procedures described in the next section.

The lower and the upper value functions of the infinite-horizon game are defined, as usual, for $s \in \mathcal{S}$, as

$$\underline{V}^*(s) = \sup_{\pi \in \Pi} \inf_{\gamma \in \Gamma} V^{\pi,\gamma}(s) \quad \text{and} \quad \overline{V}^* = \inf_{\gamma \in \Gamma} \sup_{\pi \in \Pi} V^{\pi,\gamma}(s) \tag{7}$$

respectively. We know that, in general, $\underline{V}^* \leqq \overline{V}^*$.

If $\underline{V}^* = \overline{V}^*$, and both the "sup inf" and "inf sup" are attained, we refer to this common value as the value of the game, and we note it with V^* (we refer to [13] for definitions on general dynamic zero-sum games). Likewise for the finite-horizon games.

Suppose that our games have a value. Then, the objective of the players is to find (when it exists) a pair of policies that solves, given the current state s:

$$(\pi^*(s), \gamma^*(s)) \in \arg\max_{\pi} \min_{\gamma} V^{\pi,\gamma}(s). \tag{8}$$

Such a pair of strategies $(\pi^*, \gamma^*) \in \Pi \times \Gamma$ is said to be an equilibrium.

With $R(\cdot)$ being defined in **Assumption 3**, we denote

- $\mathcal{R} = \{v \in \mathcal{M}(\mathcal{S}) : |v(s)| \leqq R(s) \text{ for all } s \in \mathcal{S}\}$;
- $\mathcal{M}_\mu(\mathcal{S})$ the linear subspace of $\mathcal{M}(\mathcal{S})$ of the functions with finite μ-weighted norm. This is a Banach space.

Define the operator $T : \mathcal{M}(\mathcal{S}) \mapsto \mathcal{M}(\mathcal{S})$ by

$$(Tv)(s) := \sup_{a\in\mathcal{A}_s} \inf_{b\in\mathcal{B}_s} \left\{ r(s,a,b) + \beta(s,a,b) \int_{\mathcal{S}} v(z) Q(\mathrm{d}z|s,a,b) \right\} , \tag{9}$$

and, given a pair of stationary policies $f \in \Pi_{\text{stat}}$ and $g \in \Gamma_{\text{stat}}$, $T^{f,g} : \mathcal{M}(\mathcal{S}) \mapsto \mathcal{M}(\mathcal{S})$

$$(T^{f,g}v)(s) := r(s,f,g) + \beta(s,f,g) \int_{\mathcal{S}} v(z) Q(\mathrm{d}z|s,f,g) . \tag{10}$$

We first state general properties of these operators.

Lemma 3.1. *Under* **Assumptions 1**, **2** *and* **3**, *the operator T maps $\mathcal{R}$ to itself. Under* **Assumptions 1**, **2** *and* **4**, *T maps $\mathcal{M}_\mu(\mathcal{S})$ to itself. If v is a bounded function, then Tv is also bounded. The same three properties hold for the operators $T^{f,g}$.*

Proof. First, observe that $\rho(LR)(s) = \sum_{t=1}^{\infty} \rho^t (L^t r_0)(s) = R(s) - r_0(s)$. Next, note that for all $v \in \mathcal{R}$,

$$|Tv(s)| \leqq r_0(s) + \rho(L|v|)(s) \leqq r_0(s) + \rho LR(s) = R(s),$$

which implies $Tv \in \mathcal{R}$. On the other hand, under **Assumption 4**, we can use (6) in (10) to obtain $||Tv||_\mu \leqq m + \rho||v||_\mu$, which implies in turn that $Tv \in \mathcal{M}_\mu(\mathcal{S})$. □

The following results provide the existence of optimal stationary strategies, then bounds and convergence results related to the operator L. First, under **Assumptions 1**, **2** and **3**, the sup inf and inf sup are attained in (9) for each $s \in \mathcal{S}$.

Lemma 3.2. *Under* **Assumptions 1, 2** *and* **3**, *for each $v \in \mathcal{R}$ there exist stationary strategies $\tilde{f} \in \Pi_{\text{stat}}$, $\tilde{g} \in \Gamma_{\text{stat}}$ such that*

$$\begin{aligned} (Tv)(s) &= r(s,\tilde{f},\tilde{g}) + \beta(s,\tilde{f},\tilde{g}) \int_{\mathcal{S}} v(z) Q(dz|s,\tilde{f},\tilde{g}) \\ &= \sup_{f\in\Pi_{\text{stat}}} \left\{ r(s,f,\tilde{g}) + \beta(s,f,\tilde{g}) \int_{\mathcal{S}} v(z) Q(dz|s,f,\tilde{g}) \right\} \\ &= \inf_{g\in\Gamma_{\text{stat}}} \left\{ r(s,\tilde{f},g) + \beta(s,\tilde{f},g) \int_{\mathcal{S}} v(z) Q(dz|s,\tilde{f},g) \right\} . \end{aligned} \tag{11}$$

Proof. According to Lemma 3.1, for each $v \in \mathcal{R}$, $Tv \in \mathcal{R}$. The proof makes use of well-known measurable selection theorems and is similar to that of Lemma 5.1 in [11], but with the weaker **Assumption 1 (f)**, which is enough to guarantee the existence of the pair of policies stated. □

Lemma 3.3. *Suppose that* **Assumptions 1, 2** *and* **3** *hold. Then, for all $s \in \mathcal{S}$, $\pi \in \Pi$, $\gamma \in \Gamma$, $v \in \mathcal{R}$ and $t \in \mathbb{N}$, we have:*

(a)

$$\mathbb{E}_s^{\pi,\gamma}[v(S_t)] \leqq (L^t v)(s). \tag{12}$$

(b)

$$\lim_{n\to\infty} \mathbb{E}_s^{\pi,\gamma}\left[\left(\prod_{k=0}^{n-1}\beta(S_k, A_k, B_k)\right) v(S_n)\right] = 0. \tag{13}$$

Proof. We prove (a) by induction on t. Consider the policies $\pi = \{f_0, f_1, \ldots\}$ and $\gamma = \{g_0, g_1, \ldots\}$. Since we are dealing with Markov policies $\pi \in \Pi$ and $\gamma \in \Gamma$, by [6, Properties 2.5 e and f, pp. 5-6], we have

$$\mathbb{E}_s^{\pi,\gamma}\left[v(S_{t+1})|S_0 = s, A_0 = a, B_0 = b, S_1 = z\right] = \mathbb{E}_z^{\pi',\gamma'}[v(S_t)] \tag{14}$$

where we denote with $\pi' = \{f'_0, f'_1, ...\}$ and $\gamma' = \{g'_0, g'_1, \ldots\}$ the shifted policies defined by $f'_t = f_{t+1}$ and $g'_t = g_{t+1}$.
For $t = 0$ the result is obvious. For $t = 1$, since $\int_{\mathcal{S}} v(z)Q(\mathrm{d}z|s, a, b) \leqq (Lv)(s)$,

$$\mathbb{E}_s^{\pi,\gamma}[v(S_1)] = \int_{\mathcal{A}}\int_{\mathcal{B}}\int_{\mathcal{S}} v(z)Q(\mathrm{d}z|s, a, b)g_0(\mathrm{d}b)f_0(\mathrm{d}a) \leqq (Lv)(s).$$

In general, using (14), we have for any t:

$$\begin{aligned}
\mathbb{E}_s^{\pi,\gamma}[v(S_{t+1})] \;&=\; \int_{\mathcal{A}}\int_{\mathcal{B}}\int_{\mathcal{S}} \mathbb{E}^{\pi,\gamma}\left[v(S_{t+1})|S_0 = s, A_0 = a, B_0 = b, S_1 = z\right] \\
&\qquad\qquad Q(\mathrm{d}z|s, a, b)g_0(\mathrm{d}b)f_0(\mathrm{d}a) \\
&=\; \int_{\mathcal{A}}\int_{\mathcal{B}}\int_{\mathcal{S}} \mathbb{E}_z^{\pi',\gamma'}[v(S_t)]Q(\mathrm{d}z|s, a, b)g'_0(\mathrm{d}b)f'_0(\mathrm{d}a) \\
&\leqq\; \int_{\mathcal{A}}\int_{\mathcal{B}}\int_{\mathcal{S}} (L^t v)(z)Q(\mathrm{d}z|s, a, b)g'_0(\mathrm{d}b)f'_0(\mathrm{d}a) \leqq (L^{t+1}v)(s).
\end{aligned}$$

For the proof of (b), we have:

$$\begin{aligned}
\mathbb{E}_s^{\pi,\gamma}\left[\left(\prod_{k=0}^{n-1}\beta(S_k, A_k, B_k)\right) v(S_n)\right] \;&\leqq\; \rho^n \mathbb{E}_s^{\pi,\gamma} v(S_n) \;\leqq\; \rho^n (L^n v)(s) \\
&\leqq\; \rho^n (L^n R)(s) \;=\; \sum_{t=n}^{\infty} \rho^t (L^t r_0)(s) \;\to\; 0,
\end{aligned}$$

as $n \to \infty$, by **Assumption 3**. □

The next result generalizes Theorems 4.3 and 4.4 from [11], proved under the more restrictive **Assumption 1** (see Remark 2.2), and **Assumptions 2** and **4**.

Theorem 3.1. *Suppose that* **Assumptions 1, 2** *and* **3** *hold. Then*

(a) For all $\pi \in \Pi$ and $\gamma \in \Gamma$, $V^{\pi,\gamma}, \overline{V}^*$ and $\underline{V}^*$ are in $\mathcal{R}$.

(b) The finite-stage games have values. We have $V_0^* \equiv 0$ and, for $n \geqq 1$, the function $V_n^* := TV_{n-1}^* \in \mathcal{R}$ is the value function for the n-stage horizon problem, and the Markovian pair of policies $\{f_0^*, f_1^*, ..., f_{n-1}^*\}$, $\{g_0^*, g_1^*, ..., g_{n-1}^*\}$, where the functions f_k^* and g_k^* are the corresponding maxi-minimizing functions, for $k = 0, ..., n-1$, form an equilibrium.

(c) The infinite-horizon game has a value, and $|V^* - V_n^*| \leqq \sum_{t=n}^{\infty} \rho^t L^t r_0 = \rho^n L^n R \to 0$ (pointwise) as $n \to \infty$.

(d) V^* is the unique function in $\mathcal{R}$ satisfying the optimality equation $TV^* = V^*$. Moreover, there exists a pair of stationary policies (f^*, g^*) which is an equilibrium pair for the infinite-horizon game.

*(e) In addition, if **Assumption 4** holds, T is a contractive operator on $\mathcal{M}_\mu(\mathcal{S})$ of modulus ρ and $||V^* - V_n^*||_\mu \leqq \frac{m\rho^n}{1-\rho}$.*

Proof. (a) By definition of r_0, for all $(s, a, b) \in \mathbb{K}$, $|r(s, a, b)| \leqq r_0(s)$, and taking $v = r_0$ in Lemma 3.3 (a), for all $t \in \mathbb{N}$,

$$\mathbb{E}_s^{\pi,\gamma}|r(S_t, A_t, B_t)| \leqq \mathbb{E}_s^{\pi,\gamma}|r_0(S_t)| \leqq (L^t r_0)(s),$$

which implies

$$|V^{\pi,\gamma}(s)| \leqq \sum_{t=0}^{\infty} \rho^t (\mathbb{E}_s^{\pi,\gamma}|r(S_t, A_t, B_t)|) \leqq \sum_{t=0}^{\infty} \rho^t (L^t r_0)(s) = R(s)$$

and therefore

$$|\overline{V}(s)| \leqq R(s), \quad |\underline{V}(s)| \leqq R(s).$$

(b) The main ideas of the proof can be sketched out as follows. By backward induction and well-known arguments, it is possible to prove that starting from $V_0^* \equiv 0 \in \mathcal{R}$, after n successive applications of the operator T, we obtain the value function of the game as well as the equilibrium strategies. In this recursive application it is crucial taking in mind that **Assumption 3** implies that T maps $\mathcal{R}$ into itself (Lemma 3.1). In this way, for each $k \leqq n$, V_k^* is in the domain of T, $\mathcal{R}$.

(c) For any $s \in \mathcal{S}$, $\pi \in \Pi$ and $\gamma \in \Gamma$, consider formulas (4) and (5) for $V^{\pi,\gamma}(s)$ and $V_n^{\pi,\gamma}(s)$. Then, using Lemma 3.3 (a) again,

$$|V^{\pi,\gamma}(s) - V_n^{\pi,\gamma}(s)| \leqq \sum_{t=n}^{\infty} \rho^t \mathbb{E}_s^{\pi,\gamma}|r(S_t, A_t, B_t)| \leqq \sum_{t=n}^{\infty} \rho^t (L^t r_0)(s). \tag{15}$$

In addition, according to Lemma A.2 in the Appendix,

$$\begin{aligned} |\overline{V}^*(s) - \overline{V}_n^*(s)| &= \left| \inf_\gamma \sup_\pi V^{\pi,\gamma}(s) - \inf_\gamma \sup_\pi V_n^{\pi,\gamma}(s) \right| \\ &\leqq \sup_\pi \sup_\gamma |V^{\pi,\gamma}(s) - V_n^{\pi,\gamma}(s)| \ . \end{aligned}$$

Then using (15),

$$|\overline{V}^*(s) - \overline{V}^*_n(s)| \ \leqq \ \sum_{t=n}^{\infty} \rho^t (L^t r_0)(s).$$

Likewise,

$$|\underline{V}^*(s) - \underline{V}^*_n(s)| \leqq \sup_{\pi} \sup_{\gamma} |V^{\pi,\gamma}(s) - V^{\pi,\gamma}_n(s)| \leqq \sum_{t=n}^{\infty} \rho^t (L^t r_0)(s).$$

From part (b) we know that each finite-stage game has a value (and then $\underline{V}^*_n(s) = \overline{V}^*_n(s) = V^*_n(s)$), and taking into account the fact that **Assumption 3** implies $\sum_{t=n}^{\infty} \rho^t (L^t r_0)(s) \to 0$ as $n \to \infty$, we have that the infinite-horizon game has a value, which verifies

$$V^*(s) = \lim_{n\to\infty} V^*_n(s),$$

with the convergence bound

$$|V^*(s) - V^*_n(s)| \ \leqq \ \sum_{t=n}^{\infty} \rho^t (L^t r_0)(s) \ = \ \rho^n (L^n R)(s) \ . \tag{16}$$

(d) First, we will prove uniqueness. That is, that there is at most one $V \in \mathcal{R}$ such that $TV = V$. For this, observe that Lemma A.2 implies the following inequality for any pair of functions u and v in $\mathcal{R}$, and for any $s \in \mathcal{S}$,

$$|(Tu - Tv)(s)| \leqq \rho \sup_{a\in\mathcal{A}_s, b\in\mathcal{B}_s} \int_{\mathcal{S}} |u(z) - v(z)| Q(\mathrm{d}z|s,a,b) = \rho (L|u-v|)(s) \ .$$

In general, for $n \in \mathbb{N}$,

$$|(T^n u - T^n v)(s)| \leqq \rho^n (L^n |u - v|)(s) \leqq 2\rho^n (L^n R)(s) \to 0,$$

as in Lemma 3.3 (b). This implies the uniqueness of fixed points in $\mathcal{R}$.

The identity $TV^* = V^*$ follows by parts (b) and (c) and the uniqueness already shown. Indeed, as $n \to \infty$,

$$T^n 0 = V^*_n \to V^* \ ,$$

and it results that $TV^* = V^*$.

Morover, since by parts (a) and (c), $V^* = \overline{V}^* \in \mathcal{R}$, applying Lemma 3.2, there exists a pair $(f^*, g^*) \in \Pi_{\text{stat}} \times \Gamma_{\text{stat}}$ of stationary strategies such that $TV^* = T^{f^*,g^*} V^*$. In consequence V^* is a fixed point of T^{f^*,g^*}, resulting, by Lemma 3.4, in

$$V^* = V^{f^*,g^*} \ .$$

(e) The contractivity is a direct consequence of (4).

For the second part, observe that from (16), $|V^*(s) - V^*_n(s)| \leqq \sum_{t=n}^{\infty} \rho^t (L^t r_0)(s)$. By using Remark 2.4,

$$\sum_{t=n}^{\infty} \rho^t \times (L^t r_0) \leqq \sum_{t=n}^{\infty} \rho^t \times m\mu = \frac{m\rho^n}{1-\rho} \ \mu \ ,$$

which gives

$$||V^* - V^*_n||_\mu \leqq \frac{m\rho^n}{1-\rho} \ .$$

□

Observe that the contractivity of T (claim (e)) is proved in [11, Lemma 6.1] under a stronger assumption than our **Assumption 1** (f).

Lemma 3.4. *Suppose that* **Assumptions 1, 2** *and* **3** *hold. Then, given a pair of stationary policies* $f \in \Pi_{\mathrm{stat}}$ *and* $g \in \Gamma_{\mathrm{stat}}$*, the value* $V^{f,g}$ *is the unique fixed point of* $T^{f,g}$*.*

Proof. The proof of the fixed point equation follows from the next calculations, for $s \in \mathcal{S}$,

$$\begin{aligned}
V^{f,g}(s) &= \mathbb{E}_s^{f,g}\left[\sum_{t=0}^{\infty}\left(\prod_{k=0}^{t-1}\beta(S_k,A_k,B_k)\right) r(S_t,A_t,B_t)\right] \\
&= \mathbb{E}_s^{f,g}\left[r(s,f,g) + \sum_{t=1}^{\infty}\left(\prod_{k=0}^{t-1}\beta(S_k,A_k,B_k)\right) r(S_t,A_t,B_t)\right] \\
&= r(s,f,g) + \beta(s,f,g)\, \mathbb{E}_s^{f,g}\left[\sum_{t=1}^{\infty}\left(\prod_{k=1}^{t-1}\beta(S_k,A_k,B_k)\right) r(S_t,A_t,B_t)\right] \\
&= r(s,f,g) + \beta(s,f,g)\int_{\mathcal{S}} V^{f,g}(z) Q(\mathrm{d}z|s,f,g) \;=\; (T^{f,g}V^{f,g})(s)\ .
\end{aligned}$$

The uniqueness follows similarly to the first part in the proof of Theorem 3.1 part (d). □

4 Approximation Procedures

4.1 Rolling Horizon Procedure

For a wide class of stochastic control or game problems, obtaining an optimal policy explicitly is a difficult task. This is why practitioners often use instead a heuristic method called the Rolling Horizon procedure (also, Receding Horizon, or Model Predictive Control), which works as follows. To the infinite-horizon game is associated a finite-stage horizon[1] game: for a given integer N (the stage-horizon length) and a state s, find:

$$(FHP) \qquad \sup_{\pi}\inf_{\gamma} \mathbb{E}_s^{\pi,\gamma}\left[\sum_{t=0}^{N-1}\left(\prod_{k=0}^{t-1}\beta(S_k,A_k,B_k)\right) r(S_t,A_t,B_t)\right].$$

Solving this problem for each initial state results in a sequence of pairs of Markovian policies

$$\pi_N^* = (f_N, f_{N-1}, \ldots, f_2, f_1),\ \gamma_N^* = (g_N, g_{N-1}, \ldots, g_2, g_1), \tag{17}$$

where $f_1(s_{N-1})$ is the best action to be applied at stage $t = N-1$ by the first player, and $g_1(s_{N-1})$ for the second one, when only one stage remains to reach the horizon, f_2 and g_2 are the best decision rules to be applied for the players when two stages remain to reach the horizon, at time $t = N-2$, and so on. In particular, $f_N(s_0)$ and $g_N(s_0)$ are the best decision rules to be applied to the initial state s_0.

The **RH** method prescribes to repeatedly solve such a finite-stage horizon problem, taking the current state as initial state. Then, only the first decision will be applied.

Specifically, the procedure to construct a **RH** policy is the following one. Fix some integer N and consider a denumerable set of epochs.

[1]For continuous-time models, it is important to distinguish a time horizon and an horizon measured in a number of control/game stages.

RH1 At iteration t, and for the current state s_t, solve the N-stage game (FHP), taking s_t as initial state. A pair of actions $f_N(s_t)$, $g_N(s_t)$ is obtained.

RH2 Apply $a_t = f_N(s_t)$, $b_t = g_N(s_t)$.

RH3 Observe the achieved state at time $t+1$: s_{t+1}.

RH4 Set $t := t+1$ and go to step 1.

The **RH** procedure does not specify how to compute the values $f_N(s_t)$ and $g_N(s_t)$. Its efficiency is based on the idea that computing $f_N(s_t)$ and $g_N(s_t)$ alone is usually much easier than computing the N decision rules in (17). On the other hand, the performance of the resulting sequence of decisions is not the optimal one, although the intuition is that when N is "large enough", the performance should be close to the optimal. The practical issue is then to choose N so as to obtain a proper compromise between precision and the computational effort needed to obtain $f_N(s_t)$ and $g_N(s_t)$. We address this issue through two formal qualitative and quantitative questions. Let $U_N(s)$ be the expected gain achieved by player 1 with the **RH** procedure with horizon length N, starting in state s:

Q1 Under which conditions on the problem is it true that $\lim_{N\to\infty} U_N(s) = V^*(s)$?

Q2 Given a state s and $\epsilon > 0$, is it possible to compute N such that $|U_N(s) - V^*(s)| < \epsilon$?

In what follows we prove the convergence of the procedure to the value of the original game, thereby answering question **Q1**. We obtain convergence bounds that can be used to answer **Q2**. The term "convergence" has to be understood in the sense that when the horizon N goes to infinity, the value obtained with the procedure approaches the value of the game. The preliminary observation, classical for studying Rolling Horizon, is that the procedure **RH** effectively implements, for both players, a stationary Markov policy. Since player 1 will repeatedly play according to the state-feedback function f_N, and player 2 will play g_N, we have $U_N = V^{f_N,g_N}$.

We begin with a technical lemma of general interest.

Lemma 4.1. *Suppose that* **Assumption 3** *holds. Let $(f,g) \in \Pi_{\text{stat}} \times \Gamma_{\text{stat}}$ be a pair of stationary strategies, with total discounted reward $V^{f,g}$, and let $v \in \mathcal{R}$ and $w \geqq 0$ be such that*

$$v \ \leqq\ T^{f,g} v + w \ .$$

Then, for any $n \in \mathbb{N}$,

$$(T^{f,g})^n v \ \leqq\ V^{f,g} + \sum_{t=n}^{\infty} \rho^t \times L^t w \ .$$

Similarly, if

$$v \ \geqq\ T^{f,g} v - w,$$

then

$$(T^{f,g})^n v \ \geqq\ V^{f,g} - \sum_{t=n}^{\infty} \rho^t \times L^t w \ .$$

Combining both inequalities: if $|v - T^{f,g} v| \leqq w$, then for any $n \in \mathbb{N}$,

$$\left|(T^{f,g})^n v \ - \ V^{f,g}\right| \ \leqq\ \sum_{t=n}^{\infty} \rho^t \times L^t w \ . \tag{18}$$

Proof. First of all, we have from (10) and any function $u \in \mathcal{M}(\mathcal{S})$,

$$\begin{aligned}(T^{f,g}(v+u))(s) &= r(s,f,g) \ + \ \beta(s,f,g)\int_{\mathcal{S}} v(z)Q(\mathrm{d}z|s,f,g) \ + \ \beta(s,f,g)\int_{\mathcal{S}} u(z)Q(\mathrm{d}z|s,f,g) \\ &= (T^{f,g}v)(s) \ + \ \beta(s,f,g)\int_{\mathcal{S}} u(z)Q(\mathrm{d}z|s,f,g) \\ &\leqq (T^{f,g}v)(s) \ + \ \beta(s,f,g)\times(L|u|)(s) \ .\end{aligned}$$

Also, in particular if $w \geqq 0$ (which implies $Lw \geqq 0$), we have:

$$T^{f,g}(v+w) \ \leqq \ T^{f,g}v \ + \ \rho \times Lw \ .$$

Then, we have successively:

$$\begin{aligned} v &\leqq T^{f,g}v + w \\ T^{f,g}v &\leqq (T^{f,g})^2 v + \rho \times Lw \end{aligned}$$

and by recurrence, it is clear that for all t:

$$(T^{f,g})^t v \ \leqq \ (T^{f,g})^{t+1}v + \rho^t \times L^t w \ .$$

Summing up these inequalities for t from n to m, we obtain:

$$(T^{f,g})^n v \quad \leqq \quad (T^{f,g})^{m+1}v + \sum_{t=n}^{m} \rho^t \times L^t w \ . \tag{19}$$

As $m \to \infty$, $(T^{f,g})^m v \to V^{f,g}$. To see that let us analyze the r.h.s. of the expression (easily derived by recurrence):

$$(T^{f,g})^m v \quad = \quad \mathbb{E}^{f,g}\left[\sum_{t=0}^{m-1}\left(\prod_{k=0}^{t-1}\beta(S_k,f,g)\right) r(S_t,f,g)\right] + \ \mathbb{E}^{f,g}\left[\left(\prod_{t=0}^{m-1}\beta(S_t,f,g)\right) v(S_m)\right] \ .$$

The first term converges to the value $V^{f,g}$ by definition. For the second one, note that

$$\mathbb{E}_s^{f,g}\left[\prod_{t=0}^{m-1}\beta(S_t,f,g)v(S_m)\right] \quad = \quad \mathbb{E}_s^{f,g}\left[\prod_{t=0}^{m-1}\beta(S_t,A_t,B_t)v(S_m)\right] \ \to \ 0$$

according to Lemma 3.3 (b).

The second term in (19) is a series of positive terms which either diverges, or converges to $\sum_{t=n}^{\infty}\rho^t \times L^t w < \infty$. In both cases, the bound follows.

The opposite inequality in the statement of the lemma follows similarly. □

Theorem 4.1. *Suppose that* **Assumptions 1, 2** *and* **3** *hold. Then,*

$$|V^* - U_N| \ \leqq \ 2\rho^N \times L^N R \ .$$

Proof. By definition of an optimal $N-1$-stage strategy and the definition (5), for all $s \in \mathcal{S}$ and $N \geqq 1$,

$$V^*_{N-1}(s) = \sup_\pi \inf_\gamma \mathbb{E}_s^{\pi,\gamma}\left[\sum_{t=0}^{N-2}\left(\prod_{k=0}^{t-1}\beta(S_k,A_k,B_k)\right) r(S_t,A_t,B_t)\right] \ .$$

If we add and subtract the term

$$\Delta := \left(\prod_{k=0}^{N-2} \beta(S_k, A_k, B_k)\right) r(S_{N-1}, A_{N-1}, B_{N-1}) \ ,$$

we obtain

$$V^*_{N-1}(s) \quad = \quad \sup_\pi \inf_\gamma \ \mathbb{E}^{\pi,\gamma}_s \left[\sum_{t=0}^{N-1} \left(\prod_{k=0}^{t-1} \beta(S_k, A_k, B_k)\right) r(S_t, A_t, B_t) - \Delta \right] \ = \ \sup_\pi \inf_\gamma \ \mathbb{E}^{\pi,\gamma}_s [h - \Delta] \ .$$

Applying Lemma A.2 with $f = \mathbb{E}^{\pi,\gamma}_s[h - \Delta]$ and $g = \mathbb{E}^{\pi,\gamma}_s[h]$, we have:

$$\left| \sup_\pi \inf_\gamma \mathbb{E}^{\pi,\gamma}_s [h - \Delta] - \sup_\pi \inf_\gamma \mathbb{E}^{\pi,\gamma}_s [h] \right| \ \leqq \ \sup_\pi \sup_\gamma \mathbb{E}^{\pi,\gamma}_s |\Delta| \ .$$

Then, recognizing that $V^*_N(s)$ is $\sup_\pi \inf_\gamma \mathbb{E}^{\pi,\gamma}_s[h]$, we obtain:

$$|V^*_N(s) - V^*_{N-1}(s)| \leqq \sup_\pi \sup_\gamma \ \mathbb{E}^{\pi,\gamma}_s \left| \left(\prod_{k=0}^{N-2} \beta(S_k, A_k, B_k)\right) r(S_{N-1}, A_{N-1}, B_{N-1}) \right| \ .$$

Since for all $(s, a, b) \in \mathbb{K}$, $r(s, a, b) \leqq r_0(s)$ by **Assumption 3**, and since $\beta(s, a, b) \leqq \rho$, we further obtain:

$$|V^*_N(s) - V^*_{N-1}(s)| \leqq \sup_\pi \sup_\gamma \ \rho^{N-1} \ \mathbb{E}^{\pi,\gamma}_s [r_0(S_{N-1})] \ .$$

With Lemma 3.3 (a), $\mathbb{E}^{\pi,\gamma}_s[r_0(S_{N-1})] \leqq (L^{N-1} r_0)(s)$, and we conclude:

$$|V^*_N - V^*_{N-1}| \leqq \rho^{N-1} \times L^{N-1} r_0. \tag{20}$$

Let us work with the inequality

$$V^*_{N-1} \leqq V^*_N + \rho^{N-1} \times L^{N-1} r_0.$$

Considering Theorem 3.1, part (b), $V^*_N = T^{f_N, g_N} V^*_{N-1}$, that is:

$$\begin{aligned} V^*_N(s) \quad &= \quad r(s, f_N, g_N) + \beta(s, f_N, g_N) \times \int_S V^*_{N-1}(z) Q(\mathrm{d}z | s, f_N, g_N) \\ &\leqq \quad r(s, f_N, g_N) + \beta(s, f_N, g_N) \times \int_S V^*_N(z) Q(\mathrm{d}z | s, f_N, g_N) \ + \ \rho^N \times L^N r_0(s) \ , \end{aligned}$$

or in other words:

$$V^*_N \ \leqq \ T^{f_N, g_N} V^*_N \ + \ \rho^N \times L^N r_0 \ . \tag{21}$$

Using (18) in Lemma 4.1 with $f \equiv f_N$, $g \equiv g_N$, $n = 0$ and $w = \rho^N L^N r_0$ (which is effectively positive), we obtain since $V^{f_N, g_N} = U_N$:

$$|V^*_N - U_N| \ \leqq \ \sum_{t=N}^{\infty} \rho^t \times L^t r_0 \ . \tag{22}$$

Finally, using Theorem 3.1 (c),

$$\begin{aligned} |V^* - U_N| \quad &\leqq \quad |V^* - V^*_N| + |V^*_N - U_N| \\ &\leqq \quad \rho^N \times L^N R + \rho^N \times L^N R \ = \ 2\rho^N \times L^N R \ . \end{aligned}$$

□

Remark 4.1. Theorem 4.1 generalizes to **SMG** the results in [5, Theorem 4.2] for discrete-time **MDP**. ■

In order to improve the bounds obtained, we consider the following Positive Value Assumption

Assumption 6. $V_1^* \geqq 0$.

Observe that this is the case, for instance, when the reward function r is positive.

Proposition 4.1. *Suppose that* **Assumption 6** *holds. Then for* $n \in \mathbb{N}$ *and for all* $s \in \mathcal{S}$, $V_n^*(s) \leqq V_{n+1}^*(s)$.

Proof. Since $V_0^* \equiv 0$, we can write **Assumption 6** as $V_0^*(s) \leqq V_1^*(s)$, for all $s \in \mathcal{S}$. If this is the case, for any pair of strategies $(f, g) \in \Pi_{\text{stat}} \times \Gamma_{\text{stat}}$, there holds $T^{f,g}V_0^* \leqq T^{f,g}V_1^*$, and by Lemma A.1, $TV_0^* \leqq TV_1^*$. The last inequality rewrites as $V_1^* \leqq V_2^*$. The result follows now by induction with similar arguments. □

Corollary 4.1. *If in Theorem 4.1* **Assumption 6** *holds, then,*

$$|V^* - U_N| \leqq \rho^N \times L^N R.$$

Proof. If **Assumption 6** holds, it is possible to improve the bound (21) into $V_N^* \leqq T^{f_N,g_N}V_N^*$, and by application of Lemma 4.1, we get $V_N^* \leqq U_N$. In other terms, Inequality (22) now reads

$$0 \ \leqq\ U_N - V_N^* \ \leqq\ \rho^N \times L^N R\ ,$$

or equivalently,

$$-\rho^N \times L^N R \ \leqq\ V_N^* - U_N \ \leqq\ 0\ .$$

Moreover, since the sequence $\{V_n^*\}_n$ is monotone by Proposition 4.1, we have $V_N^* \leqq V^*$. Accordingly, (16) can be refined into:

$$0 \ \leqq\ V^* - V_N^* \ \leqq\ \rho^N \times L^N R\ .$$

Combining the last inequalities, we get

$$-\rho^N \times L^N R \ \leqq\ V^* - U_N \ \leqq\ \rho^N \times L^N R\ ,$$

which is equivalent to the bound we look for. □

Corollary 4.2. *Suppose that* **Assumptions 1, 2** *and* **4** *hold. Then,*

$$|V^* - U_N| \ \leqq\ \frac{2m\rho^N}{1-\rho}\mu,$$

and therefore

$$||V^* - U_N||_\mu \ \leqq\ \frac{2m\rho^N}{1-\rho}\ .$$

If in addition **Assumption 6** *holds, then*

$$||V^* - U_N||_\mu \ \leqq\ \frac{m\rho^N}{1-\rho}.$$

Proof. If **Assumption 4** holds, then (see Remark 2.4), $||L^t r_0||_\mu \leqq m$ for all t, so that $\rho^N L^N R = \sum_{t=N}^{\infty} \rho^t L^t r_0 \leq \sum_{t=N}^{\infty} m \rho^t \mu = \frac{m\rho^N}{1-\rho}\mu$. The bounds follow from Theorem 4.1.

If **Assumption 6** holds, the bound follows from Corollary 4.1 using the same argument. □

When, in particular, we require **Assumption 5** instead of **Assumption 4**, bounds in Corollary 4.2 take the following form:

Corollary 4.3. *Suppose that* **Assumptions 1, 2** *and* **Assumption 5** *hold, then for all* $s \in \mathcal{S}$

$$|V^*(s) - U_N(s)| \leqq \frac{2M\rho^N}{1-\rho},$$

and therefore,

$$||V^* - U_N||_\infty \leqq \frac{2M\rho^N}{1-\rho}.$$

If in addition **Assumption 6** *holds,*

$$||V^* - U_N||_\infty \leqq \frac{M\rho^N}{1-\rho}.$$

4.2 Approximate Rolling Horizon Procedure

Suppose now that the players do not have exact information about the problem to be solved at **RH1** in the **RH** procedure, but suppose they know or are able to compute an approximation of that value. We are interested in implementing a procedure where this last approximation is used instead of the value function of the game with finite horizon. We want to estimate the error introduced.

In the first scheme we study, the source of approximation lies in the determination of the finite-stage equilibrium V^*_{N-1}. Specifically,

ARH1 Choose some function V close in some sense to V^*_{N-1}, where V^*_{N-1} is the $N-1$-stage value function.

ARH2 At iteration t, and for the current state s_t, solve

$$\max_{a\in\mathcal{A}_{s_t}} \min_{b\in\mathcal{B}_{s_t}} \left\{ r(s_t,a,b) + \beta(s_t,a,b) \int_{\mathcal{S}} V(z) Q(\mathrm{d}z|s_t,a,b) \right\}.$$

A pair of actions $\tilde{f}_N(s_t)$, $\tilde{g}_N(s_t)$ is obtained.

ARH3 Apply $a_t = \tilde{f}_N(s_t)$, $b_t = \tilde{g}_N(s_t)$.

ARH4 Observe the achieved state at time $t+1$: s_{t+1}.

ARH5 Set $t := t+1$ and go to step 2.

We will note with $\tilde{U}_N$ the total discounted reward of the pair of stationary policies $\tilde{f}_N$ and $\tilde{g}_N$. The next result gives answers to questions **Q1** and **Q2** stated in this section for the sequence of successive rewards $\tilde{U}_N$.

A priori, the space $\mathcal{R}$ is not equipped with a metric. This makes it difficult to give a precise meaning to “close in some sense to V^*_{N-1}” in step **ARH1**, for functions in $\mathcal{R}$. On the other hand, in the space $\mathcal{M}_\mu(\mathcal{S})$ the μ-norm induces a metric under **Assumption 4**. Adopting this (stronger) assumption instead of **Assumption 3**, we can prove the following results.

Theorem 4.2. *Suppose that* **Assumptions 1, 2** *and* **4** *hold, and that, for some* N, $||V_N^* - V_{N-1}^*||_\mu \leqq \varepsilon_1$. *Let* $V \in \mathcal{M}_\mu(\mathcal{S})$ *be a function such that* $||V_{N-1}^* - V||_\mu \leqq \varepsilon_2$, *consider a pair of stationary strategies* $(f,g) \in \Pi_{\text{stat}} \times \Gamma_{\text{stat}}$ *such that* $T^{f,g}V = TV$, *and let* $\tilde{U}_N = V^{f,g}$. *Then,*

$$||V^* - \tilde{U}_N||_\mu \leqq \frac{2\rho(\varepsilon_1 + \varepsilon_2)}{1-\rho} \ .$$

Proof. Let us start with the triangular inequality

$$||V^* - \tilde{U}_N||_\mu \leqq ||V^* - V_N^*||_\mu + ||V_N^* - \tilde{U}_N||_\mu \ .$$

To bound the first term in the r.h.s. of this inequality, observe that, for any $m \in \mathbb{N}$,

$$\begin{aligned}
||V_{N+m}^* - V_N^*||_\mu &\leqq \sum_{k=0}^{m-1} ||V_{N+k+1}^* - V_{N+k}^*||_\mu \\
&= \sum_{k=0}^{m-1} ||T^{k+1}V_N^* - T^{k+1}V_{N-1}^*||_\mu \\
&\leqq \sum_{k=0}^{m-1} \rho^{k+1}||V_N^* - V_{N-1}^*||_\mu \\
&= \frac{\rho\,(1-\rho^m)}{1-\rho}||V_N^* - V_{N-1}^*||_\mu \ ,
\end{aligned}$$

and, taking limits when $m \to \infty$,

$$||V^* - V_N^*||_\mu \leqq \frac{\rho\ \varepsilon_1}{1-\rho} \ , \tag{23}$$

using the bound $||V_N^* - V_{N-1}^*||_\mu \leqq \varepsilon_1$ in the hypothesis.

To bound the second term, we use the facts that $T^{f,g}V = TV$, $T^{f,g}\tilde{U}_N = \tilde{U}_N$, and that T and $T^{f,g}$ are contractions of modulus ρ (Lemma 3.1). With that,

$$\begin{aligned}
||V_N^* - \tilde{U}_N||_\mu &= ||V_N^* - T^{f,g}\tilde{U}_N||_\mu \\
&\leqq ||V_N^* - T^{f,g}V||_\mu + ||T^{f,g}V - T^{f,g}\tilde{U}_N||_\mu \\
&\leqq ||TV_{N-1}^* - TV||_\mu + \rho||V - \tilde{U}_N||_\rho \\
&\leqq \rho\varepsilon_2 + \ \rho\ (||V - V_{N-1}^*||_\mu + ||V_{N-1}^* - V_N^*||_\mu + ||V_N^* - \tilde{U}_N||_\mu) \\
&\leqq \rho\varepsilon_2 + \rho(\varepsilon_2 + \varepsilon_1) + \rho\ ||V_N^* - \tilde{U}_N||_\infty \ ,
\end{aligned}$$

which implies

$$(1-\rho)\ ||V_N^* - \tilde{U}_N||_\mu \leqq 2\rho\varepsilon_2 + \rho\varepsilon_1 \ ,$$

or

$$||V_N^* - \tilde{U}_N||_\mu \leqq \frac{2\rho\varepsilon_2 + \rho\varepsilon_1}{1-\rho} \ . \tag{24}$$

Finally, by (23) and (24),

$$||V^* - \tilde{U}_N||_\mu \leqq \frac{2\rho(\varepsilon_1 + \varepsilon_2)}{1-\rho} \ .$$

□

Corollary 4.4. *Suppose that* **Assumptions 1, 2** *and* **4** *hold. Let* $V \in \mathcal{M}_\mu(\mathcal{S})$ *be a function such that for some* $N \geqq 1$, $||V^*_{N-1} - V||_\mu \leqq \varepsilon$, *consider a pair of policies* $(\tilde{f}_N, \tilde{g}_N) \in \Pi_{\rm stat} \times \Gamma_{\rm stat}$ *such that* $T^{\tilde{f}_N,\tilde{g}_N} V = TV$, *and let* $\tilde{U}_N = V^{\tilde{f}_N,\tilde{g}_N}$. *Then,*

$$||V^* - \tilde{U}_N||_\mu \leqq \frac{2m\rho^N}{1-\rho} + \frac{2\rho\varepsilon}{1-\rho} \ .$$

If in addition **Assumption 5** *holds,*

$$||V^* - \tilde{U}_N||_\infty \leqq \frac{2M\rho^N}{1-\rho} + \frac{2\rho\varepsilon}{1-\rho} \ .$$

Proof. The proof follows by application of Theorem 4.2, with $\varepsilon_2 := \varepsilon$, and considering that, as a consequence of Inequality (20),

$$||V^*_N - V^*_{N-1}||_\mu \ \leqq \ \frac{m\rho^{N-1}}{1-\rho} \ =: \varepsilon_1 \ .$$

The second statement follows as in Corollary 4.2. □

So far in this work we have considered the situation where, given an approximate value function V, the maximizer plays the policy $\tilde{f} \in \Pi_{\rm stat}$ such that $T^{\tilde{f},\tilde{g}} V = TV$. Suppose now that this player chooses any $\tilde{f} \in \Pi_{\rm stat}$ such that

$$(T^f V)(s) = (TV)(s) \ ,$$

where

$$(T^f V)(s) := \inf_{b \in \mathcal{B}_s} \left\{ r(s,f,b) + \beta(s,f,b) \int_{\mathcal{S}} V(z) Q(\mathrm{d}z|s,f,b) \right\} \ .$$

This situation corresponds to the *worst-case scenario* for player 1. The next result gives us bounds for this second **ARH** framework. In this worst-case approach we keep the notation $\tilde{U}_N$ to denote the value function when the corresponding strategies are used.

The proof of the next result is ommited, due to its similarity with the one of Theorem 4.2.

Theorem 4.3. *Suppose that* **Assumptions 1, 2** *and* **4** *hold, and that, for some* N, $||V^*_N - V^*_{N-1}||_\mu \leqq \varepsilon_1$. *Let* $V \in \mathcal{M}_\mu(\mathcal{S})$ *be a function such that* $||V^*_{N-1} - V||_\mu \leqq \varepsilon_2$, *and consider a policy* $\tilde{f}_N \in \Pi_{\rm stat}$ *such that* $T^{\tilde{f}_N} V = TV$. *Then,*

$$||V^* - \tilde{U}_N||_\mu \leqq \frac{2\rho(\varepsilon_1 + \varepsilon_2)}{1-\rho} \ .$$

where $\tilde{U}_N = V^{\tilde{f}_N, g^*}$ *is the value of the pair* $(\tilde{f}_N, g^*)$, *and* $g^* \in \Gamma_{\rm stat}$ *is an infinite-horizon equilibrium strategy for player 2.*

Remark 4.2. Again, if **Assumptions 1, 2** and **4** hold, with $V \in \mathcal{M}_\mu(\mathcal{S})$ an in Theorem 4.3, and $\tilde{f}_N \in \Pi_{\rm stat}$ such that $T^{\tilde{f}_N} V = TV$, in the worst case scenario,

$$||V^* - \tilde{U}_N||_\mu \leqq \frac{2m\rho^N}{1-\rho} + \frac{2\rho\varepsilon}{1-\rho} \ ,$$

and if also **Assumption 5** holds,

$$||V^* - \tilde{U}_n||_\infty \leqq \frac{2M\rho^N}{1-\rho} + \frac{2\rho\varepsilon}{1-\rho} \ .$$

■

Remark 4.3. Theorems 4.2 and 4.3 generalize to **SMG** our results for **SMDP** presented in [3] and improve results for finite state discrete-time **MG** described in [2, Theorems 6, 7]. The use of finer bounds even allows us to improve their "$O((1-\rho)^{-2})$" term into a "$O((1-\rho)^{-1})$" one. ■

5 Concluding Remarks

Through this work we have dealt with semi-Markov games models with discounted payoff, under different assumptions on the reward function. We have generalized known properties of the equilibria of games for both the finite-horizon and the infinite horizon case.

In addition, we have studied the performance of the rolling horizon procedure and of approximate rolling horizon procedures. We have proved the convergence of the values related to the rolling horizon procedure to the optimal reward function. We obtain simple pointwise convergence if **Assumption 3** is verified and pointwise geometrical convergence when **Assumption 4** holds. As a particular case, we have obtained uniform geometrical convergence for the case of uniformly bounded rewards functions, i.e. under **Assumption 5**.

Finally we have discussed an approximate rolling horizon procedure, based on the possibility of the controller not having perfect prediction of the future needed to take the best immediate action, but approximations of it. Here we have completed the analysis studying the case when the maximizer deals with the *worst-case scenario*.

References

[1] Bhattacharya R., Majumdar, M.; "Controlled semi-Markov models - the discounted case". *J. Statistical Planning and Inference*, 21, 365–381, 1989.

[2] Chang H.S., Marcus, S.I.; "Two-person zero-sum games: receding horizon approach". *IEEE Trans. Automat. Control*, 48, 11, 2003, pp. 1951–1961.

[3] Della Vecchia E., Di Marco S., Jean-Marie A.; "Rolling Horizon Procedure on Controlled Semi-Markov Models. The Discounted Case". *Annals of the Simposio Argentino en Investigación Operativa* (CD version), ISSN 1850 - 2865. 2011.

[4] Hernández-Lerma O., Lasserre J.B.; "Error Bounds for Rolling Horizon Policies in Discrete-Time Markov Control Processes". *IEEE Trans. Automat. Control*, 35, 10, 1990, pp. 1118–1124.

[5] Hernández-Lerma O., Lasserre J.B.; "Value iteration and rolling plans for Markov control processes with unbounded rewards". *J. Math. Anal. Appl.*, 177, 1993, pp. 38–55.

[6] Hernández-Lerma O., *Adaptive Markov Control Processes.* Springer-Verlag 1989.

[7] Hernández-Lerma O., Lasserre J.B., *Discrete-Time Markov Control Processes.* Springer-Verlag 1996.

[8] Jaśkiewicz A., Nowak A.; "Approximations of noncooperative semi-Markov games". *J. Optim. Theory Appl.*, **131**(1), pp. 115–134, 2006.

[9] Jaśkiewicz A., Nowak A.; "Stochastic Games with Unbounded Payoffs: Applications to Robust Control in Economics". *Dyn. Games Appl.*, 1, 2011, 253–279.

[10] Kwon W., Han S.; *Receding Horizon Control. Predictive Control for State Models.* Advanced Textbooks in Control and Signal Processing, Springer, 2005.

[11] Luque-Vásquez, F.; "Zero-sum semi-Markov game in Borel spaces with discounted payoff". Research Report, Departamento de Matemáticas, Universidad de Sonora, Mexico. 2002.

[12] Minjárez-Sosa J., Luque-Vásquez F.; "Two Person Zero-Sum Semi-Markov Games with Unknown Holding Times Distribution on One Side: A Discounted Payoff Criterion", *Appl. Math. Optim.* 57, 2008, 289–305, DOI 10.1007/s00245-007-9016-7.

[13] Nowak A.; "On zero-sum stochastic games with general state space, I". *Probab. Math. Statist.* 4, 1984, 13–32.

[14] Puterman L.; *Markov Decision Processes.* Wiley and Sons, 2005.

[15] Ross, S.; *Applied Probability Models with Optimization Applications*, Holden-Day, 1970.

[16] Sethi T., Sorger G.; "A Theory of Rolling Horizon Decision Making". *Ann. Ops. Res.*, 29, 1, 1991, pp. 387–415.

A Appendix

Lemma A.1. *Consider two functions $f, g : X \times Y \to \mathbb{R}$ such that, for all $x \in X$ and $y \in Y$, $f(x,y) \leqq g(x,y)$. Then*

$$\sup_{x \in X} \inf_{y \in Y} f(x,y) \leqq \sup_{x \in X} \inf_{y \in Y} g(x,y) \ .$$

Proof. Consider the functions $F(x) = \inf_{y \in Y} f(x,y)$ and $G(x) = \inf_{y \in Y} g(x,y)$. For any $x \in X$ fixed, and for all $y \in Y$, $F(x) \leqq f(x,y) \leqq g(x,y)$, and then $F(x) \leqq \inf_{y \in Y} g(x,y) = G(x)$. Consequently, $\sup_{x \in X} F(x) \leqq \sup_{x \in X} G(x)$, which is the stated inequality. □

Lemma A.2. *Consider two functions $f, g : X \times Y \to \mathbb{R}$. Then:*

$$\left| \inf_{y \in Y} \sup_{x \in X} f(x,y) - \inf_{y \in Y} \sup_{x \in X} g(x,y) \right| \ \leqq \ \sup_{x \in X} \sup_{y \in Y} |f(x,y) - g(x,y)|$$

and

$$\left| \sup_{x \in X} \inf_{y \in Y} f(x,y) - \sup_{x \in X} \inf_{y \in Y} g(x,y) \right| \ \leqq \ \sup_{x \in X} \sup_{y \in Y} |f(x,y) - g(x,y)| \ .$$

Proof. Without losing generality, let us suppose $\inf_{y \in Y} \sup_{x \in X} f(x,y) - \inf_{y \in Y} \sup_{x \in X} g(x,y) \geqq 0$. If it is not the case, interchange f with g.

Given $\varepsilon > 0$, take $y^* \in Y$ such that

$$\sup_{x \in X} g(x, y^*) - \varepsilon \leqq \inf_{y \in Y} \sup_{x \in X} g(x,y) \ .$$

Then

$$\inf_{y \in Y} \sup_{x \in X} f(x,y) - \inf_{y \in Y} \sup_{x \in X} g(x,y) \leqq \sup_{x \in X} f(x,y^*) - \sup_{x \in X} g(x,y^*) + \varepsilon \ .$$

Now, taking $x^* \in X$ such that

$$\sup_{x \in X} f(x,y^*) \leqq f(x^*,y^*) + \varepsilon,$$

$$\sup_{x\in X} f(x,y^*) - \sup_{x\in X} g(x,y^*) + \varepsilon \leqq f(x^*,y^*) - g(x^*,y^*) + 2\varepsilon \ ,$$

which implies

$$\left|\inf_{y\in Y}\sup_{x\in X} f(x,y) - \inf_{y\in Y}\sup_{x\in X} g(x,y)\right| \ \leqq \ \sup_{x\in X}\sup_{y\in Y} |f(x,y) - g(x,y)| + 2\varepsilon \ .$$

Since ε is any arbitrary positive number, the first inequality is proved.

The second inequality follows in the same manner, taking, for $\varepsilon > 0$, $x^* \in X$ such that

$$\sup_{x\in X}\inf_{y\in Y} f(x,y) \leqq \inf_{y\in Y} f(x^*,y) + \varepsilon$$

and $y^* \in Y$ such that

$$g(x^*,y^*) - \varepsilon \leqq \inf_{y\in Y} g(x^*,y) \ .$$

Indeed, if $\sup_{x\in X}\inf_{y\in Y} f(x,y) - \sup_{x\in X}\inf_{y\in Y} g(x,y) \geqq 0$,

$$\begin{aligned}\sup_{x\in X}\inf_{y\in Y} f(x,y) - \sup_{x\in X}\inf_{y\in Y} g(x,y) \ &\leqq \ \inf_{y\in Y} f(x^*,y) - \inf_{y\in Y} g(x^*,y) \\ &\leqq \ f(x^*,y^*) - g(x^*,y^*) + 2\varepsilon \ ,\end{aligned}$$

and

$$\left|\sup_{x\in X}\inf_{y\in Y} f(x,y) - \sup_{x\in X}\inf_{y\in Y} g(x,y)\right| \ \leqq \ \sup_{x\in X}\sup_{y\in Y} |f(x,y) - g(x,y)| + 2\varepsilon \ .$$

□

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399