Structural approximations in discounted semi-Markov games

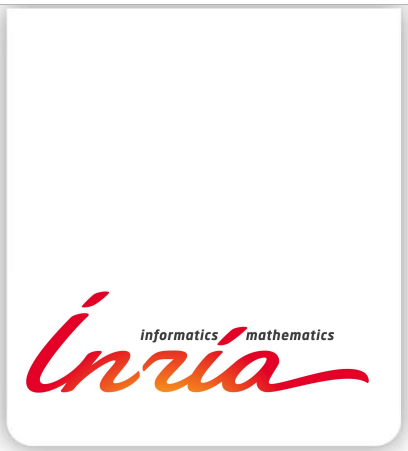
Eugenio Della Vecchia, Silvia C. Di Marco, Alain Jean-Marie

Structural approximations in discounted semi-Markov games

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

RESEARCH REPORT
N° 8162
December 2012
Project-Team Maestro

ISSN 0249-6399 ISRN INRIA/RR--8162--FR+ENG



Structural approximations in discounted semi-Markov games

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

Project-Team Maestro

Research Report n° 8162 — December 2012 — 19 pages

Abstract: We consider the problem of approximating the values and the equilibria in two-person zero-sum discounted semi-Markov games with infinite horizon and compact action spaces, when several uncertainties are present about the parameters of the model. Specifically: on the one hand, we study approximations made on the transition probabilities, the discount factor and the reward functions when the state space is a borelian set. On the other hand, we study approximations on the state space for denumerable ones. Our results are based on those of Tidball and Altman on generic zero-sum games [9]. We provide conditions under which these results can be applied. We also discuss the application of such approximations for finite-horizon games, in relation with the Approximate Rolling Horizon procedure proposed in [3].

Key-words: Game theory, Semi-Markov games, Zero-sum games

* CONICET - UNR, Pellegrini 250, Rosario, Argentina, eugenio@fceia.unr.edu.ar.
† CONICET - UNR, Pellegrini 250, Rosario, Argentina, dimarco@fceia.unr.edu.ar .
‡ INRIA and LIRMM, CNRS/Université Montpellier 2, 161 Rue Ada, 34392 Montpellier, France, ajm@lirmm.fr.

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Approximations structurelles dans les jeux semi-Markoviens actualisés

Résumé : Nous considérons le problème de l'approximation des valeurs et des équilibres d'un jeu semi-Markovien actualisé, en horizon infini avec des ensembles d'actions compacts, en présence d'incertitude sur plusieurs paramètres du modèle. Spécifiquement: d'une part nous étudions les approximations sur les probabilités de transition, sur le facteur d'actualisation et sur les coûts, quand l'espace d'états est un ensemble Borélien. D'autre part, nous étudions les approximations de l'ensemble d'états quand celui-ci est dénombrable. Nos résultats sont basés sur ceux de Tidball et Altman [9]. Nous donnons des conditions sous lesquelles ces résultats peuvent être appliqués. Nous discutons aussi de l'application de telles approximations à des jeux en horizon fini, en relation avec la procédure de l'horizon roulant approchée, proposée dans [3].

Mots-clés : Théorie des jeux, jeux semi-Markoviens, jeux à somme nulle

1 Introduction

In this work we analyze several approximation procedures applied on zero-sum semi-Markov games with the expected total discounted reward as the performance criterion.

Semi-Markov games (**SMG**) generalize Markov games in continuous-time (**MG**) by allowing the decision maker to choose actions whenever the system state changes and allowing the time spent in a particular state to follow an arbitrary probability distribution.

Semi-Markov discounted games are studied, for example, in [5], [6] and [7]. In particular, we study some approximation issues in [3], including a widely applied heuristic method, the Rolling Horizon procedure.

In this paper we analyze approximations of the value function and of the equilibrium policies of the infinite horizon game, when it can be considered the limit game of a sequence of approximating games. The approximating games are designed by considering approximations of the parameters of the original game. In particular we work with approximations on the transition and the holding time probabilities of the models, approximations on the discount factor, approximations on the reward functions, and with approximations on the space of the states, conveniently truncated. In all cases we estimate the errors arising in the approximation. To do that, we use a key result on approximations for zero-sum games, and other ideas presented in [9]. As application, we adapt also these results for the finite horizon game, which gives approximations needed when applying the Approximate Rolling Horizon procedure studied in [3].

In Section 2, we introduce the notation and the model, and we state preliminary results, including the "Key Theorem" of [9] on which we base our analysis. In Section 3, we develop approximation results for infinite-horizon games. In Section 4, we apply these results to finite-horizon games. Section 5 is devoted to the concluding remarks.

2 Preliminaries and notations

We consider a semi-Markov game of the form

$$G := (\mathcal{S}, \mathcal{A}, \mathcal{B}, \{\mathcal{A}_s : s \in \mathcal{S}\}, \{\mathcal{B}_s : s \in \mathcal{S}\}, Q, F, \ell, \alpha) \tag{1}$$

where $\mathcal{S}$ is the state space, and, for each $s \in \mathcal{S}$, $\mathcal{A}_s$ and $\mathcal{B}_s$ denote the sets of actions available in state s for players 1 and 2. $\mathcal{A} = \bigcup_{s\in\mathcal{S}} \mathcal{A}_s$ and $\mathcal{B} = \bigcup_{s\in\mathcal{S}} \mathcal{B}_s$. Let $\mathbb{K} := \{(s, a, b) : s \in \mathcal{S}, a \in A_s, b \in B_s\}$. Moreover, $Q(\cdot|s, a, b)$ is a stochastic kernel on $\mathcal{S}$ given $\mathbb{K}$ called the transition law, and $F(\cdot|s, a, b)$ is a probability distribution on $[0, \infty)$ given $\mathbb{K}$ called the transition time distribution. The real function ℓ on $\mathbb{K}$ represents the reward for player 1 and the cost function for player 2. Finally, α is a discount factor.

The game is played as follows: if s is the state of the game at some decision (or transition) epoch, the players independently choose actions $a \in \mathcal{A}_s$ and $b \in \mathcal{B}_s$. Then several things happen. First, the system moves to a new state according to the probability measure $Q(.|s, a, b)$. Second, the time until the next transition/decision occurs is determined. This time interval is a random variable having the distribution function $F(\cdot|s, a, b)$. This defines a sequence of 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 stage n. During two transition epochs, player 1 receives a constant reward rate $\ell(s, a, b)$, discounted over time at rate α, and player 2 incurs in a cost rate $\ell(s, a, b)$, also discounted. We consider that the reward just depends on the current state and the actions but not on time, i.e. ℓ stationary.

Let $\mathcal{M}(\mathcal{S})$ denote the space of measurable functions on $\mathcal{S}$. For Borel sets X and Y, $\mathbb{P}(X)$ denotes the family of probability measures on X endowed with the weak topology and $\mathbb{P}(X|Y)$ is the family of transition probabilities from Y to X.

The space H_n of admissible histories of the process at the n-th decision epoch, consists of sequences of states, decisions and holding 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 $H_n = (\mathbb{K} \times \mathbb{R}^+)^n \times \mathcal{S}$ 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 Markov strategy (or Markov policy) for player 1 is a sequence $\pi = \{\pi_n\}$ of stochastic kernels $\pi_n \in \mathbb{P}(\mathcal{A}|H_n)$ such that for every $h_n \in H_n$ and $n \in \mathbb{N}$, $\pi_n(\mathcal{A}_{s_n}|h_n) = 1$. We denote by Π the set of all Markov strategies of player 1. A Markov strategy $\pi = \{\pi_n\}$ is called stationary if there exists $f \in \mathbb{P}(\mathcal{A}|\mathcal{S})$ such that $f(s) \in \mathbb{P}(\mathcal{A}_s)$ and $\pi_n = f$ 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.

Similarly, a Markov strategy for player 2 is a sequence $\gamma = \{\gamma_n\}$, where $\gamma_n \in \mathbb{P}(\mathcal{B}|H_n)$, such that for every $h_n \in H_n$ and $n \in \mathbb{N}$, $\gamma_n(\mathcal{B}_{s_n}|h_n) = 1$. In this case we note with Γ the set of all Markov strategies of player 2. A Markov strategy γ is called stationary if there exists $g \in \mathbb{P}(\mathcal{B}|S)$ such that $g(s) \in \mathbb{P}(\mathcal{B}_s)$ and $\gamma_n = g$ for all $s \in \mathcal{S}$ and $n \in \mathbb{N}$. For player 2, we denote Γ_{stat} the set of its stationary strategies.

We note

$$\beta(s, a, b) := \int_0^\infty e^{-\alpha t} F(\mathrm{d}t|s, a, b) \tag{2}$$

and

$$\vartheta(s, a, b) = \frac{1 - \beta(s, a, b)}{\alpha} \ . \tag{3}$$

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 policies f and g.

As mentioned previously, in order to evaluate the performance of policies, we use a total discounted criterion. We assume that rewards are continuously discounted over time 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 by

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

where $T_0 = 0$ and $T_n = T_{n-1} + \delta_n$. The infinite-horizon total expected α-discounted payoff is

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

where we adopt the usual conventions that $\prod_{k=0}^{-1} X_k = 1$ and $\sum_{t=0}^{-1} Y_t = 0$.

At this point, we 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

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

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

We shall make further assumptions under which we work.

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) r is a bounded function on $\mathbb{K}$, i.e. there exist $M > 0$ such that, for all $(s,a,b) \in \mathbb{K}$, $|r(s,a,b)| \leqq M$.

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

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

(f) 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 $\mathcal{A}_s \times \mathcal{B}_s$.

(g) $\int_0^\infty tF(dt|\cdot)$ is continuous on $\mathbb{K}$.

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

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

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

respectively. We know that, in general $\underline{V}^* \leqq \overline{V}^*$. If $\underline{V}^* = \overline{V}^*$, we refer to this common value as the value of the game, and we note it with V^*. Similar values are defined 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:

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

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

A pair of strategies $\tilde{\pi} \in \Pi$ and $\tilde{\gamma} \in \Gamma$ is said to be an *ε-equilibrium* (or an *almost equilibrium*) pair if it holds

$$\inf_{\gamma\in\Gamma} V^{\tilde{\pi},\gamma} \geqq \inf_{\gamma\in\Gamma} V^{\pi,\gamma} - \varepsilon, \quad \text{for all } \pi \in \Pi \ ,$$

and

$$\sup_{\pi\in\Pi} V^{\pi,\tilde{\gamma}} \leqq \sup_{\pi\in\Pi} V^{\pi,\gamma} + \varepsilon, \quad \text{for all } \gamma \in \Gamma \ .$$

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

$$(Tv)(s) := \sup_{a\in\mathcal{A}_s} \inf_{b\in\mathcal{B}_s} \left\{ r(s,a,b) + \beta(s,a,b) \int_{\mathcal{S}} v(z) Q(\mathrm{d}z|s,a,b) \right\} \ , \tag{6}$$

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

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

Tidball and Altman study in [9] under which conditions a sequence of games converges to a given game.

Consider a sequence G_n, $n = 1, 2, ...$ of generic zero-sum games and a game G. We will note, for a pair of policies $\pi \in \Pi$ and $\gamma \in \Gamma$, $U_n^{\pi,\gamma}$ the reward produced for the pair in the game G_n and $U^{\pi,\gamma}$ in the game G.

Let also the upper and lower values of the game G_n be defined as

$$\overline{U}_n = \inf_{\pi\in\Pi} \sup_{\gamma\in\Gamma} U_n^{\pi,\gamma} \qquad \text{and} \qquad \underline{U}_n = \sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} U_n^{\pi,\gamma} \ ,$$

respectively. The authors of [9] establish the following **Key Theorem**, which we state here for the sake of completeness.

Theorem 1 (**Key Theorem**, [9, Theorem 2.1., p. 4])**.** *Let us suppose that the original game G has a value, and that for the sequence of games $\{G_n\}$ and the game G it is verified that*

$\limsup_{n\to\infty} U_n^{\pi,\gamma} - U^{\pi,\gamma} \leqq 0$, *uniformly in $\pi \in \Pi$ for each $\gamma \in \Gamma$,*

$\liminf_{n\to\infty} U_n^{\pi,\gamma} - U^{\pi,\gamma} \geqq 0$, *uniformly in $\gamma \in \Gamma$ for each $\pi \in \Pi$.*

Then

1. $\lim_{n\to\infty} \underline{U}_n = \lim_{n\to\infty} \overline{U}_n = U^*$.
2. *Suppose that there exists N such that for $n \geqq N$, π_n^* and γ_n^* form an ε-equilibrium for the n-th game. Then, for any $\varepsilon' > \varepsilon$, π_n^* and γ_n^* form an ε-equilibrium limit game.*
3. *Let π^* and γ^* be an ε-equilibrium for the limit game. Then for all $\varepsilon' > \varepsilon$, there exist $N(\varepsilon')$ such that π^* and γ^* form an ε'-equilibrium for all $n \geqq N(\varepsilon')$.*

In what follows we are interested in approximating the infinite horizon game defined in (1) through suitably defined games, which satisfy the assumptions of the **Key Theorem** above.

3 Approximations for infinite-horizon games

3.1 Approximations on general state space games

In the context of stochastic games, players have to take simultaneous decisions, based on the knowledge of the current state, but without the certainty of the dynamics of the system, which will be governed by distributions of probabilities on the space of states, known *a priori* for both of them.

Instead, in many situations, players may not have the exact information on these probability distributions (on the system transitions or on the holding times) because it is known only through some statistical method, for example. This lack of information could stem from imprecision on the measure of quantities involved, and it could be improved by some investment of effort or money. Assessing whether this spending is necessary or excessive is an interesting practical issue. In these cases then, there arises the necessity of having bounds to the errors involved when players choose their actions (and then their policies) considering the inexact probability distributions. It is also interesting to study the errors produced by uncertainties on other parameters of the model.

In [4, Section 2.4] the author provides bounds for errors when dealing with approximate transition probabilities and approximate reward functions, on infinite horizon discrete-time **MDP**s, by defining a non-stationary Value Iteration algorithm. In this work we propose a different approach making a sensitivity analysis on these elements (as well as others typical of **SMDP**s), by using the **Key Theorem** stated in the previous section.

Observe that from [3, Theorem 3.1 (c)], under **Assumptions 1** and **2**, the infinite horizon **SMG** has a value V^*, and the existence hypothesis for the value function of the original game in the **Key Theorem** is satisfied.

Through this section, we shall work with approximating games of the form

$$G_n := (\mathcal{S}, \mathcal{A}, \mathcal{B}, \{\mathcal{A}_s : s \in \mathcal{S}\}, \{\mathcal{B}_s : s \in \mathcal{S}\}, Q_n, F_n, \ell_n, \alpha_n) \ , \tag{7}$$

all differing in the transition and holding time probabilities, the reward functions and the discount factors.

In order to quantify the approximations, we shall need adequate norms. On spaces of transition probabilities, we shall use the total variation norm (see for example [4, 8]).

Definition 3.1. *The total variation norm between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{S}$ is defined by:*

$$\|\mathbb{P} - \mathbb{Q}\|_{tv} \ := \sup_{A \subset \mathcal{S} : A \ measurable} |\mathbb{P}(A) - \mathbb{Q}(A)| \ .$$

It is known that (see *e.g.* [8, Proposition D.1., p. 349 and the remarks below, p. 350]):

Proposition 1. *If $\mathbb{P}$ and $\mathbb{Q}$ are two probabilities distributions on $\mathcal{S}$, then*

$$\|\mathbb{P} - \mathbb{Q}\|_{tv} \ = \ \frac{1}{2} \sup_{v, ||v||_\infty \leqq 1} \left| \int_{\mathcal{S}} v \ d\mathbb{P} - \int_{\mathcal{S}} v \ d\mathbb{Q} \right| \ .$$

For the approximate discount factors and holding time distributions, we define the quantities $\beta_n(s,a,b)$ in a similar way to (2):

$$\beta_n(s,a,b) := \int_0^\infty e^{-\alpha_n t} F_n(\mathrm{d}t|s,a,b) \ ,$$

and $\vartheta_n(s,a,b)$ similar to (3)

$$\vartheta_n(s,a,b) := \frac{1-\beta_n(s,a,b)}{\alpha_n} \ .$$

Finally, define

$$r_n(s,a,b) \ = \ \vartheta_n(s,a,b)\ \ell_n(s,a,b)$$

and put $\rho_n := \sup_{(s,a,b)\in\mathbb{K}} \beta_n(s,a,b)$.

We can now make precise the sense in which G_n approximates G.

Assumption 3.

(a) The sequence of transition probabilities Q_n satisfies $||Q(\cdot|s,a,b) - Q_n(\cdot|s,a,b)||_{tv} \to 0$ uniformly for $(s,a,b) \in \mathbb{K}$;

(b) The sequence of probability distributions F_n satisfies $||F(\cdot|s,a,b) - F_n(\cdot|s,a,b)||_{tv} \to 0$ uniformly for $(s,a,b) \in \mathbb{K}$;

(c) The sequence of discount factors α_n satisfies $\alpha_n \to \alpha$;

(d) $\rho_n < 1$ for all n;

(e) The sequence of reward functions ℓ_n satisfies $\ell_n \to \ell$ uniformly on $\mathbb{K}$;

(f) The functions r_n are bounded: there exist $M_n > 0$ such that, for all $(s,a,b) \in \mathbb{K}$, $|r_n(s,a,b)| \leqq M_n$.

For the games defined by (7) we define, given a pair $f \in \Pi_{\text{stat}}$ and $g \in \Gamma_{\text{stat}}$, the corresponding dynamic programming operators

$$(T_n^{f,g}v)(s) = r_n(s,f,g) + \beta_n(s,f,g)\int_{\mathcal{S}} v(z)Q_n(\mathrm{d}z|s,f,g) \ ,$$

and we will denote with $U_n^{f,g}(s)$ the solution of the equation

$$T_n^{f,g}v \ = \ v \ . \tag{8}$$

Lemma 1. *Under* **Assumption 3** (b)-(c), *β_n converges to β uniformly on $\mathbb{K}$. In consequence ϑ_n also converges uniformly to ϑ on $\mathbb{K}$ and ρ_n converges to ρ. With the additional* **Assumption 3** (e), *the sequence r_n converges to $r = \vartheta\ell$, uniformly on $\mathbb{K}$. With the additional* **Assumption 3** (f), *the bounds M_n and M can be chosen satisfying $M_n \to M$.*

Proof. First observe that, if $\alpha_n \to \alpha$, then $e^{-\alpha_n t} \to e^{-\alpha t}$ uniformly on $t \in [0,\infty)$. To see that, consider the functions $h_n(t) = e^{-\alpha t} - e^{-\alpha_n t}$. For each $n \in \mathbb{N}$,

$$\frac{dh_n}{\mathrm{d}t} = -\alpha e^{-\alpha t} + \alpha_n e^{-\alpha_n t} \ ,$$

which vanishes at $t_n^* = \frac{1}{\alpha-\alpha_n}\log\frac{\alpha}{\alpha_n} \geq 0$. At those points

$$h_n(t_n^*) = \left(\frac{\alpha}{\alpha_n}\right)^{\frac{-\alpha}{\alpha-\alpha_n}} - \left(\frac{\alpha}{\alpha_n}\right)^{\frac{-\alpha_n}{\alpha-\alpha_n}} \to \frac{1}{e} - \frac{1}{e} = 0 \ ,$$

as $n \to \infty$, since $\alpha_n \to \alpha$, and then $h_n \to 0$ uniformly on $[0,\infty)$ because $|h_n(t)| \leq h_n(t_n^*)$ for all n and t.

Also observe that, for each $n \in \mathbb{N}$, by Proposition 1, we have

$$\left| \int_0^\infty e^{-\alpha_n t} F(\mathrm{d}t|s,a,b) - e^{-\alpha_n t} F_n(\mathrm{d}t|s,a,b) \right|$$
$$\leqq \sup_{v, ||v||_\infty \leqq 1} \left| \int_0^\infty v(t)\ F(\mathrm{d}t|s,a,b) - v(t)\ F_n(\mathrm{d}t|s,a,b) \right| \leqq 2\ ||F(\cdot|s,a,b) - F_n(\cdot|s,a,b)||_{tv}\ .$$

With the previous observations, given $\varepsilon > 0$, let us consider $N = N(\varepsilon)$, such that, for $n \geqq N$, $\sup_{t \geq 0} |e^{-\alpha t} - e^{-\alpha_n t}| \leqq \frac{\varepsilon}{2}$ and $||F(\cdot|s,a,b) - F_n(\cdot|s,a,b)||_{tv} \leqq \frac{\varepsilon}{4}$. Then

$$\begin{aligned}
|\beta(s,a,b) - \beta_n(s,a,b)| &= \left| \int_0^\infty e^{-\alpha t} F(\mathrm{d}t|s,a,b) - \int_0^\infty e^{-\alpha_n t} F_n(\mathrm{d}t|s,a,b) \right| \\
&= \left| \int_0^\infty \left(e^{-\alpha t} - e^{-\alpha_n t}\right) F(\mathrm{d}t|s,a,b) + \int_0^\infty e^{-\alpha_n t} F(\mathrm{d}t|s,a,b) - e^{-\alpha_n t} F_n(\mathrm{d}t|s,a,b) \right| \\
&\leqq \int_0^\infty |e^{-\alpha t} - e^{-\alpha_n t}| F(\mathrm{d}t|s,a,b) + \left| \int_0^\infty e^{-\alpha_n t} F(\mathrm{d}t|s,a,b) - e^{-\alpha_n t} F_n(\mathrm{d}t|s,a,b) \right| \\
&\leqq \sup_{[0,\infty]} |e^{-\alpha t} - e^{-\alpha_n t}| + 2\ ||F(\cdot|s,a,b) - F_n(\cdot|s,a,b)||_{tv} \leqq \varepsilon
\end{aligned}$$

which shows the stated convergence for β_n. The remaining statements follow easily. □

The following result is easily verified.

Lemma 2. *Under* **Assumption 1 (c)**, for any of stationary strategies $f \in \Pi_{\mathrm{stat}}$ and $g \in \Gamma_{\mathrm{stat}}$, the reward for game G, $U^{f,g}$, defined by (4), and for game G_n, $U_n^{f,g}$, defined by (8), satisfy:

$$||U^{f,g}||_\infty \ \leqq\ \frac{M}{1-\rho}, \qquad ||U_n^{f,g}||_\infty \ \leqq\ \frac{M_n}{1-\rho_n}\ .$$

Theorem 2. *Consider the games G_n defined in (7) and G like in (1). Assume that* **Assumptions 1** and **2** *hold. Let us define, for any pair of stationary strategies $f \in \Pi_{\mathrm{stat}}$ and $g \in \Gamma_{\mathrm{stat}}$, the reward for the limit game G, $U^{f,g}$, by (4), and for the approximating game G_n, $U_n^{f,g}$, by (8). Then, for any pair $f \in \Pi_{\mathrm{stat}}$ and $g \in \Gamma_{\mathrm{stat}}$:*

$$||U^{f,g} - U_n^{f,g}||_\infty \ \leqq\ \varepsilon := \frac{||r - r_n||_\infty}{1-\rho} + \frac{2M\rho}{(1-\rho)^2} \sup_{s \in \mathcal{S}} ||Q(\cdot|s,f,g) - Q_n(\cdot|s,f,g)||_{tv} + \frac{M_n ||\beta - \beta_n||_\infty}{(1-\rho)(1-\rho_n)}$$

and then

$$||V^* - \underline{U}_n||_\infty \ \leqq\ \varepsilon \qquad \textit{and} \qquad ||V^* - \overline{U}_n||_\infty \ \leqq\ \varepsilon\ .$$

If **Assumption 3** *holds in addition, then $U_n^{f,g}$ converges uniformly to $U^{f,g}$ as $n \to \infty$, and the hypotheses of the* **Key Theorem** *hold.*

Proof. For a given pair of policies $f \in \Pi_{\mathrm{stat}}$ and $g \in \Gamma_{\mathrm{stat}}$, and for $s \in \mathcal{S}$ we estimate the difference between $U^{f,g}(s)$ and $U_n^{f,g}(s)$ as follows:

$$\begin{aligned}
|U^{f,g}(s) - U_n^{f,g}(s)| &= \Big| r(s,f,g) + \beta(s,f,g) \int_{\mathcal{S}} U^{f,g}(z) Q(\mathrm{d}z|s,f,g) \\
&\qquad - r_n(s,f,g) - \beta_n(s,f,g) \int_{\mathcal{S}} U_n^{f,g}(z) Q_n(\mathrm{d}z|s,f,g) \Big| \\
&\leqq |r(s,f,g) - r_n(s,f,g)| \ +\ \Delta_n
\end{aligned}$$

where

$$\Delta_n = \left|\beta(s,f,g)\int_{\mathcal{S}} U^{f,g}(z)Q(\mathrm{d}z|s,f,g) - \beta_n(s,f,g)\int_{\mathcal{S}} U_n^{f,g}(z)Q_n(\mathrm{d}z|s,f,g)\right| . \quad (9)$$

Introducing intermediate terms in the difference in (9), we have:

$$\begin{aligned}
\Delta_n \leqq & \left|\beta(s,f,g)\int_{\mathcal{S}} U^{f,g}(z)Q(\mathrm{d}z|s,f,g) - \beta(s,f,g)\int_{\mathcal{S}} U^{f,g}(z)Q_n(\mathrm{d}z|s,f,g)\right| \\
& + \left|\beta(s,f,g)\int_{\mathcal{S}} U^{f,g}(z)Q_n(\mathrm{d}z|s,f,g) - \beta(s,f,g)\int_{\mathcal{S}} U_n^{f,g}(z)Q_n(\mathrm{d}z|s,f,g)\right| \\
& + \left|\beta(s,f,g)\int_{\mathcal{S}} U_n^{f,g}(z)Q_n(\mathrm{d}z|s,f,g) - \beta_n(s,f,g)\int_{\mathcal{S}} U_n^{f,g}(z)Q_n(\mathrm{d}z|s,f,g)\right| \\
\leqq & \rho\left|\int_{\mathcal{S}} U^{f,g}(z)Q(\mathrm{d}z|s,f,g) - \int_{\mathcal{S}} U^{f,g}(z)Q_n(\mathrm{d}z|s,f,g)\right| \\
& + \rho\left|\int_{\mathcal{S}} (U^{f,g}(z) - U_n^{f,g}(z))Q_n(\mathrm{d}z|s,f,g)\right| \\
& + |\beta(s,f,g) - \beta_n(s,f,g)|\left|\int_{\mathcal{S}} U_n^{f,g}(z)Q_n(\mathrm{d}z|s,f,g)\right| .
\end{aligned}$$

Using $||U^{f,g}||_\infty \leqq M/(1-\rho)$ (Lemma 2), the first term can be bounded using Proposition 1:

$$\begin{aligned}
& \left|\int_{\mathcal{S}} U^{f,g}(z)Q(\mathrm{d}z|s,f,g) - \int_{\mathcal{S}} U^{f,g}(z)Q_n(\mathrm{d}z|s,f,g)\right| \\
= & \frac{M}{1-\rho}\left|\int_{\mathcal{S}}\left(\frac{1-\rho}{M}U^{f,g}(z)\right)Q(\mathrm{d}z|s,f,g) - \int_{\mathcal{S}}\left(\frac{1-\rho}{M}U^{f,g}(z)\right)Q_n(\mathrm{d}z|s,f,g)\right| \\
\leqq & \frac{M}{1-\rho}\sup_{v,||v||_\infty\leqq 1}\left|\int_{\mathcal{S}} v(z)\ Q(\mathrm{d}z|s,f,g) - \int_{\mathcal{S}} v(z)\ Q_n(\mathrm{d}z|s,f,g)\right| \\
\leqq & \frac{2M}{1-\rho}||Q(\cdot|s,f,g) - Q_n(\cdot|s,f,g)||_{tv} .
\end{aligned}$$

With the bound on $U_n^{f,g}$ in the second term of Δ_n, we further obtain:

$$\begin{aligned}
\Delta_n \leqq & \frac{2M\rho}{1-\rho}||Q(\cdot|s,f,g) - Q_n(\cdot|s,f,g)||_{tv} \\
& + \rho||U^{f,g} - U_n^{f,g}||_\infty + |\beta(s,f,g) - \beta_n(s,f,g)|\frac{M_n}{1-\rho_n} .
\end{aligned}$$

Finally, for all $s \in \mathcal{S}$,

$$\begin{aligned}
|U^{f,g}(s) - U_n^{f,g}(s)| \leqq & |r(s,f,g) - r_n(s,f,g)| + \frac{2M\rho}{1-\rho}||Q(\cdot|s,f,g) - Q_n(\cdot|s,f,g)||_{tv} \\
& + \rho||U^{f,g} - U_n^{f,g}||_\infty + |\beta(s,f,g) - \beta_n(s,f,g)|\frac{M_n}{1-\rho_n} .
\end{aligned}$$

This implies

$$\begin{aligned}
||U^{f,g} - U_n^{f,g}||_\infty \leqq & ||r - r_n||_\infty + \frac{2M\rho}{1-\rho}\sup_{s\in\mathcal{S}}||Q(\cdot|s,f,g) - Q_n(\cdot|s,f,g)||_{tv} \\
& + \rho||U^{f,g} - U_n^{f,g}||_\infty + \frac{M_n||\beta - \beta_n||_\infty}{1-\rho_n} ,
\end{aligned}$$

which implies in turn the claim:

$$||U^{f,g} - U_n^{f,g}||_\infty \leqq \frac{||r - r_n||_\infty}{1-\rho} + \frac{2M\rho}{(1-\rho)^2} \sup_{s\in\mathcal{S}} ||Q(\cdot|s,f,g) - Q_n(\cdot|s,f,g)||_{tv} + \frac{M_n||\beta - \beta_n||_\infty}{(1-\rho)(1-\rho_n)} .$$

The proposed bounds on the lower and the upper values of the approximate games are justified by [3, Lemma A.2.].

The uniform convergence follows from **Assumption 3** and Lemma 1. □

Remark 1. *Clearly, if the approximated games G_n have a value U_n^*, then*

$$||U_n^* - V^*||_\infty \leqq \frac{||r - r_n||_\infty}{1-\rho} + \frac{2M\rho}{(1-\rho)^2} \sup_{s\in\mathcal{S}} ||Q(\cdot|s,f,g) - Q_n(\cdot|s,f,g)||_{tv} + \frac{M_n||\beta - \beta_n||_\infty}{(1-\rho)(1-\rho_n)}.$$

3.2 Approximations on denumerable state space games

In this subsection, **Assumption 1**(a) is replaced by

(a') The state space $\mathcal{S}$ is denumerable,

and consequently, **Assumption 1**(f) takes the particular form

(f') For each $s \in \mathcal{S}$ and each bounded function v on $\mathcal{S}$, the function $(a,b) \mapsto \sum_{z\in\mathcal{S}} v(z)Q(z|s,a,b)$ is continuous on $\mathcal{A}_s \times \mathcal{B}_s$.

In this context other kind of approximations can be done by reducing the space of states conveniently and computing the values of the new games.

A review of the history and the motivation of the theory of state truncation in control problems is done in the introduction of [1, Chapter 16, p. 205], devoted to this kind of approximation on constrained **MDP**s.

The approximations we consider on infinite-horizon games are based on some increasing sequence $\mathcal{S}_n \subset S_{n+1} \subset \mathcal{S}$ of subsets of the state space, with $\mathcal{S}_0 \neq \emptyset$. Typically (but not necessarily for the results we state) this sequence converges to the original state space, i.e., $\bigcup_{n\in\mathbb{N}} \mathcal{S}_n = \mathcal{S}$.

Depending on the convenience of the case we can assume one of the following assumptions:

Assumption 4. *For all integers ν,*

$$\varepsilon(\nu,n) = \sup_{s\in\mathcal{S}_\nu, a\in\mathcal{A}_s, b\in\mathcal{B}_s} \left\{ \sum_{z\notin\mathcal{S}_n} Q(z|s,a,b) \right\} \to 0$$

as $n \to \infty$.

Assumption 5. *From any state $k \in \mathcal{S}$, only a finite set of states $\mathcal{S}_{(k)}$ can be reached:*

$$\forall k \in \mathcal{S}, \quad \#\{z \in \mathcal{S}, \exists a \in \mathcal{A}_k, \exists b \in \mathcal{B}_k, Q(z|k,a,b) > 0\} \ < \ \infty .$$

Suppose now that we have chosen one way construct the sequence of state spaces $\{\mathcal{S}_n\}$, subsets of the original space $\mathcal{S}$. Associated to each set $\mathcal{S}_n$, we can construct a new game

$$G_n := (\mathcal{S}_n, \mathcal{A}_n, \mathcal{B}_n, \{\mathcal{A}_s : s \in \mathcal{S}_n\}, \{\mathcal{B}_s : s \in \mathcal{S}_n\}, Q_n, F, r, \alpha) \tag{10}$$

with $\mathcal{A}_n = \bigcup_{s\in\mathcal{S}_n} \mathcal{A}_s$, $\mathcal{B}_n = \bigcup_{s\in\mathcal{S}_n} \mathcal{B}_s$ as the corresponding action sets for players 1 and 2.

At this point we must define the transition laws for these new games. That is, given $s, z \in \mathcal{S}_n$, $a \in \mathcal{A}_s$ and $b \in \mathcal{B}_s$, assign a new value $Q_n(z|s,a,b)$. We consider two different schemes.

In the first one, we eliminate transitions outside the sets $\mathcal{S}_n$ and redirect them to a specific, constant state $s^* \in \mathcal{S}_0$. Formally, for $s \in \mathcal{S}_n$, $a \in \mathcal{A}_s$ and $b \in \mathcal{B}_s$,

$$Q_n(z|s,a,b) = \begin{cases} Q(z|s,a,b) + \sum_{w\notin\mathcal{S}_n} Q(w|s,a,b) & z = s^* \\ Q(z|s,a,b) & z \in \mathcal{S}_n \setminus \{s^*\} \\ 0 & z \notin \mathcal{S}_n \ , \end{cases} \tag{11}$$

while $Q_n(z|s,a,b) = Q(z|s,a,b)$ for $s \notin \mathcal{S}_n$, $a \in \mathcal{A}_s$ and $b \in \mathcal{B}_s$.

It is argued in [1, Section 16.3, p. 211] that the way probabilities are defined in (11) may not be convenient. This motivates the introduction of the following more general scheme.

In this second way of redefining the transition probabilities we put

$$Q_n(z|s,a,b) = \begin{cases} Q(z|s,a,b) + q_n(z|s,a,b) & z \in \mathcal{S}_n \\ 0 & z \notin \mathcal{S}_n \end{cases} \tag{12}$$

where $q_n(.|s,a,b)$ is any family of positive numbers which satisfy $\sum_{z\in\mathcal{S}_n} Q(z|s,a,b)+q_n(z|s,a,b) = 1$. Hence,

$$\sum_{z\in S_n} q_n(z|s,a,b) = \sum_{z\notin S_n} Q(z|s,a,b). \tag{13}$$

As before, $Q_n(z|s,a,b) = Q(z|s,a,b)$ for $s \notin \mathcal{S}_n$, $a \in \mathcal{A}_s$ and $b \in \mathcal{B}_s$.

Once the Q_n have been defined, we obtain the corresponding dynamic programming operators, for each f and g:

$$(T_n^{f,g}v)(s) = \begin{cases} r(s,f,g) + \beta(s,f,g)\sum_{z\in\mathcal{S}_n} v(z)Q_n(z|s,f,g) & \text{if } s \in \mathcal{S}_n \\ 0 & \text{if } s \notin \mathcal{S}_n \end{cases}$$

and we denote with $U_n^{f,g}(s)$ the solution of the equation

$$T_n^{f,g}v \ = \ v. \tag{14}$$

The following observation will be repeatedly used: under **Assumption 2**, for any n and any $s \in \mathcal{S}_n$,

$$\begin{aligned} |U^{f,g}(s) - U_n^{f,g}(s)| \ &= \ \Big| r(s,f,g) + \beta(s,f,g)\sum_{z\in\mathcal{S}} U^{f,g}(z)Q(z|s,f,g) \\ &\qquad - r(s,f,g) - \beta(s,f,g)\sum_{z\in\mathcal{S}} U_n^{f,g}(z)Q_n(z|s,f,g)\Big| \\ &= \ \beta(s,f,g)\sum_{z\in\mathcal{S}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \\ &\leqq \ \rho\sum_{z\in\mathcal{S}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \ . \end{aligned} \tag{15}$$

We can prove the following theorem of approximation:

Theorem 3. *Consider the games G_n defined in (10) and G like in (1), with the transition probabilities Q_n defined as in (12). Under* **Assumptions 1**, **2**, *and* **4**, *all the hypotheses of the* **Key Theorem** *hold where for a given pair of stationary policies $f \in \Pi_{\text{stat}}$ and $g \in \Gamma_{\text{stat}}$, the reward $U^{f,g}$ is defined by (4) for the limit game, and $U_n^{f,g}$ by (14) for the approximating game G_n.*

Proof. In the proof of this result, we follow the ideas presented in [2] and [1] for **MDP** models and in [9] for the **MG** case. Particularly, we adopt most the notation from the last one.

Let us for $\varepsilon > 0$ note $g^0(\varepsilon,\nu) = \nu$ and for $k = 1, 2, ...$, $g^k(\varepsilon,\nu) = g(\varepsilon, g^{k-1}(\varepsilon,\nu))$ where

$$g(\varepsilon,\nu) = \min\{m : \varepsilon(\nu,m) \leqq \varepsilon\}$$

and the value $\varepsilon(\nu,m)$ taken from **Assumption 4**. The fact that all terms in this sequence are finite is precisely a consequence of this assumption.

Note that, for any $\nu \geqq 0$,

$$\sum_{z \notin \mathcal{S}_{g^{l+1}(\varepsilon,\nu)}} Q(z|s,f,g) \leqq \varepsilon, \quad \text{for all } s \in \mathcal{S}_{g^l(\varepsilon,\nu)},\ l \geqq 0,\ f \in \Pi_{\text{stat}},\ g \in \Gamma_{\text{stat}} \tag{16}$$

and that $g^l(\varepsilon,\nu)$ need not be increasing in l.

For any subset $\mathcal{J} \subset \mathcal{S}$, denote $\tau(\mathcal{J}) = \min\{m : \mathcal{J} \subset \mathcal{S}_m\}$. In the remainder of the proof, we shall consider a fixed set $\mathcal{J}$ such that $\tau(\mathcal{J}) < \infty$. Finally, define

$$m_k(\varepsilon,\tau(\mathcal{J})) = \max\{\tau(\mathcal{J}), g(\varepsilon,\tau(\mathcal{J})), ..., g^k(\varepsilon,\tau(\mathcal{J}))\}, \quad k = 0, 1, 2, ... \tag{17}$$

To simplify the notation, since from now on, ε and $\mathcal{J}$ will be supposed fixed, we shall write g^l instead of $g^l(\varepsilon,\tau(\mathcal{J}))$ and m_k instead of $m_k(\varepsilon,\tau(\mathcal{J}))$.

Starting from (15), for any pair of stationary strategies $f \in \Pi_{\text{stat}}$ and $g \in \Gamma_{\text{stat}}$, any n and $s \in \mathcal{S}_n$, we have

$$\begin{aligned}
|U^{f,g}(s) - U_n^{f,g}(s)| &\leqq \rho \sum_{z \in \mathcal{S}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \\
&= \rho \sum_{z \in \mathcal{S}_{g^1}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \\
&\quad + \rho \sum_{z \in \mathcal{S} \setminus \mathcal{S}_{g^1}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \ .
\end{aligned}$$

We now consider successively the two terms in this last expression. Observe that for $n \geqq g^1$, $z \in \mathcal{S}_{g^1}$, implies $Q_n(z|s,a,b) = Q(z|s,a,b) + q_n(z|s,a,b)$ according to (12). It follows that, for all $n \geqq g^1$ and $s \in S_0$, using (12) together with the identity (13), property (16) and the bound

on $U_n^{f,g}$ provided by Lemma 2,

$$\begin{aligned}
&\sum_{z\in\mathcal{S}_{g^1}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \\
&= \sum_{z\in\mathcal{S}_{g^1}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)q_n(z|s,f,g)| \\
&\leqq \sum_{z\in\mathcal{S}_{g^1}} |U^{f,g}(z) - U_n^{f,g}(z)|Q(z|s,f,g) + \sum_{z\in\mathcal{S}_{g^1}} |U_n^{f,g}(z)|q_n(z|s,f,g) \\
&\leqq \sup_{z\in\mathcal{S}_{g^1}} |U^{f,g}(z) - U_n^{f,g}(z)| + \frac{M}{1-\rho}\sum_{z\in\mathcal{S}_{g^1}} q_n(z|s,f,g) \\
&= \sup_{z\in\mathcal{S}_{g^1}} |U^{f,g}(z) - U_n^{f,g}(z)| + \frac{M}{1-\rho}\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} Q(z|s,f,g) \\
&\leqq \sup_{z\in\mathcal{S}_{g^1}} |U^{f,g}(z) - U_n^{f,g}(z)| + \frac{\varepsilon M}{1-\rho}\ .
\end{aligned}$$

For the second term, if $z \notin \mathcal{S}_{g^1}$, by (12), $Q_n(z|s,f,g)$ is either 0 (if $z \notin \mathcal{S}_n$) or $Q(z|s,f,g) + q_n(z|s,f,g)$ (if $z \in \mathcal{S}_n \setminus \mathcal{S}_{g^1}$). In both cases $Q_n(z|s,f,g) \leqq Q(z|s,f,g)+q_n(z|s,f,g)$, which gives, still for $n \geqq g^1$ and $s \in \mathcal{S}_{g^0}$:

$$\begin{aligned}
&\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \\
&\leqq \sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U^{f,g}(z)|Q(z|s,f,g) + \sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U_n^{f,g}(z)|Q_n(z|s,f,g) \\
&\leqq \sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U^{f,g}(z)|Q(z|s,f,g) + \sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U_n^{f,g}(z)|(Q(z|s,f,g) + q_n(z|s,f,g)) \\
&= \sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U^{f,g}(z)|Q(z|s,f,g) + \sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U_n^{f,g}(z)|Q(z|s,f,g) + \sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U_n^{f,g}(z)|q_n(z|s,f,g) \\
&\leqq \frac{M}{1-\rho}\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} Q(z|s,f,g) + \frac{M}{1-\rho}\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} Q(z|s,f,g) + \frac{M}{1-\rho}\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} q_n(z|s,f,g) \\
&\leqq \frac{2\varepsilon M}{1-\rho} + \frac{M}{1-\rho}\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} q_n(z|s,f,g)\ ,
\end{aligned}$$

using again the bounds of Lemma 2, and the fact that $\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} Q(z|s,f,g) \leqq \varepsilon$, from (16). At this point observe also that

$$\begin{aligned}
\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} q_n(z|s,f,g) &= \sum_{z\in\mathcal{S}_n\setminus\mathcal{S}_{g^1}} q_n(z|s,f,g) \\
&\leqq \sum_{z\in\mathcal{S}_n} q_n(z|s,f,g) = \sum_{z\in\mathcal{S}\setminus\mathcal{S}_n} Q(z|s,f,g) \\
&\leqq \sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} Q(z|s,f,g) \leqq \varepsilon\ ,
\end{aligned}$$

and in consequence

$$\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \leqq \frac{3\varepsilon M}{1-\rho} \ . \tag{18}$$

Summarizing, for any $n \geq g^1$ and any $s \in \mathcal{S}_{g^0}$,

$$|U^{f,g}(s) - U_n^{f,g}(s)| \quad \leqq \quad \rho \sup_{z\in\mathcal{S}_{g^1}} |U^{f,g}(z) - U_n^{f,g}(z)| + \frac{4\varepsilon M\rho}{1-\rho} \ . \tag{19}$$

Using the same reasoning, we readily obtain that for $n \geqq g^2$ and $z \in \mathcal{S}_{g^1}$,

$$|U^{f,g}(z) - U_n^{f,g}(z)| \quad \leqq \quad \rho \sup_{w\in\mathcal{S}_{g^2}} |U^{f,g}(w) - U_n^{f,g}(w)| + \frac{4\varepsilon M\rho}{1-\rho} \ , \tag{20}$$

and more generally, for any $k \in \mathbb{N}$, $n \geqq g^{k+1}$ and $z \in \mathcal{S}_{g^k}$,

$$|U^{f,g}(z) - U_n^{f,g}(z)| \quad \leqq \quad \rho \sup_{w\in\mathcal{S}_{g^{k+1}}} |U^{f,g}(w) - U_n^{f,g}(w)| + \frac{4\varepsilon M\rho}{1-\rho} \ .$$

Then, combining (19) and (20), for any $n \geqq m_2 = \max(g^0, g^1, g^2)$ and $s \in \mathcal{S}_{g^0}$,

$$|U^{f,g}(s) - U_n^{f,g}(s)| \quad \leqq \quad \rho^2 \sup_{w\in\mathcal{S}_{g^2}} |U^{f,g}(w) - U_n^{f,g}(w)| + \frac{4\varepsilon M\rho^2}{1-\rho} + \frac{4\varepsilon M\rho}{1-\rho} \ .$$

In general, for any $k \in \mathbb{N}$, $n \geqq m_k$ and $s \in \mathcal{S}_{g^0}$ we obtain

$$\begin{aligned} |U^{f,g}(s) - U_n^{f,g}(s)| \quad &\leqq \quad \rho^k \sup_{w\in\mathcal{S}_{g^k}} |U^{f,g}(w) - U_n^{f,g}(w)| + \frac{4\varepsilon M\rho}{1-\rho} \sum_{\nu=0}^{k-1} \rho^\nu \\ &\leq \quad \rho^k \frac{2M}{1-\rho} \ + \ \frac{4\varepsilon M\rho}{1-\rho} \ \frac{1-\rho^k}{1-\rho} \ . \end{aligned}$$

This bound is independent of f and g and can be made as small as needed. This proves the uniform convergence needed in the **Key Theorem**. □

Remark 2. *If in Theorem 3 we assume the redefinition of transition probabilities given by (11), we have, for $z \notin \mathcal{S}_{g^1}$, $Q_n(z|s,f,g) \leqq Q(z|s,f,g)$, and (18) can be put in the tighter form*

$$\sum_{z\in\mathcal{S}\setminus\mathcal{S}_{g^1}} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \leqq \frac{2\varepsilon M}{1-\rho} \ .$$

Consequently in this case, adapting the proof of the theorem, for any $k \in \mathbb{N}$, $n \geqq m_k$ and $s \in \mathcal{S}_{g^0}$,

$$|U^{f,g}(s) - U_n^{f,g}(s)| \quad \leqq \quad \rho^k \frac{2M}{1-\rho} \ + \ \frac{3\varepsilon M\rho}{1-\rho} \ \frac{1-\rho^k}{1-\rho} \ .$$

With the aim to compute an uniform bound on the approximation error on a given set of states $\mathcal{S}_0 = \mathcal{J}$, let us define the sequence $\{\mathcal{S}_n\}$ by

$$\mathcal{S}_0 = \mathcal{J}, \quad \mathcal{S}_{n+1} = \bigcup_{s \in \mathcal{S}_n} \mathcal{Y}(s) \bigcup \mathcal{S}_n \ , \tag{21}$$

where $\mathcal{Y}(s) = \{z : Q(z|s,f,g) > 0 \text{ for some } f,\, g\}$. The set $\mathcal{S}_n$ is exactly the set of states that can be reached with positive probability from one state of $\mathcal{J}$ after exactly n transitions, under any stationary policy. Under **Assumptions 5**, this is a sequence of finite sets, provided that $\mathcal{J}$ be finite.

In the last scheme, we modify the transition probabilities eliminating transitions outside the sets $\mathcal{S}_n$, but without changing any other probability in that set:

$$Q_n(z|s,a,b) = \begin{cases} Q(z|s,a,b) & z \in \mathcal{S}_n \\ 0 & z \notin \mathcal{S}_n \ , \end{cases} \tag{22}$$

and again $Q_n(z|s,a,b) = Q(z|s,a,b)$ for $s \notin \mathcal{S}_n$, $a \in \mathcal{A}_s$ and $b \in \mathcal{B}_s$. This results in possibly defective transition probabilities. Because of this, we do not obtain, strictly speaking, a **SMG**. See the arguments of by [4, Section 2.4, pp. 32–33], related to finite-state approximations on discounted **MDP** models.

On the other hand, looking at the proof of [3, Theorem 3.1], on value functions and optimal policies of semi-Markov games, the fact that $Q(\cdot|s,a,b)$ be a probability measure is not necessary. It is used only the inequality $Q(\mathcal{S}|s,a,b) \leqq 1$. It follows that results such as the **Key Theorem** can be applied to "games" with possibly defective transition probabilities.

Under this scheme with the construction of the subsets given by (21), we prove the following result.

Theorem 4. *Under* **Assumptions 1**, **2** *and* **5**, *all statements of the* **Key Theorem** *hold for the approximation considered with the construction of the truncated space of states given by (21), where for a given pair of stationary policies $f \in \Pi_{\text{stat}}$ and $g \in \Gamma_{\text{stat}}$, the reward $U^{f,g} = V^{f,g}$ is defined in (4) for the limit game, and for the approximating game G_n, $U_n^{f,g}$ is defined by (14), where the transition probabilities are redefined by (22). Moreover, for $n \in \mathbb{N}$,*

$$||V^* - \underline{U}_n||_\infty \leqq \frac{2M\rho^n}{1-\rho} \qquad \textit{and} \qquad ||V^* - \overline{U}_n||_\infty \leqq \frac{2M\rho^n}{1-\rho} \ .$$

Proof. Let us fix any state $s \in \mathcal{J}$, and consider a pair of stationary policies $f \in \Pi_{\text{stat}}$ and $g \in \Gamma_{\text{stat}}$. Then, since from s only the states of $\mathcal{Y}(s)$ are reachable (either with Q or Q_n), (15) gives:

$$\begin{aligned} |U^{f,g}(s) - U_n^{f,g}(s)| &\leqq \rho \sum_{z \in \mathcal{Y}(s)} |U^{f,g}(z)Q(z|s,f,g) - U_n^{f,g}(z)Q_n(z|s,f,g)| \\ &\leqq \rho \sum_{z \in \mathcal{Y}(s)} |U^{f,g}(z) - U_n^{f,g}(z)|Q(z|s,f,g) \ \leqq \ \rho \sup_{z \in \mathcal{S}_1} |U^{f,g}(z) - U_n^{f,g}(z)| \\ &\leqq \rho^2 \sup_{z \in \mathcal{S}_1} \sum_{w \in \mathcal{Y}(z)} |U^{f,g}(w) - U_n^{f,g}(w)|Q(w|z,f,g) \\ &\leqq \rho^2 \sup_{w \in \mathcal{S}_2} |U^{f,g}(w) - U_n^{f,g}(w)| \ . \end{aligned}$$

Recursively, we obtain for all $n \in \mathbb{N}$ and $s \in \mathcal{J}$,

$$|U^{f,g}(s) - U_n^{f,g}(s)| \leqq \rho^n ||U^{f,g} - U_n^{f,g}||_\infty \leqq \frac{2M\rho^n}{1-\rho} \ .$$

The bounds for the lower and the upper bounds follow from [3, Lemma A.2]. □

4 Approximations on finite horizon games and approximated rolling horizon procedure

So far in this work, we have approximated in several ways the value function and the equilibria of infinite-horizon zero-sum semi-Markov games. In the present section we are interested in the application of such approximations to finite-horizon games of this type.

Consider a semi-Markov game with all the parameters defined in (1) but with finite stage-horizon N. In Section 2, for a given pair of policies $\pi \in \Pi$ and $\gamma \in \Gamma$, the discounted reward criterion $V_n^{\pi,\gamma}$ was defined by Equation (5).

It is known (see, for instance [5], [7]) that under **Assumptions 1** and **2** the value function V_N^* of the N-stage game exists and can be obtained by successive application of the corresponding dynamic programming operator defined by Equation (6), in the recursion $V_0^* = 0$ and $V_k^* = TV_{k-1}^*$, $k = 1, 2, ..., N$, and that any pair of policies $\{f_0^*, f_1^*, ..., f_{N-1}^*\}$, $\{g_0^*, g_1^*, ..., g_{N-1}^*\}$ constructed with the successive maximinimizing actions forms an equilibrium pair for this finite-stage horizon game.

In order to apply the results from the previous sections to finite-horizon games, we make the following observation. The finite horizon model is equivalent to the following infinite horizon model with enlarged state space, similar to the construction made in [9, Section 5, p. 16-17] for **MG** models. Define the game $\hat{G} := (\hat{\mathcal{S}}, \hat{\mathcal{A}}, \hat{\mathcal{B}}, \{\hat{\mathcal{A}}_{\hat{s}} : \hat{s} \in \hat{\mathcal{S}}\}, \{\hat{\mathcal{B}}_{\hat{s}} : \hat{s} \in \hat{\mathcal{S}}\}, \hat{Q}, \hat{F}, \hat{\ell}, \alpha)$ as

- $\hat{\mathcal{S}} = \mathcal{S} \times \{0, 1, ..., N\}$
- $\hat{\mathcal{A}}_{(s,k)} = \mathcal{A}_s$, $\hat{\mathcal{B}}_{(s,k)} = \mathcal{B}_s$, for all $(s,k) \in \hat{\mathcal{S}}$
- $\hat{r}((s,k),a,b) = r(s,a,b)$, for all $(s,k) \in \hat{\mathcal{S}}$, $a \in \mathcal{A}_s$, $b \in \mathcal{B}_s$
- $\hat{Q}((z,l)|(s,k),a,b) = \begin{cases} Q(z|s,a,b) & k+1 = l \leqq N \\ 0 & \text{otherwise} \end{cases}$
- $\hat{F}(\cdot|(s,k),a,b) = F(\cdot|s,a,b)$ for all $(s,k) \in \hat{\mathcal{S}}$, $a \in \mathcal{A}_s$, $b \in \mathcal{B}_s$.

Observe that from the definition of the elements of this new game, it follows that $\hat{\beta}((s,k),a,b) = \beta(s,a,b)$, $\hat{\vartheta}((s,k),a,b) = \vartheta(s,a,b)$, and $\hat{r}((s,k),a,b) = r(s,a,b)$ for all $(s,k) \in \hat{\mathcal{S}}$, $a \in \mathcal{A}_s$, $b \in \mathcal{B}_s$.

Observe also that, by construction, the transition probabilities are defective from states of the form (s,N), causing the game to effectively terminate after a finite number of stages.

The construction of the different sets of strategies for each player is similar to the one made in Section 2. Noting with $\hat{\Pi}$ and $\hat{\Gamma}$ the set of Markov policies for both players in this new game, for a pair of policies $\hat{\pi} \in \hat{\Pi}$ and $\hat{\gamma} \in \hat{\Gamma}$, for $\hat{s} \in \hat{\mathcal{S}}$, the reward for the N-stage horizon game is defined by

$$\hat{V}_N^{\hat{\pi},\hat{\gamma}}(s) := \mathbb{E}_{\hat{s}}^{\hat{\pi},\hat{\gamma}} \sum_{k=0}^{N-1} e^{-\alpha T_k} r(\hat{S}_k, \hat{A}_k, \hat{B}_k) \ , \tag{23}$$

where $\{\hat{S}_k\}$, $\{\hat{A}_k\}$ and $\{\hat{B}_k\}$ and $\{\hat{T}_k\}$ are the stochastic processes on $\hat{\mathcal{S}}$, $\hat{\mathcal{A}}$, $\hat{\mathcal{B}}$ and $[0,\infty)$ respectively, from application of policies $\hat{\pi}$ and $\hat{\gamma}$.

Let us note that there is a one to one correspondence between stationary policies in the new infinite horizon game and Markov policies in the original one. If $\hat{\pi}$ and $\hat{\gamma}$ are stationary in the new game, then we can obtain Markov policies in the original one given by

$$\pi_t(.|s) = \hat{\pi}(.|(s,t)), \quad \gamma_t(.|s) = \hat{\gamma}(.|(s,t)) \tag{24}$$

for $t = 0, 1, ..., N$, and vice versa, and the same holds for stationary strategies $\hat{f}$ and $\hat{g}$ for the augmented game.

Applying the construction given by (21) to the enlarged spaces we obtain the sequence, for an initial subset $\hat{\mathcal{J}} \subset \hat{\mathcal{S}}$,

$$\hat{\mathcal{S}}_0 = \hat{\mathcal{J}}, \quad \hat{\mathcal{S}}_{n+1} = \bigcup_{(s,k)\in\hat{\mathcal{S}}_n} \mathcal{Y}((s,k)) \bigcup \hat{\mathcal{S}}_n \ , \tag{25}$$

where $\mathcal{Y}(s) = \{(z,l) : \hat{Q}((z,l)|(s,k), \hat{f}, \hat{g}) > 0 \ \text{ for some } \hat{f}, \hat{g}\}$. Actually, if $\mathcal{J} \subset \mathcal{S} \times \{0\}$ as in the preceding construction, then $\hat{\mathcal{S}}_n = S_n \times \{n\}$.

In this case, for $\hat{U}_n$, the functions which satisfies, for all $(s,k) \in \hat{\mathcal{S}}_n$

$$\hat{U}_n((s,k)) = \sup_{\hat{f}} \inf_{\hat{g}} \left\{ r((s,k), \hat{f}, \hat{g}) + \beta((s,k), \hat{f}, \hat{g}) \sum_{(z,l)\in\hat{\mathcal{S}}_n} \hat{U}_n((z,l)) \hat{Q}_n((z,l)|(s,k), \hat{f}, \hat{g}) \right\} \ ,$$

and $\hat{U}_n((s,k)) = 0$ if $(s,k) \notin \hat{\mathcal{S}}_n$, according to Theorem 4, for any $(s,k) \in \hat{\mathcal{J}}$,

$$|\hat{U}_n((s,k)) - \hat{V}_N^*(s)| \leqq \frac{2M\rho^n}{1-\rho}.$$

As an application of the previous construction let us consider the Approximate Rolling Horizon (**ARH**) procedure, used to obtain approximated optimal policies in many optimization problems. In particular, in our games, it formulates as follows.

ARH1 Choose some function V a priori near V_N^* where V_N^* is the N-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 $\tilde{f}_N(s_t)$, $\tilde{g}_N(s_t)$ are 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 $s_t := s_{t+1}$ and go to step 2.

In [3, Corollary 4.4], under **Assumptions 1** and **2**, we give bounds to the error produced by utilization of the **ARH** strategies in the infinite horizon games, as a function of the error between V and V_N^*.

Also from the **ARH** framework, given an approximate value function V of V_N^*, the maximizer can take his decision on the supposition that minimizer actually plays in his worst-case scenario. In that case, we have studied the errors in [3, Corollary 4.5] under the same assumptions.

As it was stated in **ARH1**, approximations of the value of a finite-stage horizon game, with the corresponding error bound, are necessary in order to utilize this procedure and be able to compute the final bounds of the error incurred.

5 Concluding remarks

Through this work we have dealt with zero-sum semi-Markov games models with discounted payoff and bounded rewards.

In Section 3 we studied approximations of the value function of the infinite-horizon game and their equilibria, by considering it as a limit of a sequence of approximating games. Several ways of constructing this approximating games were considered, each of them perturbing some parameter of the game. Specifically, in Subsection 3.1 we have considered the original game as the limit of a sequence of games with approximate transition and holding time probabilities, reward functions and discount factors converging to the original ones. In Subsection 3.2, for the case of denumerable space of states, we study the approximations by convenient truncation of it.

In Section 4 we applied all the approximations made on the previous ones to the finite-stage horizon game, by application of the corresponding infinite-horizon theorems on a suitable enlarged space of states.

Finally, we applied the results obtained Section 4 to obtain approximate values functions to finite-stage horizon games needed to initialize the Approximate Rolling Horizon method stated in [3], to obtain new approximations for the infinite-horizon game.

References

[1] Altman E., *Constrained Markov Decision Processes.* Chapman and Hall, 1999.

[2] Cavazos-Cadena R.; "Finite-state approximations for denumerable state discounted Markov Decision Processes". *Journal of Applied Mathematics and Optimization*, 14, 1986, pp. 27–47. *IEEE Transactions on Automatic Control*, 48, 11, 2003, pp. 1951–1961.

[3] Della Vecchia E., Di Marco S., Jean-Marie A., "Rolling horizon procedures in Semi-Markov Games: The Discounted Case", INRIA Research Report 8019, July 2012, `http://hal.inria.fr/hal-00720351`.

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

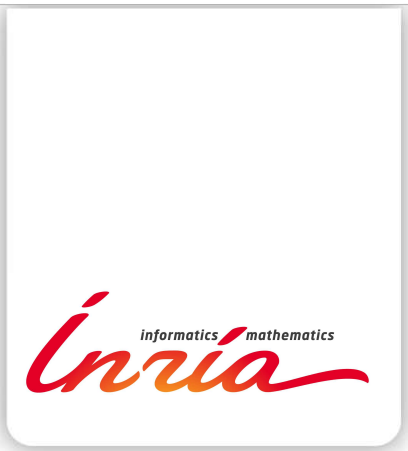
[5] Jaskiewicz A., Nowak A.; "Approximations of noncooperative semimarkov games". *Journal of optimization, theory and applications*, **131**(1), 2006, pp. 115–134.

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

[7] Luque-Vásquez F.; *Zero-sum semi-Markov game in Borel spaces with discounted payoff.* Universidad de Sonora, Mexico, 2002.

[8] Robert P.; *Réseaux et files d'attente: méthodes probabilistes.* Mathématiques et Applications, Vol. 35, Springer Verlag, 2000.

[9] Tidball M., Altman E.; "Approximations in dynamics zero-sum games". *SIAM J. Control and Optimization*, 34, 1, 1996, pp. 311–328.



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