TESIS DOCTORAL

2015

Aproximaciones de modelos de cadenas de Markov controladas y juegos markovianos en tiempo continuo.

Approximation results for continuous-time controlled Markov chains and Markov games

José María Lorenzo Magán

Licenciado en Ciencias Matemáticas
Diploma de Estudios Avanzados en el Programa de Doctorado
"Sistemas Estocásticos y su Control Óptimo", UNED.

Universidad Nacional de Educación a Distancia, UNED
Facultad de Ciencias
Departamento de Estadística, Investigación Operativa
y Cálculo Numérico

Programa de Doctorado
"Sistemas Estocásticos y su Control Óptimo"

Director: Dr. Tomás Prieto Rumeau

Universidad Nacional de
Educación a Distancia, UNED

Facultad de Ciencias

Departamento de Estadística, Investigación Operativa
y Cálculo Numérico

Aproximaciones de modelos de cadenas de Markov controladas y juegos markovianos en tiempo continuo.

Approximation results for continuous-time controlled Markov chains and Markov games

por José María Lorenzo Magán

Licenciado en Ciencias Matemáticas
Diploma de Estudios Avanzados en el Programa de Doctorado
"Sistemas Estocásticos y su Control Óptimo", UNED.

Para optar al título de Doctor en Matemáticas por la Universidad Nacional de Educación a Distancia.

Director: Dr. Tomás Prieto Rumeau

Noviembre de 2015

Indice

Resumen

Esta tesis estudia métodos de aproximación para cadenas de Markov controladas en tiempo continuo y para juegos markovianos bipersonales de suma cero en tiempo continuo. Estos modelos dinámicos ya han sido estudiados desde el punto de vista teórico pero, en general, no es posible obtener explícitamente los valores óptimos de los problemas ni las estrategias óptimas, debido a la complejidad de las correspondientes ecuaciones de optimalidad. Es por ello que se introducen aquí métodos de aproximación que permitan aproximar numéricamente dichos valores óptimos y las correspondientes estrategias óptimas.

En un contexto más general, la idea es proponer una definición de convergencia de una sucesión $\{\mathcal{M}_n\}_{n\geq 1}$ de modelos de cadenas de Markov controladas a un modelo $\mathcal{M}$, cuya solución óptima se quiere aproximar. Se darán entonces condiciones bajo las cuales la convergencia $\mathcal{M}_n \to \mathcal{M}$ implique la convergencia de los valores óptimos y de las políticas óptimas de $\mathcal{M}_n$ a los de $\mathcal{M}$. Esta misma problemática se abordará para la convergencia $\mathcal{G}_n \to \mathcal{G}$ para juegos de Markov de suma nula.

Los modelos de control y juegos considerados tienen espacio de estados numerable, espacios de acciones de Borel, y sus tasas de transición y pago pueden ser no acotadas. Se estudiarán los criterios de optimalidad del pago descontado y del pago promedio. Las hipótesis principales que se harán sobre estos modelos incluyen desigualdades de tipo Lyapunov sobre las tasas de transición, continuidad del pago y de las tasas de transición, y compacidad de los conjuntos de acciones. Además de los resultados de convergencia de los valores óptimos de los modelos de control $\mathcal{M}$ y juegos $\mathcal{G}$, se estudiarán las tasas de convergencia de los valores óptimos de $\mathcal{M}_n$ y $\mathcal{G}_n$, cuando estos se definen mediante una truncación finita de los espacios de estados y acciones de los modelos originales. Se probará que estas tasas están estrechamente relacionadas con el máximo exponente para el que se obtiene una desigualdad de Lyapunov.

Los resultados teóricos obtenidos se ilustran con varias aplicaciones a modelos de poblaciones y procesos de nacimiento y muerte. De esta manera, se prueba también que los métodos de aproximación estudiados son una herramienta potente que permite estimar con precisión la solución óptima de modelos estocásticos de decisión complejos.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The goal of this thesis is to propose techniques to approximate continuous-time controlled Markov chains and Markov games. The motivation for such approximations is mainly practical. Indeed, the aforementioned control and game models have been extensively studied theoretically, but it is not possible in general to determine explicitly the optimal value or the optimal strategies. The purpose of the thesis is precisely to tackle the problem of approximating numerically such optimal solutions.

In a more general framework, the idea is to propose a definition of convergence of a sequence of control models $\{\mathcal{M}_n\}_{n\geq 1}$ to a control model $\mathcal{M}$, written $\mathcal{M}_n \to \mathcal{M}$, such that this convergence implies convergence of the corresponding optimal value functions and optimal policies. (This will be done, similarly, for game models $\{\mathcal{G}_n\}_{n\geq 1}$ and $\mathcal{G}$.) To some extent, such results can be viewed as a sort of "continuity results" for control models, ensuring that the functions that map the control model into its optimal value function and its optimal policy, say $\mathbb{V}$ and $\mathbb{P}$, respectively, are continuous, meaning that $\mathcal{M}_n \to \mathcal{M}$ implies

$$\mathbb{V}(\mathcal{M}_n) \to \mathbb{V}(\mathcal{M}) \quad \text{and} \quad \mathbb{P}(\mathcal{M}_n) \to \mathbb{P}(\mathcal{M}).$$

In this sense, it is out of the scope of this thesis to define some kind of topology on the family of control models ensuring the continuity of $\mathbb{V}$ and $\mathbb{P}$. Our point of view in this thesis will be to start from a so-called original control model $\mathcal{M}$ and then study sequences of control models $\{\mathcal{M}_n\}$ converging to $\mathcal{M}$, ensuring convergence of optimal values and policies.

Concerning the application of this technique, one should think of the sequence $\{\mathcal{M}_n\}_{n\geq 1}$ converging to $\mathcal{M}$ as a sequence of simpler control models that, in principle, we are able to solve explicitly. Then, the above mentioned convergences $\mathbb{V}(\mathcal{M}_n) \to \mathbb{V}(\mathcal{M})$ and $\mathbb{P}(\mathcal{M}_n) \to \mathbb{P}(\mathcal{M})$ allow to obtain approximations of the optimal value function and the optimal policies by letting n tend to infinity. Therefore, one of our main objectives is to show how we can construct such sequence of approximating control models, starting from a given control model $\mathcal{M}$. Then, issues such as the convergence rates will be studied as well. All that has

been said for control models will be analyzed also for game models.

Now we describe briefly the control and game models we will be concerned with. We will consider continuous-time controlled Markov chains with denumerable state space, Borel action space and compact action sets. The transition and reward rates are continuous and they are allowed to be unbounded in state. We are interested in the discounted and average reward optimality criteria. For game models, we will deal with two-person zero-sum continuous-time Markov games. The underlying dynamical system is of the same nature as for the control model: countable state space, Borel action space, compact action sets, and continuous and possibly unbounded transition and reward rates.

Approximation results for control models are studied in Chapter 2, while such approximation results for game models are the purpose of Chapter 3. There is apparently a close parallelism between the results obtained in both chapters. In fact, although both chapters have the same motivation and the same structure, the techniques used in the proofs in Chapters 2 and 3 are quite different, due precisely to the fact that a control problem (with a "single player") and a two-person zero-sum game (two players with opposite goals) are of a different nature. The conclusions and some interesting open issues are mentioned in Chapter 4.

1.2 Motivation and state of the art

Control models. When solving a control problem by following the dynamic programming approach, one usually ends up with a so-called optimality equation (also known as the Bellman or the Hamilton-Jacobi-Bellman equation, depending on the nature of the control problem under study). Except for some particular cases (as, for instance, linear-quadratic control problems), such optimality equations cannot be explicitly solved because they are "highly" non-linear. Moreover, in the case of a countable state space, there is an infinite amount of such equations.

Concerning continuous-time controlled Markov chains, there exist also algorithms that are shown to converge to the optimal reward and policies of the control model. These include the value iteration algorithm —developed in [13, 17] for discounted reward controlled Markov chains— and the policy iteration algorithm —introduced in [14] for average reward controlled Markov chains. For the models we shall deal with (with countable state space and general action space), the value iteration and the policy iteration algorithms are not viable in practice because they require to perform a "denumerable" amount of calculations at each step and, in addition, a maximization over a "general" set. This shows the necessity for numerical methods to approximate the optimal solutions of controlled Markov chains.

In this same vein, and as can be seen in the references, in particular, [5, 35, 36], there are several approaches to show the existence of optimal policies, but it is not clear at all how to compute these policies and the corresponding optimal rewards. More precisely, in [5, Chapter 5] and [36], continuous-time controlled Markov chains with bounded reward and transition rates are analyzed. The uniformization technique (which reduces the continuous-time controlled Markov chain to a discrete-time one) is used. In our case, how-

ever, this approach is not possible because we consider unbounded reward and transition rates. Similarly, in [35, Chapter 11], an algorithm to determine the optimal policies and the optimal gain of a continuous-time controlled Markov chain is proposed for the finite state and action case.

One usual tool to obtain numerical solutions to the dynamic programming optimality equation is by means of the Markov chain approximating method. The idea is to define, starting from the original control model, a controlled Markov chain with finite state space whose optimal reward and policies approximate the optimal reward and policies of the original control model. Such methods have been developed to approximate, e.g., controlled diffusions [23, 38], discrete-time finite horizon and infinite horizon discounted controlled Markov chains [24, 40], average reward discrete-time controlled Markov chains [25], or discrete-time control models involving constraints [2], among others.

As in the Markov chain approximation scheme [23], this suggests the idea of considering finite-state and finite-action control models $\mathcal{M}_n$ whose optimal reward and policies we can explicitly compute (by using, for instance, the value or the policy iteration algorithms). Then, the optimal reward and policies of $\mathcal{M}_n$ are used as approximations of those of the original control model $\mathcal{M}$. Following this approach, we will introduce a finite state and action truncation technique to obtain the approximating control models $\mathcal{M}_n$. Similar discretization procedures can be found in, e.g., [2, 20]. The above mentioned more general framework of convergence of control models $\mathcal{M}_n \to \mathcal{M}$ has already been used in [24] for finite horizon and infinite horizon discounted discrete-time controlled Markov chains, and in [3, 39] for constrained discrete-time models.

It is also interesting to mention the reference [19], which proposes approximation techniques for discounted cost Markov decision processes with constraints. Their setting is similar to ours, in the sense that they propose a definition of convergence for control models. The technique proofs in [19] mainly rely on linear programming, while here use dynamic programming arguments.

Game models. We will deal with a two-person zero-sum continuous-time Markov game with denumerable state space, general action spaces, and possibly unbounded payoff and transition rates. The optimality criterion consists in finding a Nash equilibrium for the total expected discounted payoff, and for the long-run expected average payoff of the players. The existence of such Nash equilibria, as well as the existence of optimal strategies for the players, has been established in [15, 16]. In these references, it is shown that the value of the game is the solution of an optimality equation (also referred to as the Shapley equation).

Now we explain, somehow loosely, the form of this Shapley equation. Let $i \in S$ be the state of the system, and denote by $a \in A$ and $b \in B$ the actions of the players, that take values in some Borel spaces A and B. Let $\mathcal{P}(A)$ and $\mathcal{P}(B)$ be the family of probability measures on A and B, respectively. There is some operator $\mathbf{H}$ that maps, for each fixed $a \in A$ and $b \in B$, a function $\{u(i)\}_{i\in S}$ into the function $\{(\mathbf{H}u)(i,a,b)\}_{i\in S}$ such that the value of the game $\{V(i)\}_{i\in S}$ —either for the discounted or the average payoff criterion—

is the unique solution of the equations

$$V(i) = \sup_{\varphi\in\mathcal{P}(A)}\inf_{\psi\in\mathcal{P}(B)}\int_A\int_B(\mathbf{H}V)(i,a,b)\psi(db)\varphi(da) \tag{1.2.1}$$

$$= \inf_{\psi\in\mathcal{P}(B)}\sup_{\varphi\in\mathcal{P}(A)}\int_A\int_B(\mathbf{H}V)(i,a,b)\psi(db)\varphi(da) \tag{1.2.2}$$

for all $i \in S$. It should be clear that one cannot expect to solve, in general, the equations (1.2.1)–(1.2.2) explicitly. For computational purposes, therefore, one should use some kind of discretization technique to, at least, approximate the value of the game and the optimal strategies of the players. This is precisely the goal of this chapter.

Recall that we will let $\mathcal{G}$ be the "original" game model and $\{\mathcal{G}_n\}_{n\geq 1}$ be a sequence of game models. We propose a definition of the convergence $\mathcal{G}_n \to \mathcal{G}$ which, under adequate conditions, implies that the value of the games $\mathcal{G}_n$ and the corresponding optimal strategies converge to the value and the optimal strategies of the game $\mathcal{G}$. Then, for computational purposes, we show how we can construct, starting from the game model $\mathcal{G}$, a sequence of game models $\{\mathcal{G}_n\}_{n\geq 1}$ with finite state and action spaces that converge to $\mathcal{G}$. Such finite models can be solved explicitly and, hence, we can provide computable approximations of the value of the game model $\mathcal{G}$.

As far as we know, this is the first attempt to provide such computable approximations for continuous-time Markov games with denumerable state space and general action spaces. The reader interested in related works can consult [21, 28], in which the idea of approximating a game model $\mathcal{G}$ with "simpler" models has been studied. The reference [9] also considers computational issues for a continuous-time game with general state space and finite action spaces.

At this point, it is interesting to make a comparison between the approximation approaches for control and game models. Approximating a game model by means of finite state and actions game models is, from a technical point of view, more complicated than such approximations for control models. The analogous to (1.2.1)–(1.2.2) for a control model in which the state space is S and the action space of the controller is A, is the optimality (or dynamic programming) equation

$$V(i) = \sup_{a\in A}\{(\mathbf{H}V)(i,a)\} \quad \text{for } i \in S. \tag{1.2.3}$$

When making a finite approximation, one roughly considers an optimality equation as in (1.2.3) with finite S and A. Then, one can use, for instance, the policy iteration algorithm that solves this optimality equation in a finite number of steps. For a game model, however, the equations (1.2.1)–(1.2.2) are, even in the case of finite S, A, and B, of a continuous nature because we are optimizing on a set of probability measures (say, a simplex). This makes the computational problems less straightforward. Here, we combine linear programming with a "value iteration" or a "policy iteration" algorithm to solve such problems. Moreover, from a computational perspective, the maximum of a function (as in (1.2.3)) is easier to approximate than the saddle point of a function (as in (1.2.1)–(1.2.2)).

We shall address these issues for two-person continuous-time Markov games under both the discounted and the average payoff optimality criteria.

1.3 Contribution

The basic results on control models, namely, the existence of optimal policies and characterization of the optimal value as the solution of the Bellman equation, are already known results; these are given in Section 2.1. The convergence results for discounted game models in Section 2.2 are borrowed from [30], but we present them here for completeness. The results of the convergence rates for discounted models in Section 2.3.2 are, however, original. The analysis of convergence of control models under the average reward optimality criterion in Sections 2.2 and 2.3 is also an original contribution, and it is mainly drawn from [33].

Concerning the basic results for game models, these are already known facts, and they are given in Section 3.1. The rest of the material in this chapter (Sections 3.2, 3.3, and 3.4) is an original contribution, and it is based on [26] for the average payoff criterion, and on [34] for the discounted payoff criterion.

1.4 Notation and preliminary results

We define some notation that will be used throughout.

The real numbers set is denoted by $\mathbb{R}$. Given a topological space X, its Borel σ-algebra is the smallest σ-algebra containing its open sets. It will be denoted by $\mathbb{B}(X)$. In what follows, measurability issues (sets, functions, measures) will be always referred to the Borel σ-algebras. Given $D \subseteq X$, the indicator function of D is $\mathbf{I}_D$, with $\mathbf{I}_D(x) = 1$ if $x \in D$, and $\mathbf{I}_D(x) = 0$ if $x \notin D$. Sometimes it will be also written $\mathbf{I}\{x \in D\}$.

A Polish space is a complete and separable metric space. A Borel space is a measurable subset of a Polish space. Given a probability measure μ on some Borel space $(X, \mathbb{B}(X))$ and a real-valued measurable function f on X, the integral of f with respect to μ, provided that it is well defined, will be denoted

$$\mu(f) = \int_X f d\mu.$$

The Dirac probability measure supported on some point $x \in X$ is denoted by δ_x; that is, $\delta_x(B) = \mathbf{I}_B(x)$ for all $B \in \mathbb{B}(X)$. The constant function on X equal to 1 will be denoted by $\mathbf{1}$.

We will use the Landau notation O. As an illustration, given real-valued sequences $\{f(n)\}_{n\geq 1}$ and $\{g(n)\}_{n\geq 1}$, the latter being positive, we say that $f(n) = \mathrm{O}(g(n))$ as $n \to \infty$ when

$$\limsup_{n\to\infty} \frac{|f(n)|}{g(n)} < \infty.$$

The Kronecker delta is δ_{ij}, which equals 1 whenever $i = j$, and 0 otherwise. Given real numbers x and y, we will use the notation $x \vee y = \max\{x, y\}$. Finally, the symbol $:=$ refers to an equality by definition.

The Hausdorff distance. Suppose that (X, d_X) is a metric space. Given two nonempty subsets C and D of X we define

$$\rho_X(C, D) = \sup_{y \in C} \inf_{x \in D} \{d_X(x, y)\} \ \vee \ \sup_{x \in D} \inf_{y \in C} \{d_X(x, y)\}.$$

If C and D are closed sets, then $\rho_X(C, D)$ is referred to as the Hausdorff distance between C and D. The Hausdorff distance satisfies all the properties of a metric except that it might not be finite. For $\{C_n\}_{n \geq 1}$ and C closed subsets of X we say that $\{C_n\}_{n \geq 1}$ converges to C in the Hausdorff metric when $\rho_X(C_n, C) \to 0$ as $n \to \infty$.

Convergence of probability measures. Now we recall some facts on convergence of probability measures; see, e.g., [6, Chapter 1] or [7, Chapter 8]. Given a metric space (X, d_X), let $\mathcal{P}(X)$ be the family of probability measures on $(X, \mathbb{B}(X))$. We say that the sequence $\{\mu_n\} \subseteq \mathcal{P}(X)$ converges weakly to $\mu \in \mathcal{P}(X)$, and we will write $\mu_n \xrightarrow{d} \mu$, if

$$\lim_{n \to \infty} \mu_n(f) = \mu(f) \tag{1.4.4}$$

for all bounded and continuous functions $f : X \to \mathbb{R}$. We will use the following definition.

Definition 1.4.1 *We say that the function $f : X \to \mathbb{R}$ is Lipschitz continuous if there exists a constant $L \geq 0$ such that $|f(x) - f(y)| \leq L \cdot d_X(x, y)$ for all $x, y \in X$. In this case, we say that f is L-Lipschitz continuous. Let $\mathrm{Lip}_1(X)$ be the set of all 1-Lipschitz continuous functions on X.*

As a consequence of the Portmanteau theorem (see Theorems 1.2 and 2.1 in [6]), to have weak convergence it suffices that (1.4.4) holds for all bounded and Lipschitz continuous functions $f : X \to \mathbb{R}$. (Although this is not the usual statement of the Portmanteau theorem, observe that the function constructed in [6, Theorem 1.2] is bounded and Lipschitz continuous, and then proceed as in the proof of [6, Theorem 2.1]. Another reference for this result is [7, Remark 8.3.1].)

In case that X is a compact metric space, we have that weak convergence is metrizable with the Wasserstein distance

$$d_W(\mu, \nu) = \sup_{f \in \mathrm{Lip}_1(X)} \big\{\mu(f) - \nu(f)\big\} = \inf_{\lambda} \int_{X \times X} d_X(x, x') \lambda(dx, dx'), \tag{1.4.5}$$

for $\mu, \nu \in \mathcal{P}(X)$, where the infimum ranges over the set of all probability measures λ on $X \times X$ with marginals μ and ν (see Theorems 8.3.2 and 8.10.45, and Section 8.10(viii) in [7]). With this metric, we have that $(\mathcal{P}(X), d_W)$ is a compact metric space [7, Theorem

8.9.3(i)]. In addition, if $\{x_1, x_2, \ldots\}$ is a countable dense subset of X, then the countable family of probability measures

$$\sum_{j=1}^{k} \beta_j \delta_{x_j}$$

for all $k \geq 1$ and rational $\beta_1, \ldots, \beta_k \geq 0$ with $\sum \beta_j = 1$ is dense in $(\mathcal{P}(X), d_W)$; see [7, Theorem 8.9.4(ii)] or [8].

Chapter 2

Approximation of control models

We give an overview of this chapter. In Section 2.1 we introduce the control model we will be dealing with. In particular, we give the basic results on the existence of the controlled Markov chain model, and on the discounted and average reward optimality criteria. This section is mainly based on [17, 18, 31].

Section 2.2 gives the definition of convergence of control models and establishes the first theoretical convergence results. In Section 2.3 we present finite state and action truncations of the original control model. Convergence is studied and convergence rates are also analyzed. These sections are based on [26, 34].

Finally, we give some numerical applications for a controlled population system and a controlled birth-and-death system in Section 2.4.

2.1 Basic results

In this section we give the definition of the control model $\mathcal{M}$ and recall some basic results on the existence of the controlled process, and on the discounted and average reward optimality criteria. The results in this section are mainly drawn from [17, 31].

2.1.1 The control model $\mathcal{M}$

We define the control model we will be dealing with. Let

$$\mathcal{M} := \{S, A, \mathbb{K}, q, r\},$$

which consists of the following elements:

- The state space of the system is the denumerable set S. We suppose that $S = \{0, 1, 2, \ldots\}$ is the set of nonnegative integers.
- The action space of the controller is A, assumed to be a Borel space, that is, a measurable subset of a complete and separable metric space. Here, measurability is always referred to the corresponding Borel σ-algebra.

- The action set in state $i \in S$ is $A(i)$, which is a nonempty measurable subset of A. The family of feasible state-action pairs is defined as

$$\mathbb{K} := \{(i, a) \in S \times A : a \in A(i)\}.$$

- The transition rates of the system are given by $q = \{q_{ij}(a)\}$. We interpret $q_{ij}(a)$ as the transition rate from the state $i \in S$ to the state $j \in S$ under the action $a \in A(i)$. We assume that $a \mapsto q_{ij}(a)$ is a measurable function on $A(i)$ for each fixed $i, j \in S$. The transition rates verify that $q_{ij}(a) \geq 0$ for every $(i, a) \in \mathbb{K}$ and $j \neq i$. Finally, we suppose that the transition rates are conservative, i.e.,

$$\sum_{j \in S} q_{ij}(a) = 0 \quad \text{for all } (i, a) \in \mathbb{K},$$

  and stable, i.e.,

$$q(i) := \sup_{a \in A(i)} \{-q_{ii}(a)\} < \infty \quad \text{for all } i \in S.$$

- The reward rate function is $r : \mathbb{K} \to \mathbb{R}$. It is assumed that $a \mapsto r(i, a)$ is measurable on $A(i)$ for each $i \in S$.

This continuous-time controlled Markov chain model can also be found in, e.g., [14, 16, 33].

The dynamics of the control model can be roughly described as follows. Suppose that the system is in state $i \in S$ at some time $t \geq 0$. The controller takes an action $a \in A(i)$ and then, on the small time interval $[t, t + dt]$, the following happens:

- the controller receives an infinitesimal reward $r(i, a)dt$, and
- the system remains in state $i \in S$ with probability $1 + q_{ii}(a)dt$ or makes a transition to the state $j \neq i$ with probability $q_{ij}(a)dt$.

This procedure is carried on over all the time horizon $t \in [0, \infty)$.

Control policies. Now we describe the control policies available to the decision-maker. Let Φ be the family of functions

$$\varphi \ \equiv \ \{\varphi_t(B|i) \ : \ t \geq 0, \ i \in S, \ B \in \mathbb{B}(A(i))\}$$

that verify the following properties:

(i) The mapping $B \mapsto \varphi_t(B|i)$ is a probability measure on $(A(i), \mathbb{B}(A(i)))$ for each $t \geq 0$ and $i \in S$;

(ii) The function $t \mapsto \varphi_t(B|i)$ is measurable on $[0, \infty)$ for every $i \in S$ and $B \in \mathbb{B}(A(i))$.

We say that $\varphi \in \Phi$ is a randomized Markov policy or a Markov policy, for short. Such policies are also sometimes referred to as relaxed controls. The interpretation is, loosely, that when the state of the system is $i \in S$ at time $t \geq 0$, the actions taken by the controller are randomized according to the probability distribution $\varphi_t(\cdot|i)$.

If the Markov policy $\varphi \in \Phi$ is such that $\varphi_t(B|i)$ does not depend on $t \geq 0$ then we say that $\varphi = \{\varphi(B|i)\}$ is a (randomized) stationary policy. The class of such policies is denoted by Φ^s. This means that the controller follows the same probability rule at every time $t \geq 0$.

Finally, if the stationary policy φ is such that the probability measure $\varphi(B|i)$ is a Dirac measure, then we say that φ is a deterministic stationary policy. It should be clear that the class of deterministic stationary policies can be identified with the family of functions $f : S \to A$ with $f(i) \in A(i)$ for all $i \in S$, by letting $\varphi(\cdot|i) = \delta_{f(i)}(\cdot)$. The set of such functions will be denoted by $\mathbb{F}$. Clearly, we have the following inclusions: $\mathbb{F} \subseteq \Phi^s \subseteq \Phi$.

We introduce some notation. For each Markov policy $\varphi \in \Phi$ we define the corresponding transition rates as

$$q_{ij}(t, \varphi) := \int_{A(i)} q_{ij}(a)\varphi_t(da|i) \quad \text{for all } i, j \in S \text{ and } t \geq 0, \tag{2.1.1}$$

which is just the average transition rate from i to j at time t when using the control policy φ. The so-defined transition rates are finite because the $q_{ij}(a)$ are conservative and stable. In particular, $|q_{ij}(t, \varphi)| \leq q(i)$ for all $i, j \in S$ and $t \geq 0$. The corresponding reward rates are

$$r(t, i, \varphi) = \int_{A(i)} r(i, a)\varphi_t(da|i) \quad \text{for all } i \in S \text{ and } t \geq 0,$$

which are given a similar interpretation. Later, we will give conditions ensuring that these reward rates are well defined and finite.

In the particular case when $f \in \mathbb{F}$ is a deterministic stationary policy, we will write

$$q_{ij}(f) = q_{ij}(f(i)) \quad \text{and} \quad r(i, f) = r(i, f(i))$$

for $i, j \in S$.

The controlled process. We recall that a family of nonnegative real-valued functions $P_{ij}(s, t)$, for $0 \leq s \leq t$ and $i, j \in S$, is a (nonhomogeneous) transition function when the following conditions hold:

- $P_{ij}(s, s) = \delta_{ij}$ (the Kronecker delta) for all $i, j \in S$ and $s \geq 0$.
- $\sum_{j \in S} P_{ij}(s, t) \leq 1$ for all $i \in S$ and $0 \leq s \leq t$.
- The Chapman-Kolmogorov equation holds:

$$\sum_{k \in S} P_{ik}(s, z)P_{kj}(z, t) = P_{ij}(s, t)$$

for all $i, j \in S$ and $s \leq z \leq t$.

- In addition, the transition function is said to be regular when $\sum_{j\in S} P_{ij}(s,t) = 1$ for all $i \in S$ and $0 \le s \le t$.

Given some initial state in S at time 0, a regular transition function allows the construction of a probability measure on $S^{[0,\infty)}$, with the product σ-algebra, with conditional distributions given by the transition function itself.

For each Markov policy $\varphi \in \Phi$, consider the family of matrices $[q_{ij}(t,\varphi)]_{i,j}$, for $t \ge 0$, which is a nonhomogeneous Q^φ-matrix. By Proposition C.4 in [17, Appendix C], there exists a nonhomogeneous transition function

$$P_{ij}^{\varphi}(s,t) \quad \text{for } i,j \in S \text{ and } t \ge s \ge 0$$

whose transition rates are given by (2.1.1), that is,

$$\lim_{h\downarrow 0} \frac{P_{ij}^{\varphi}(t,t+h) - \delta_{ij}}{h} = q_{ij}(t,\varphi) \quad \text{for all } t \ge 0 \text{ and } i,j \in S. \tag{2.1.2}$$

To ensure that this transition function is unique and regular we impose the Assumption 2.1.2 below, which uses the notion of a Lyapunov function, defined next.

Definition 2.1.1 *(a) We say that $w : S \to [1,\infty)$ is a Lyapunov function on S when w is monotone nondecreasing and, in addition, $\lim_{i\to\infty} w(i) = +\infty$.*

(b) Let $\mathcal{B}_w(S)$ denote the family of functions $u : S \to \mathbb{R}$ such that

$$||u||_w = \sup_{i\in S}\{|u(i)|/w(i)\} < \infty.$$

We have that $||\cdot||_w$ is a norm on $\mathcal{B}_w(S)$, under which it is a Banach space.

Now we are ready to state our first assumption on the control model $\mathcal{M}$.

Assumption 2.1.2 *There exist a Lyapunov function w on S, and constants $c_1 \in \mathbb{R}$ and $b_1 \ge 0$ such that*

$$\sum_{j\in S} q_{ij}(a)w(j) \le -c_1 w(i) + b_1 \quad \textit{for all } (i,a) \in \mathbb{K}.$$

In addition, for each $i \in S$ we have $q(i) \le w(i)$.

We will usually refer to an equality such as $\sum q_{ij}(a)w(j) \le -c_1 w(i) + b_1$ as to a Lyapunov condition on the function w. Under this assumption, we have the following existence theorem. We omit its proof and the interested reader is referred to [17, Theorem 2.3].

Theorem 2.1.3 *Suppose that the control model $\mathcal{M}$ satisfies Assumption 2.1.2.*

(i) For every Markov policy $\varphi \in \Phi$ there exists a unique regular transition function

$$\{P_{ij}^{\varphi}(s,t)\}_{i,j\in S, 0\le s\le t}$$

with transition rates given by the $q_{ij}(t,\varphi)$; recall (2.1.2).

Let $\Omega = \mathbb{K}^{[0,\infty)} = \{(x(t), a(t))\}_{t\ge 0}$ be endowed with the product σ-algebra $\mathcal{F}$.

(ii) Given an initial state $i \in S$ at time 0 and a Markov policy $\varphi \in \Phi$, there exists a unique probability measure $P^{i,\varphi}$ on $(\Omega, \mathcal{F})$ that satisfies the following properties:

- *For every $A_0 \in \mathbb{B}(A(i))$ we have $P^{i,\varphi}\{x(0) = i, a(0) \in A_0\} = \varphi_0(A_0|i)$.*
- *For any $n \ge 1$, given $0 \le s_1 < s_2 < \ldots < s_n$ and, on the other hand, $i_k \in S$ and $A_k \in \mathbb{B}(A(i_k))$ for $k = 1, \ldots, n$, we have*

$$P^{i,\varphi}\{x(s_1) = i_1, a(s_1) \in A_1, \ldots, x(s_n) = i_n, a(s_n) \in A_n\} = \prod_{k=1}^{n} P_{i_{k-1}i_k}^{\varphi}(s_{k-1}, s_k)\varphi_{s_k}(A_k|i_k),$$

where we make the convention that $i_0 = i$ and $s_0 = 0$.

The corresponding expectation operator will be denoted by $E^{i,\varphi}$.

The above theorem ensures the existence of the controlled Markov chain model itself. Assumption 2.1.2 is used to ensure regularity and uniqueness of the transition function. In particular, the process $\{x(t)\}_{t\ge 0}$ is nonexplosive under any Markov policy $\varphi \in \Phi$. Assumption 2.1.2 ensures, as well, that the (non homogeneous) backward and forward Kolmogorov differential equations hold.

We have the following bound on the expected growth of $w(x(t))$. As a consequence of Assumption 2.1.2 and [17, Lemma 6.3], for every initial state $i \in S$ and every Markov policy $\varphi \in \Phi$

$$E^{i,\varphi}[w(x(t))] \le e^{-c_1 t} w(i) + \frac{b_1}{c_1}(1 - e^{-c_1 t}) \quad \text{for all } t \ge 0. \tag{2.1.3}$$

When $c_1 = 0$, the above inequality reads $E^{i,\varphi}[w(x(t))] \le w(i) + b_1 t$.

2.1.2 The discounted reward optimality criterion

Let us now focus on the total expected discounted reward optimality criterion. We suppose that the rewards earned by the controller are depreciated at a constant discount rate $\alpha > 0$.

Assumption 2.1.4 *The control model $\mathcal{M}$ satisfies the following conditions.*

(i) The discount rate $\alpha > 0$ is such that $\alpha + c_1 > 0$, where $c_1 \in \mathbb{R}$ is the constant in Assumption 2.1.2.

(ii) There exists a constant $M > 0$ such that $|r(i,a)| \leq Mw(i)$ for all $(i,a) \in \mathbb{K}$.

The total expected discounted reward (or, in short, the discounted reward) of the Markov policy $\varphi \in \Phi$ when $i \in S$ is the initial state is defined as

$$V^{\alpha}(i,\varphi) := E^{i,\varphi}\left[\int_0^{\infty} e^{-\alpha t} r(x(t), a(t))dt\right] = E^{i,\varphi}\left[\int_0^{\infty} e^{-\alpha t} r(t, x(t), \varphi)dt\right].$$

Under Assumptions 2.1.2 and 2.1.4, and recalling the inequality (2.1.3), we have that the discounted reward verifies

$$|V^{\alpha}(i,\varphi)| \leq \frac{Mw(i)}{\alpha + c_1} + \frac{b_1 M}{\alpha(\alpha + c_1)} \quad \text{for all } i \in S \text{ and } \varphi \in \Phi;$$

in particular, the fact that $\alpha + c_1 > 0$ is used to ensure that the integral of the exponential function is finite. Therefore, the optimal discounted reward, defined as

$$V^{\alpha}(i) := \sup_{\varphi \in \Phi} V^{\alpha}(i,\varphi) \quad \text{for all } i \in S$$

is finite. We deduce also that $V^{\alpha}(\cdot,\varphi)$ and V^{α} are in $\mathcal{B}_w(S)$ and, by letting $\mathfrak{M} := \frac{M(b_1+\alpha)}{\alpha(c_1+\alpha)}$, we obtain

$$||V^{\alpha}(\cdot,\varphi)||_w \leq \mathfrak{M} \quad \text{for all } \varphi \in \Phi, \quad \text{and} \quad ||V^{\alpha}||_w \leq \mathfrak{M}. \tag{2.1.4}$$

Finally, we say that a Markov policy $\varphi \in \Phi$ is discount optimal if it satisfies

$$V^{\alpha}(i,\varphi) = V^{\alpha}(i) \quad \text{for all } i \in S.$$

In order to characterize the optimal discounted reward as the solution of a dynamic programming optimality equation, we need to introduce further assumptions.

Assumption 2.1.5 *The control model $\mathcal{M}$ verifies the following conditions.*

(i) The action sets $A(i)$ are compact for every $i \in S$.

(ii) The functions $a \mapsto q_{ij}(a)$ and $a \mapsto r(i,a)$ are continuous on $A(i)$ for all $i, j \in S$.

(iii) There are constants $c_2 \in \mathbb{R}$ and $b_2 \geq 0$ with

$$\sum_{j \in S} q_{ij}(a) w^2(j) \leq -c_2 w^2(i) + b_2 \quad \textit{for all } (i,a) \in \mathbb{K}.$$

The conditions (i) and (ii) in Assumption 2.1.5 above impose the usual compactness-continuity requirements, while part (iii), which is just a Lyapunov condition on the function w^2, is used to ensure the use of Dynkin's formula.

Our next result will be useful in the sequel.

Corollary 2.1.6 *Under Assumptions 2.1.5(ii)–(iii), for every $u \in \mathcal{B}_w(S)$ the function $a \mapsto \sum_{j\in S} q_{ij}(a)u(j)$ is continuous on $A(i)$ for each $i \in S$.*

Proof. Fix $i \in S$ and let $k > i$. Note that, for each $a \in A(i)$,

$$\Big|\sum_{j=k}^{\infty} q_{ij}(a)u(j)\Big| \leq ||u||_w \sum_{j=k}^{\infty} q_{ij}(a)w(j).$$

On the other hand, by the monotonicity of the Lyapunov function w we have

$$\sum_{j=k}^{\infty} q_{ij}(a)w(j) \leq \frac{1}{w(k)} \sum_{j=k}^{\infty} q_{ij}(a)w^2(j).$$

Since $k > i$,

$$\sum_{j=k}^{\infty} q_{ij}(a)w^2(j) \leq \sum_{j\neq i} q_{ij}(a)w^2(j) = \sum_{j\in S} q_{ij}(a)w^2(j) - q_{ii}(a)w^2(i),$$

and so, by Assumption 2.1.5(iii)

$$\sum_{j=k}^{\infty} q_{ij}(a)w^2(j) \leq -c_2 w^2(i) + b_2 + q(i)w^2(i).$$

Summarizing, for all $a \in A(i)$,

$$\Big|\sum_{j=k}^{\infty} q_{ij}(a)u(j)\Big| \leq \frac{||u||_w}{w(k)}\Big(- c_2 w^2(i) + b_2 + q(i)w^2(i)\Big).$$

Therefore,

$$\lim_{k\to\infty} \sup_{a\in A(i)} \Big|\sum_{j=k}^{\infty} q_{ij}(a)u(j)\Big| = 0$$

and so the series $\sum_{j\in S} q_{ij}(a)u(j)$ of continuous functions converges uniformly and it is therefore itself continuous. □

Our next result summarizes the main results on the dynamic programming optimality equation for $\mathcal{M}$ and the existence of discount optimal policies.

Theorem 2.1.7 *Let the control model $\mathcal{M}$ satisfy the Assumptions 2.1.2, 2.1.4, and 2.1.5.*

(i) The optimal discounted reward V^α is the unique solution u in $\mathcal{B}_w(S)$ of the discounted reward optimality equation

$$\alpha u(i) = \max_{a\in A(i)} \big\{r(i,a) + \sum_{j\in S} q_{ij}(a)u(j)\big\} \quad \textit{for all } i \in S.$$

(ii) A deterministic stationary policy $f \in \mathbb{F}$ is discount optimal if and only if it attains the maximum in the discounted reward optimality equation, i.e.,

$$\begin{aligned} \alpha V^{\alpha}(i) &= \max_{a \in A(i)} \big\{ r(i,a) + \sum_{j \in S} q_{ij}(a) V^{\alpha}(j) \big\} \\ &= r(i,f) + \sum_{j \in S} q_{ij}(f) V^{\alpha}(j) \end{aligned}$$

for all $i \in S$, and such $f \in \mathbb{F}$ indeed exist.

The proof of Theorem 2.1.7 is made in [13, Theorem 3.2] and [17, Chapter 6] by using the value iteration algorithm. In [29, Theorem 1], however, Theorem 2.1.7 is established by showing the convergence of the policy iteration algorithm.

Notice that, as a consequence of Corollary 2.1.6, we can indeed take the max as $a \in A(i)$ in the discounted reward optimality equation, instead of taking the sup as $a \in A(i)$.

2.1.3 The average reward optimality criterion

We will suppose now that the controller is interested in maximizing his long-run expected average reward. To deal with this optimality criterion, some of the assumptions made so far on the control model $\mathcal{M}$ must be strengthened. First of all, the Lyapunov condition in Assumption 2.1.2 is replaced with the following drift condition.

Assumption 2.1.8 *There exist a Lyapunov function w on S, constants $c_1 > 0$ and $b_1 \geq 0$, and a finite set $D \subset S$ such that*

$$\sum_{j \in S} q_{ij}(a) w(j) \leq -c_1 w(i) + b_1 \mathbf{I}_D(i) \quad \textit{for all } (i,a) \in \mathbb{K}.$$

Moreover, for each $i \in S$ we have $q(i) \leq w(i)$.

It should be clear that Assumption 2.1.8 implies Assumption 2.1.2, and so Theorem 2.1.3 applies. In particular, the inequality (2.1.3) is also valid.

Our next assumption is similar to Assumption 2.1.4, except that part (i) has been suppressed.

Assumption 2.1.9 *There exists a constant $M > 0$ such that $|r(i,a)| \leq M w(i)$ for all $(i,a) \in \mathbb{K}$.*

Given a control policy $\varphi \in \Phi$ and an initial state $i \in S$, the long-run expected average reward (or average reward, for short) is defined as

$$J(i,\varphi) = \limsup_{T \to \infty} \frac{1}{T} E^{i,\varphi}\Big[\int_0^T r(x(t), a(t)) dt \Big] = \limsup_{T \to \infty} \frac{1}{T} E^{i,\varphi}\Big[\int_0^T r(t, x(t), \varphi) dt \Big].$$

Under Assumptions 2.1.8 and 2.1.9, and using (2.1.3), it is easily seen that the average reward is finite and bounded, with

$$|J(i,\varphi)| \leq \frac{Mb_1}{c_1} \quad \text{for all } i \in S \text{ and } \varphi \in \Phi. \tag{2.1.5}$$

The optimal expected average reward for the initial state $i \in S$ is then defined as

$$J(i) = \sup_{\varphi \in \Phi} J(i,\varphi) \quad \text{for } i \in S,$$

and it verifies as well $|J(i)| \leq Mb_1/c_1$ for all $i \in S$. We say that a Markov policy is average reward optimal when $J(i,\varphi) = J(i)$ for every initial state $i \in S$.

Our next assumption is an extension of Assumption 2.1.5, used for the discounted reward optimality criterion. It uses the following terminology. We say that a deterministic stationary policy $f \in \mathbb{F}$ is irreducible when the controlled process $\{x(t)\}_{t\geq 0}$, under the policy $f \in \mathbb{F}$, can travel with positive probability between any two states. In terms of transition rates this is equivalently stated as follows.

Definition 2.1.10 *The deterministic stationary policy $f \in \mathbb{F}$ is irreducible when, given arbitrary distinct states $i, j \in S$, there exist states $i = i_0, i_1, \ldots, i_n = j$ with $q_{i_{k-1}i_k}(f) > 0$ for all $k = 1, \ldots, n$.*

Note that items (i)–(iii) in our next assumption are the same as in Assumption 2.1.5; for ease of reference, however, we prefer to state them again.

Assumption 2.1.11 *The control model $\mathcal{M}$ verifies the following conditions.*

(i) The action sets $A(i)$ are compact for every $i \in S$.

(ii) The functions $a \mapsto q_{ij}(a)$ and $a \mapsto r(i,a)$ are continuous on $A(i)$ for all $i, j \in S$.

(iii) There are constants $c_2 \in \mathbb{R}$ and $b_2 \geq 0$ with

$$\sum_{j \in S} q_{ij}(a) w^2(j) \leq -c_2 w^2(i) + b_2 \quad \textit{for all } (i,a) \in \mathbb{K}.$$

(iv) Every deterministic stationary policy $f \in \mathbb{F}$ is irreducible.

Under Assumptions 2.1.8 and 2.1.11(iv), we have that for each deterministic stationary policy $f \in \mathbb{F}$, the Markov chain $\{x(t)\}_{t\geq 0}$ has a unique invariant probability measure on S, that will be denoted by μ_f. The probabilities $\mu_f(i)$, for $i \in S$, are characterized as the unique nonnegative solutions x_i of the linear equations

$$\sum_{i \in S} x_i q_{ij}(f) = 0 \quad \text{for all } j \in S$$

such that $\sum_{i\in S} x_i = 1$. In addition, the invariant probabilities satisfy $\mu_f(i) > 0$ for all $i \in S$ and, moreover, we have $\mu_f(w) = \sum_{i\in S} \mu_f(i)w(i) < \infty$. These results can be found in [31, Theorem 2.5]. It also follows that the expected average reward of a policy $f \in \mathbb{F}$ is constant (that is, it does not depend on the initial state of the system), with

$$J(i, f) = \sum_{j\in S} r(j, f)\mu_f(j) =: g(f) \quad \text{for all } i \in S,$$

where the constant $g(f) \in \mathbb{R}$ is usually referred to as the gain of $f \in \mathbb{F}$.

Exponential ergodicity. An important consequence of the above assumptions is the so-called uniform exponential ergodicity property. More precisely, under Assumptions 2.1.8 and 2.1.11, the control model $\mathcal{M}$ is uniformly exponentially ergodic on $\mathbb{F}$, meaning that there exist constants $R > 0$ and $\gamma > 0$ such that

$$\sup_{f\in\mathbb{F}} \left|E^{i,f}[u(x(t))] - \mu_f(u)\right| \leq Re^{-\gamma t}||u||_w w(i) \tag{2.1.6}$$

for all $u \in \mathcal{B}_w(S)$, $i \in S$, and $t \geq 0$. For a proof, see [30, Theorem 2.11] or [32]. This means that the expected value of $u(x(t))$, under the policy $f \in \mathbb{F}$, approaches its limiting average value $\mu_f(u)$ at an exponential speed, in the w-norm. Moreover, the constants in the exponential decay are uniform in $f \in \mathbb{F}$.

Additionally, under Assumption 2.1.9, given a deterministic stationary policy $f \in \mathbb{F}$ and an initial state $i \in S$, we define the bias of f at i as

$$h_f(i) = \int_0^\infty \left[E^{i,f}[r(x(t), f)] - g(f)\right]dt.$$

As a direct consequence of (2.1.6) we obtain that the bias h_f is in $\mathcal{B}_w(S)$ with

$$||h_f||_w \leq \frac{RM}{\gamma}, \tag{2.1.7}$$

and note that the bound on the w-norm of h_f is uniform in $f \in \mathbb{F}$. Moreover, the expectation of the bias with respect to the invariant probability measure is zero: $\mu_f(h_f) = 0$.

It is not possible, generally speaking, to derive an explicit expression (depending directly on the elements of the control model $\mathcal{M}$) for the constants R and γ in (2.1.6). A particular case is known, however, for which such explicit expressions are indeed available; see [14, 27] or [30, Theorem 2.8].

Remark 2.1.12 *Suppose that the control model $\mathcal{M}$ satisfies the Assumptions 2.1.8 and 2.1.11, with the following additional features.*

(a) The set D in Assumption 2.1.8 is $D = \{0\}$.

(b) For any $f \in \mathbb{F}$, the state process $\{x(t)\}$ is stochastically ordered in its initial value, meaning that

$$\sum_{j=k}^{\infty} q_{ij}(f) \leq \sum_{j=k}^{\infty} q_{i+1,j}(f)$$

for all $i, k \in S$ with $k \neq i+1$.

(c) For each $f \in \mathbb{F}$ and every $0 < i < j$, the process $\{x(t)\}_{t\geq 0}$ can travel with positive probability from i to $\{j, j+1, \ldots\}$ without passing through 0. Equivalently, there exist nonzero states $i = i_0, i_1, \ldots, i_n$, with $i_n \geq j$, such that $q_{i_{k-1}i_k}(f) > 0$ for all $k = 1, \ldots, n$.

Under these additional conditions, the constants in (2.1.6) are

$$R = 2(1 + b_1/c_1) \quad \textit{and} \quad \gamma = c_1.$$

The condition (b) means, roughly, that the total transition rate to the states in the set $\{k, k+1, \ldots\}$ is an increasing function of the initial state. This is not a restrictive requirement since for, e.g., a population system in which the state space models the size of the population, it seems quite natural that visiting the states in $\{k, k+1, \ldots\}$ becomes more likely as the initial state of the system is itself larger. Similarly, the condition (c) is not restrictive, as long as the Markov chain has a sufficiently "rich" communication structure. So, in practice, the more restrictive condition in Remark 2.1.12 is (a).

The optimality equation. Next, we address the characterization of the optimal average reward $J(i)$ as a solution of an optimality equation. We say that the pair $(g, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ is a solution to the average reward optimality equation for the control model $\mathcal{M}$ if

$$g = \max_{a \in A(i)} \{r(i,a) + \sum_{j \in S} q_{ij}(a)h(j)\} \quad \text{for all } i \in S.$$

This is the main theorem on the average reward optimality criterion.

Theorem 2.1.13 *Let the control model $\mathcal{M}$ satisfy Assumptions 2.1.8, 2.1.9, and 2.1.11.*

(i) The optimal average reward $J(i)$ is constant and we will write $g^* = J(i)$ for all $i \in S$.

(ii) There exist solutions $(g, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ to the average reward optimality equation for $\mathcal{M}$.

If $(g, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ is a solution to the average reward optimality equation then $g = g^*$, the optimal average reward, and h is unique up to additive constants.

(iii) A deterministic stationary policy $f \in \mathbb{F}$ is average optimal if and only if it attains the maximum in the average reward optimality equation, that is,

$$\begin{aligned} g^* &= \max_{a \in A(i)} \{r(i,a) + \sum_{j \in S} q_{ij}(a)h(j)\} \\ &= r(i,f) + \sum_{j \in S} q_{ij}(f)h(j) \end{aligned}$$

for all $i \in S$, and such $f \in \mathbb{F}$ indeed exist.

In Theorem 2.1.13(ii), the statement on h means that if $(g^*, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ and $(g^*, h') \in \mathbb{R} \times \mathcal{B}_w(S)$ are solutions to the average reward optimality equation, then the function $h - h'$ is constant on S. In particular, the family of $f \in \mathbb{F}$ that attain the maximum in the optimality equation (as in part (iii)) does not depend on the particular solution h. We will usually refer to $g^* \in \mathbb{R}$ as to the optimal gain of the control model $\mathcal{M}$.

The Poisson equation. We conclude this section by recalling some results that will be needed in the sequel. Given a deterministic stationary policy $f \in \mathbb{F}$, we say that the pair $(g, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ is a solution to the Poisson equation for f if

$$g = r(i, f) + \sum_{j \in S} q_{ij}(f) h(j) \quad \text{for all } i \in S.$$

The next result characterizes such solutions. For a proof, see Proposition 3.14 in [31].

Proposition 2.1.14 *Suppose that the control model $\mathcal{M}$ satisfies Assumptions 2.1.8, 2.1.9, and 2.1.11. Given any $f \in \mathbb{F}$, the solutions of the Poisson equation for $f \in \mathbb{F}$ are of the form*

$$(g(f), h_f + \lambda \mathbf{1}) \quad \textit{for all } \lambda \in \mathbb{R}.$$

It follows that the pair given by the gain $g(f)$ and the bias h_f of the policy $f \in \mathbb{F}$ is the unique solution $(g(f), h)$ of the Poisson equation for f such that $\mu_f(h) = 0$.

Moreover, if $(g^*, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ is a solution to the average reward optimality equation for $\mathcal{M}$, and $f \in \mathbb{F}$ is an average reward optimal policy, it follows that (g^*, h) is a solution to the Poisson equation for f. Therefore, the function h in the average reward optimality equation can be chosen to be the bias of an optimal policy, which therefore satisfies the bound (2.1.7). This is summarized next.

Corollary 2.1.15 *Suppose that the control model $\mathcal{M}$ satisfies Assumptions 2.1.8, 2.1.9, and 2.1.11, and let R and γ be the constants for the uniform exponential ergodicity of $\mathcal{M}$; recall (2.1.6). There exists a solution $(g^*, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ to the average reward optimality equation for $\mathcal{M}$ with $||h||_w \leq RM/\gamma$.*

We propose the following definition of convergence of policies.

Definition 2.1.16 *Given deterministic stationary policies $\{f_n\}_{n \geq 1} \subseteq \mathbb{F}$, we say that $\{f_n\}$ converges to $f \in \mathbb{F}$, and we will write $f_n \to f$, if $\lim_{n \to \infty} f_n(i) = f(i)$ for all $i \in S$.*

Given $\{f_n\}_{n \geq 1}$ and $f \in \mathbb{F}$, we say that f is a limit policy of $\{f_n\}_{n \geq 1}$ if there exists a subsequence n' such that $f_{n'} \to f$.

We note that if the action sets $A(i)$ are compact (Assumption 2.1.11(i)) then $\mathbb{F}$ is metrizable and compact with this definition of convergence. The corresponding metric is

$$d(f, f') = \sum_{i \in S} \frac{1}{2^i} \cdot \frac{d_A(f(i), f'(i))}{1 + d_A(f(i), f'(i))}$$

for $f, f' \in \mathbb{F}$. The proof of the next result can be found in [31, Theorem 3.17].

Lemma 2.1.17 *Let the control model $\mathcal{M}$ satisfy Assumptions 2.1.8, 2.1.9, and 2.1.11. If the sequence $\{f_n\}_{n\geq 1}$ converges to $f \in \mathbb{F}$ then $g(f_n) \to g(f)$.*

This lemma simply states that the gain function $f \mapsto g(f)$ is continuous on $\mathbb{F}$.

2.2 Convergence of control models

The previous section was devoted to analyze the control model $\mathcal{M}$. In this section we shall consider a sequence of control models $\{\mathcal{M}_n\}_{n\geq 1}$, that we will interpret as approximations of the control model $\mathcal{M}$. In the sequel, we will sometimes refer to $\mathcal{M}$ as to the "original" control model, whose optimal value and optimal policies we want to approximate, and to the $\mathcal{M}_n$ as to the approximating control models.

2.2.1 Definition

The control model $\mathcal{M}_n$, for each $n \geq 1$, is given by the following elements:

$$\mathcal{M}_n := \{S_n, A, \mathbb{K}_n, q_n, r_n\},$$

where:

- The state space S_n is a subset (either finite or infinite) of S, the state space of the original control model $\mathcal{M}$.
- The action space is the Borel space A, which is the same as for the control model $\mathcal{M}$.
- The action sets are $A_n(i)$ for $i \in S_n$. We assume that $A_n(i)$ is a nonempty measurable subset of $A(i)$. The family of feasible state-action pairs is

$$\mathbb{K}_n := \{(i,a) \in S \times A \ : \ i \in S_n, \ a \in A_n(i)\} \subseteq \mathbb{K}.$$

- The transition rates of $\mathcal{M}_n$ are given by $q_{ij}^n(a)$ for $i, j \in S_n$ and $a \in A_n(i)$. They are measurable in a and they verify $q_{ij}^n(a) \geq 0$ when $(i,a) \in \mathbb{K}_n$ and $j \neq i$, and they are also assumed to be conservative and stable, meaning that

$$\sum_{j \in S_n} q_{ij}^n(a) = 0 \quad \text{for all } (i,a) \in \mathbb{K}_n$$

  and

$$q_n(i) := \sup_{a \in A_n(i)} \{-q_{ii}^n(a)\} < \infty \quad \text{for all } i \in S_n.$$

- The reward rate function for the control model $\mathcal{M}_n$ is $r_n : \mathbb{K}_n \to \mathbb{R}$. We assume that $a \mapsto r_n(i,a)$ is measurable on $A_n(i)$ for all $i \in S_n$.

Therefore, the control models $\mathcal{M}_n$ are of the same nature as $\mathcal{M}$, with the particular feature that the state and action sets of $\mathcal{M}_n$ are subsets of the corresponding sets for $\mathcal{M}$.

Control policies. The family of Markov policies for $\mathcal{M}_n$, for $n \geq 1$, is denoted by Φ_n. Its definition is the same as for $\mathcal{M}$, but now accounting for the control model $\mathcal{M}_n$. Namely, Φ_n is the set of functions $\varphi = \{\varphi_t(B|i)\}$, for $t \geq 0$, $i \in S_n$, and $B \in \mathbb{B}(A_n(i))$, such that $B \mapsto \varphi_t(B|i)$ is a probability measure on $A_n(i)$ and such that $t \mapsto \varphi_t(B|i)$ is measurable.

The family of stationary (randomized) policies is Φ_n^s, while the set of deterministic stationary policies is identified with $\mathbb{F}_n$, the family of functions $f : S_n \to A$ with $f(i) \in A_n(i)$ for all $i \in S_n$.

The notation used for $\mathcal{M}_n$ is basically the same as for $\mathcal{M}$, by adding a subscript or superscript n where needed. For instance,

$$q_{ij}^n(t, \varphi) = \int_{A_n(i)} q_{ij}^n(a)\varphi_t(da|i) \quad \text{and} \quad r_n(t, i, \varphi) = \int_{A_n(i)} r_n(i, a)\varphi_t(da|i)$$

for $i, j \in S_n$, $t \geq 0$, and $\varphi \in \Phi_n$, while $q_{ij}^n(f) = q_{ij}^n(f(i))$ and $r_n(i, f) = r_n(i, f(i))$ for $f \in \mathbb{F}_n$.

The controlled process. To ensure the existence of the controlled process itself, for the control model $\mathcal{M}_n$, we shall impose some assumptions. Given $n \geq 1$, we say that $w : S_n \to [1, \infty)$ is a Lyapunov function when w is monotone nondecreasing and

$$\lim_{i \to \infty, i \in S_n} w(i) = +\infty$$

(in particular, the latter condition holds whenever S_n is finite). The w-norm of a function $u : S_n \to \mathbb{R}$ is then defined as

$$||u||_w = \sup_{i \in S_n} \{|u(i)|/w(i)\};$$

cf. Definition 2.1.1. Note that we use the same notation for the w-norm on S and on S_n. This will not lead to confusion. The family of functions $u : S_n \to \mathbb{R}$ with finite w-norm is denoted by $\mathcal{B}_w(S_n)$.

Under a condition similar to Assumption 2.1.2, but now for the control model $\mathcal{M}_n$, an analogous to Theorem 2.1.3 holds. In particular, for any initial state $i \in S_n$ and any control policy $\varphi \in \Phi_n$, there exists a unique probability measure $P_n^{i,\varphi}$ on the canonical space $\mathbb{K}_n^{[0,\infty)}$ that models the controlled process $\mathcal{M}_n$. Its expectation operator will be written $E_n^{i,\varphi}$.

Definition of convergence. After having described the notation for the sequence of approximating control models $\{\mathcal{M}_n\}_{n \geq 1}$, now we give the definition of convergence of $\{\mathcal{M}_n\}_{n \geq 1}$ to the original control model $\mathcal{M}$.

Definition 2.2.1 *Consider the control models $\mathcal{M}$ and $\{\mathcal{M}_n\}_{n \geq 1}$ defined above. We say that $\{\mathcal{M}_n\}_{n \geq 1}$ converges to $\mathcal{M}$ as $n \to \infty$, and we will write $\mathcal{M}_n \to \mathcal{M}$, when the following conditions are fulfilled.*

(a) The sequence of states $\{S_n\}_{n\geq 1}$ *is monotone nondecreasing and its limit is* S*. This means that*

$$S_1 \subseteq S_2 \subseteq S_3 \subseteq \dots \quad \text{with} \quad \bigcup_{n=1}^{\infty} S_n = S.$$

Define $n(i) = \min\{n \geq 1 : i \in S_n\}$ *for* $i \in S$*. Therefore,* $n \geq n(i)$ *if and only if* $i \in S_n$*.*

(b) For each $i \in S$ *we have the following convergence in the Hausdorff sense:*

$$\lim_{n\to\infty} \rho_A(A(i), A_n(i)) = \lim_{n\to\infty} \Big[\sup_{a\in A(i)} \inf_{a'\in A_n(i)} \{d_A(a, a')\} \Big] = 0.$$

Given $i \in S$*, if* $\{a_n\}_{n\geq n(i)}$ *is a sequence in* A *with* $a_n \in A_n(i)$ *for all* $n \geq n(i)$ *and such that, in addition,* $\lim_n a_n = a$ *for some* $a \in A(i)$*, then:*

(c) $\lim_{n\to\infty} q_{ij}^n(a_n) = q_{ij}(a)$ *for all* $j \in S$*;*

(d) $\lim_{n\to\infty} r_n(i, a_n) = r(i, a)$*.*

Let us make some comments on this definition. Note that given a state $i \in S$, we have that $n(i)$ is the first n such that the state i is in S_n. Observe that, in item (b), $\rho_A(A(i), A_n(i))$ is properly defined only for $n \geq n(i)$ but, since we are dealing with the limit as $n \to \infty$, this will not be explicit in the notation. Similarly, in (c), we require that $n \geq n(i) \vee n(j)$ but this is neither explicit in the notation.

Let us make some further comments on Definition 2.2.1. Note that, here, we allow all the elements of the control models $\mathcal{M}_n$ (namely, the state space, the action sets, and the transition and reward rates) to depend on $n \geq 1$.

When dealing with related definitions of convergence of control models, the state space is usually allowed to depend on n; see [2, 20, 33]. The transition and reward rates may as well depend on n. In this case, the "uniform convergence" property in Definition 2.2.1(c)–(d) is a usual requirement; see, for instance, the condition (2) in [2, Theorem 6.1], and Assumptions 3.1(c) and 3.3(c) in [3].

The notion of the Kuratowski convergence for the approximation of control models was used in [24]. In our context, imposing the Kuratowski convergence of $A_n(i)$ to $A(i)$ would consist in assuming that for each $i \in S$

$$\lim_{n\to\infty} \inf_{a'\in A_n(i)} \{d_A(a, a')\} = 0 \quad \text{for all } a \in A(i),$$

which is weaker than the requirement in Definition 2.2.1(b). In our context, however, since we will assume later that $A(i)$ is compact, Hausdorff and Kuratowski convergences will be, in this case, equivalent.

Let us also mention that the Kuratowski convergence of the actions sets $A_n(i)$ is related to the discretization of the state space made in [20, Section 6.3] for a discrete-time Markov control process. We note however that, in the references [2, 3, 20, 33], the actions sets of $\mathcal{M}_n$ are the same as the action sets of the original control model $\mathcal{M}$.

2.2.2 The discounted reward criterion

Consider the original control model $\mathcal{M}$ and the sequence of control models $\{\mathcal{M}_n\}_{n\geq 1}$ defined previously. We deal now with the total expected discounted reward optimality criterion and we let $\alpha > 0$ be the discount rate (the same for all the control models). In Section 2.1.2 we gave conditions ensuring that the discounted reward problem for $\mathcal{M}$ is well posed, and under which the discounted reward optimality equation for $\mathcal{M}$ holds.

Assumptions. Our next assumption states the conditions we will impose on the sequence of control models $\{\mathcal{M}_n\}_{n\geq 1}$. We suppose that the original control model $\mathcal{M}$ satisfies the Assumptions 2.1.2, 2.1.4, and 2.1.5.

Assumption 2.2.2 *Let w be the Lyapunov function in Assumption 2.1.2. The following conditions hold for every $n \geq 1$.*

(i) With the constants $c_1 > -\alpha$ and $b_1 \geq 0$ as in Assumption 2.1.2, we have

$$\sum_{j\in S_n} q_{ij}^n(a)w(j) \leq -c_1 w(i) + b_1 \quad \text{for all } (i,a) \in \mathbb{K}_n,$$

with $q_n(i) \leq w(i)$ for each $i \in S_n$.

(ii) With the constant $M > 0$ taken from Assumption 2.1.4(ii), we have

$$|r_n(i,a)| \leq Mw(i) \quad \text{for all } (i,a) \in \mathbb{K}_n.$$

(iii) The action sets $A_n(i)$ are compact, and the functions $a \mapsto q_{ij}^n(a)$ and $a \mapsto r_n(i,a)$ are continuous on $A_n(i)$ for every $i, j \in S_n$.

(iv) Taking $c_2 \in \mathbb{R}$ and $b_2 \geq 0$ from Assumption 2.1.5(iii), the following inequality holds for every $(i,a) \in \mathbb{K}_n$:

$$\sum_{j\in S_n} q_{ij}^n(a)w^2(j) \leq -c_2 w^2(i) + b_2.$$

It should be clear from its definition that if w is a Lyapunov function for $\mathcal{M}$ then its restriction to S_n is as well a Lyapunov function for $\mathcal{M}_n$ for every $n \geq 1$. The conditions imposed in Assumption 2.2.2 mean, roughly, that the hypotheses for $\mathcal{M}$ are satisfied by the $\mathcal{M}_n$ "uniformly" in $n \geq 1$. Indeed, we are imposing that the constants taken from the assumptions on $\mathcal{M}$ are valid for the corresponding assumptions on the $\mathcal{M}_n$.

Under Assumption 2.2.2, we can use Theorem 2.1.3 for the control model $\mathcal{M}_n$ to ensure the existence of the controlled process, and we can define the discounted reward problem for the control models $\mathcal{M}_n$. We introduce some more notation. As already mentioned, the notation for $\mathcal{M}_n$ consists in adding a subscript n to the corresponding notation for $\mathcal{M}$. Given an initial state $i \in S_n$ and a control policy $\varphi \in \Phi_n$, its total expected discounted reward is

$$V_n^\alpha(i,\varphi) := E_n^{i,\varphi}\left[\int_0^\infty e^{-\alpha t} r_n(x(t),a(t))dt\right] = E_n^{i,\varphi}\left[\int_0^\infty e^{-\alpha t} r_n(t,x(t),\varphi)dt\right].$$

The optimal discounted reward is

$$V_n^\alpha(i) := \sup_{\varphi\in\Phi_n} V_n^\alpha(i,\varphi) \quad \text{for all } i \in S_n,$$

and the bounds in the w-norm (cf. (2.1.4))

$$||V_n^\alpha(\cdot,\varphi)||_w \leq \mathfrak{M} \quad \text{for all } \varphi \in \Phi_n, \quad \text{and} \quad ||V_n^\alpha||_w \leq \mathfrak{M}, \tag{2.2.1}$$

with $\mathfrak{M} := \frac{M(b_1+\alpha)}{\alpha(c_1+\alpha)}$ still hold (here, we make use that the constants M, b_1, c_1 are the same for every control model $\mathcal{M}_n$).

Note also that Theorem 2.1.7 remains valid for the control models $\mathcal{M}_n$, and so the optimal discounted reward $V_n^\alpha \in \mathcal{B}_w(S_n)$ as well as discount optimal policies in $\mathbb{F}_n$ can be characterized by means of the corresponding discounted reward optimality equation, which takes the form

$$\alpha V_n^\alpha(i) = \max_{a\in A_n(i)} \left\{ r_n(i,a) + \sum_{j\in S_n} q_{ij}^n(a) V_n^\alpha(j) \right\} \tag{2.2.2}$$

for each $i \in S_n$.

Preliminary results. Before proving our main results on convergence for discounted control models, we establish some useful results.

Lemma 2.2.3 *Suppose that $\mathcal{M}$ satisfies Assumptions 2.1.2 and 2.1.5(iii), and that the control models $\{\mathcal{M}_n\}_{n\geq 1}$ verify Assumptions 2.2.2(i) and (iv). Given $i \in S$ and $\epsilon > 0$, there exists some $K > i$ such that*

(i) for all $a \in A(i)$ we have $\sum_{j\geq K} q_{ij}(a)w(j) < \epsilon$

(ii) for all $n \geq n(i)$ and every $a \in A_n(i)$ we have $\sum_{j\in S_n, j\geq K} q_{ij}^n(a)w(j) < \epsilon$.

Suppose, in addition, that Assumption 2.1.5 holds and that $\mathcal{M}_n \to \mathcal{M}$. Then for every $i \in S$ and considering $n \geq n(i)$

(iii) $\lim_{n\to\infty} \sup_{a\in A_n(i)} |r_n(i,a) - r(i,a)| = 0.$

(iv) $\lim_{n\to\infty} \sup_{a\in A_n(i)} \sum_{j\in S_n} |q_{ij}^n(a) - q_{ij}(a)|w(j) = 0.$

Proof. (i)–(ii). Choose any $K > i$. Observe that for all $a \in A(i)$

$$\sum_{j\geq K} q_{ij}(a)w(j) \leq \frac{1}{w(K)} \sum_{j\geq K} q_{ij}(a)w^2(j),$$

where we make use of the monotonicity of w. On the other hand, since $K > i$, all the terms $q_{ij}(a)$ for $j \geq K$ are nonnegative and thus

$$\begin{aligned} \sum_{j\geq K} q_{ij}(a)w^2(j) &\leq \sum_{j\in S} q_{ij}(a)w^2(j) - q_{ii}(a)w^2(i) \\ &\leq -c_2 w^2(i) + b_2 + w^3(i). \end{aligned}$$

Similarly, if $n \geq 1$ is such that $i \in S_n$ or, equivalently, $n \geq n(i)$, then we can repeat the above arguments to obtain that, for each $a \in A_n(i)$,

$$\begin{aligned}\sum_{j\in S_n, j\geq K} q_{ij}^n(a)w(j) &\leq \frac{1}{w(K)} \sum_{j\in S_n, j\geq K} q_{ij}^n(a)w^2(j) \\ &\leq \frac{1}{w(K)}\Big(\sum_{j\in S_n} q_{ij}^n(a)w^2(j) - q_{ii}^n(a)w^2(i)\Big) \\ &\leq \frac{1}{w(K)}\Big(-c_2 w^2(i) + b_2 + w^3(i)\Big).\end{aligned}$$

Therefore, it suffices to choose $K > i$ such that

$$\frac{1}{w(K)}\Big(-c_2 w^2(i) + b_2 + w^3(i)\Big) < \epsilon$$

to obtain the stated result.

(iii). The proof is by contradiction. Suppose then that for some $i \in S$ there are $\epsilon > 0$ and actions $a_{n'} \in A_{n'}(i)$ for some subsequence n' such that $|r(i, a_{n'}) - r_n(i, a_{n'})| > \epsilon$. The action set $A(i)$ being compact, we can choose a further subsequence n'' with $a_{n''} \to a$ for some $a \in A(i)$. By continuity of the reward rate r we have $r(i, a_{n''}) \to r(i, a)$, while by Definition 2.2.1(d) we have $r_{n''}(i, a_{n''}) \to r(i, a)$, which leads to a contradiction.

(iv). We proceed by contradiction. Hence, for some given $i \in S$ we will suppose that there exists some $\epsilon > 0$ and some subsequence n', larger than $n(i)$, such that, for some $a_{n'} \in A_{n'}(i)$, we have

$$\sum_{j\in S_{n'}} |q_{ij}^{n'}(a_{n'}) - q_{ij}(a_{n'})|w(j) > \epsilon.$$

By parts (i) and (ii) of this lemma, there exists some $K > i$ such that for every n'

$$\sum_{j\geq K, j\in S_{n'}} q_{ij}(a_{n'})w(j) \leq \epsilon/3 \quad \text{and} \quad \sum_{j\geq K, j\in S_{n'}} q_{ij}^{n'}(a_{n'})w(j) \leq \epsilon/3.$$

Consider the n' such that $\{0, 1, \ldots, K-1\} \subseteq S_{n'}$, so that we have

$$\sum_{j=0}^{K-1} |q_{ij}^{n'}(a_{n'}) - q_{ij}(a_{n'})|w(j) > \epsilon/3 \quad \text{for all } n'. \tag{2.2.3}$$

Since the action sets $A(i)$ are compact, choose a further subsequence n'' such that $a_{n''} \to a$ for some $a \in A(i)$. Take now the limit through n'' in (2.2.3) and use continuity of the transition rates $q_{ij}(\cdot)$ and Definition 2.2.1(c) to reach a contradiction; indeed, both $q_{ij}^{n''}(a_{n''})$ and $q_{ij}(a_{n''})$ converge to $q_{ij}(a)$. □

Note that, under Assumptions 2.1.5(i)–(ii), parts (iii) and (iv) in this lemma imply items (c) and (d) in Definition 2.2.1. Under the assumptions of this lemma, these statements are equivalent, and we may use (iii)–(iv) in Lemma 2.2.3 in lieu of Definition 2.2.1(c)–(d). We introduce some more terminology.

Definition 2.2.4 *Suppose that $\{u_n\}_{n\geq 1}$ is a sequence of functions with $u_n \in \mathcal{B}_w(S_n)$ for each $n \geq 1$. We say that $\{u_n\}_{n\geq 1}$ converges pointwise to $u \in \mathcal{B}_w(S)$ if*

$$\lim_{n\to\infty} u_n(i) = u(i) \quad \textit{for each } i \in S.$$

Note that the expression $u_n(i)$ is defined provided that $n \geq n(i)$. Since the above definition is concerned with the limit as $n \to \infty$, we will not make it explicit in the notation.

We extend the definition of convergence of deterministic stationary policies given in Definition 2.1.16. Given a sequence of policies $\{f_n\}_{n\geq 1}$ with $f_n \in \mathbb{F}_n$ for every $n \geq 1$, we say that $\{f_n\}_{n\geq 1}$ converges to $f \in \mathbb{F}$ if

$$\lim_{n\to\infty} f_n(i) = f(i) \quad \text{for each } i \in S.$$

Once again, note that $f_n(i)$ is defined only for $n \geq n(i)$, but this is not made explicit in the notation since we are dealing with the limit as $n \to \infty$. In this case, we will also write $f_n \to f$. The notion of limit policy is then similar to that given in Definition 2.1.16.

Lemma 2.2.5 *(i) Suppose that the sequence $u_n \in \mathcal{B}_w(S_n)$, for every $n \geq 1$, satisfies $\sup_{n\geq 1} ||u_n||_w < \infty$. Then there exists a subsequence n' and $u \in \mathcal{B}_w(S)$ such that $\{u_{n'}\}$ converges pointwise to u.*

(ii) If Assumption 2.1.5(i) is satisfied then, given arbitrary $f_n \in \mathbb{F}_n$, for $n \geq 1$, there exist a subsequence n' and $f \in \mathbb{F}$ such that $\{f_{n'}\}$ converges to f.

Proof. This lemma follows from a standard diagonal argument. Indeed, for every $i \in S$, the sequence $\{u_n(i)\}_{n\geq n(i)}$ is bounded, and hence has a convergent subsequence. Similarly, we have that $\{f_n(i)\}_{n\geq n(i)}$ is a sequence in the compact metric space $A(i)$, and hence has a convergent subsequence. Use then the fact that S is countable to construct a subsequence n' such that $u_{n'}(i)$ or $f_{n'}(i)$ are all convergent for $i \in S$. □

We state our final preliminary result.

Lemma 2.2.6 *Suppose that $\mathcal{M}$ satisfies Assumptions 2.1.2 and 2.1.5(iii), and that the control models $\{\mathcal{M}_n\}_{n\geq 1}$ verify Assumptions 2.2.2(i) and (iv). Let $u_n \in \mathcal{B}_w(S_n)$ for $n \geq 1$ be such that $\sup_{n\geq 1} ||u_n||_w < \infty$ and such that $\{u_n\}_{n\geq 1}$ converges pointwise to some $u \in \mathcal{B}_w(S)$. Let $f_n \in \mathbb{F}_n$, for $n \geq 1$, be such that $f_n \to f$ for some $f \in \mathbb{F}$. Under these conditions, if $\mathcal{M}_n \to \mathcal{M}$ then*

$$\lim_{n\to\infty} \Big[r_n(i, f_n) + \sum_{j\in S_n} q_{ij}^n(f_n) u_n(j) \Big] = r(i, f) + \sum_{j\in S} q_{ij}(f) u(j) \quad \textit{for all } i \in S.$$

Proof. Let $\mathfrak{c} > 0$ be such that $||u_n||_w \leq \mathfrak{c}$ for all $n \geq 1$, and so $||u||_w \leq \mathfrak{c}$. Fix $i \in S$ and consider indices n such that $n \geq n(i)$.

The fact that $r_n(i, f_n)$ converges to $r(i, f)$ follows directly from Definition 2.2.1(d). Let us now analyze the second term in the limit. Fix $\epsilon > 0$ and for the small constant $\epsilon/4\mathfrak{c}$, let $K > i$ be as in Lemma 2.2.3. If $n \geq n(i)$ is such that, in addition, $\{0, 1, \ldots, K-1\} \subseteq S_n$, then

$$\begin{aligned}
&\Big| \sum_{j \in S_n} q_{ij}^n(f_n) u_n(j) - \sum_{j \in S} q_{ij}(f) u(j) \Big| \\
&\leq \Big| \sum_{j=0}^{K-1} \big[q_{ij}^n(f_n) u_n(j) - q_{ij}(f) u(j) \big] \Big| + \mathfrak{c} \cdot \sum_{j \in S_n, j \geq K} q_{ij}^n(f_n) w(j) + \mathfrak{c} \cdot \sum_{j \geq K} q_{ij}(f) w(j) \\
&\leq \Big| \sum_{j=0}^{K-1} \big[q_{ij}^n(f_n) u_n(j) - q_{ij}(f) u(j) \big] \Big| + \frac{\epsilon}{2}.
\end{aligned}$$

Therefore, since by Definition 2.2.1(c) we have $q_{ij}^n(f_n) \to q_{ij}(f)$ and as, by hypothesis, $u_n(j) \to u(j)$ for $0 \leq j < K$, choosing n large enough makes

$$\Big| \sum_{j=0}^{K-1} \big[q_{ij}^n(f_n) u_n(j) - q_{ij}(f) u(j) \big] \Big| < \epsilon/2$$

(here, note that K does not depend on n; cf. Lemma 2.2.3(b)). This completes the proof. □

Main result. Now we are ready to prove our main result on the convergence of the discounted control models.

Theorem 2.2.7 *Suppose that the control model $\mathcal{M}$ satisfies Assumptions 2.1.2, 2.1.4, and 2.1.5, and that the control models $\{\mathcal{M}_n\}_{n \geq 1}$ verify Assumption 2.2.2. If $\mathcal{M}_n \to \mathcal{M}$ then*

(i) For every $i \in S$ we have $\lim_{n \to \infty} V_n^\alpha(i) = V^\alpha(i)$.

(ii) If $f_n \in \mathbb{F}_n$, for $n \geq 1$, is a discount optimal policy for $\mathcal{M}_n$, then any limit policy $f \in \mathbb{F}$ of $\{f_n\}_{n \geq 1}$ is discount optimal for $\mathcal{M}$.

Proof. (i). Recalling that the optimal discounted rewards V_n^α are uniformly bounded in the w-norm (see (2.2.1)) and using Lemma 2.2.5, we deduce that there exists a subsequence n' such that $\{V_{n'}^\alpha\}$ converges pointwise to some $v \in \mathcal{B}_w(S)$.

Fix now an arbitrary state $i \in S$ and any action $a \in A(i)$, and consider indices $n' \geq n(i)$. By Hausdorff convergence of $A_{n'}(i)$ to $A(i)$, there exist actions $a_{n'} \in A_{n'}(i)$ such that $a_{n'} \to a$; recall Definition 2.2.1(b). From the discounted reward optimality equation for the control model $\mathcal{M}_{n'}$, given in (2.2.2), we obtain

$$\alpha V_{n'}^\alpha(i) \geq r_{n'}(i, a_{n'}) + \sum_{j \in S_{n'}} q_{ij}^{n'}(a_{n'}) V_{n'}^\alpha(j).$$

We can use Lemma 2.2.6 and take the limit as $n' \to \infty$ to obtain

$$\alpha v(i) \geq r(i,a) + \sum_{j \in S} q_{ij}(a) v(j).$$

Since this is valid for every $(i,a) \in \mathbb{K}$ we have thus established that

$$\alpha v(i) \geq \max_{a \in A(i)} \left\{ r(i,a) + \sum_{j \in S} q_{ij}(a) v(j) \right\} \quad \text{for every } i \in S. \tag{2.2.4}$$

To prove the reverse inequality, fix $i \in S$ and, for indices $n' \geq n(i)$, let $f_{n'} \in \mathbb{F}_{n'}$ be a discount optimal deterministic stationary policy for $\mathcal{M}_{n'}$. Using again Lemma 2.2.5, there exists a further subsequence n'' and some $f \in \mathbb{F}$ such that $f_{n''} \to f$. The policies $f_{n''}$ being optimal for $\mathcal{M}_{n''}$, they attain the maximum in the corresponding discounted reward optimality equation (use Theorem 2.1.7 for the control model $\mathcal{M}_{n''}$), that is,

$$\alpha V^{\alpha}_{n''}(i) = r_{n''}(i, f_{n''}) + \sum_{j \in S_{n''}} q^{n''}_{ij}(f_{n''}) V^{\alpha}_{n''}(j). \tag{2.2.5}$$

Take the limit as $n'' \to \infty$ and use Lemma 2.2.6 to obtain

$$\alpha v(i) = r(i,f) + \sum_{j \in S} q_{ij}(f) v(j).$$

Combining this equation with (2.2.4), we have thus proved that

$$\alpha v(i) = \max_{a \in A(i)} \left\{ r(i,a) + \sum_{j \in S} q_{ij}(a) v(j) \right\} \quad \text{for every } i \in S,$$

that is, we have established that $v \in \mathcal{B}_w(S)$ is indeed a solution to the discounted reward optimality equation for $\mathcal{M}$ and, therefore, $v = V^{\alpha}$, the optimal discounted reward of $\mathcal{M}$.

Therefore, we have proved that the limit of V^{α}_n through any pointwise convergent subsequence n' is V^{α}. Then, we necessarily have that V^{α}_n converges pointwise to V^{α}, establishing part (i) of the theorem.

(ii). To prove this part, let $f \in \mathbb{F}$ be any limit policy of optimal policies $\{f_n\}$ for $\mathcal{M}_n$. Proceed as in (2.2.5) to derive that f attains the maximum in the discounted reward optimality equation for $\mathcal{M}$ and, hence, it is discount optimal. □

As a consequence of item (ii) in this theorem, if the discount optimal policy for $\mathcal{M}$ is unique and $\{f_n\}$ are optimal policies for $\mathcal{M}_n$, then we necessarily have $f_n \to f$. Otherwise, it is not possible to ensure convergence of $\{f_n\}$ although, as it has been shown, any limit policy is optimal for $\mathcal{M}$.

Summarizing the results in this section, starting from a control model $\mathcal{M}$ satisfying suitable hypotheses, and also from a sequence of control models $\{\mathcal{M}_n\}$ that verify similar conditions, we have shown that convergence $\mathcal{M}_n \to \mathcal{M}$ implies convergence of the optimal

discounted reward and of optimal policies. Thus, the approximating control models $\mathcal{M}_n$ can indeed be used to approximate the original control model $\mathcal{M}$.

There remains, however, an important open issue. Usually, the decision-maker is given the control model $\mathcal{M}$ and he is interested in approximating its optimal solution. But, generally, the decision-maker is not given the sequence of approximating control models $\{\mathcal{M}_n\}$. The question, rather, is whether starting from the original control model $\mathcal{M}$ the controller is able to construct a sequence of "simpler" approximating control models $\{\mathcal{M}_n\}$ which in principle he is able to solve. This is the issue addressed in Section 2.3.

Another important remark is the following. Theorem 2.2.7 gives plain convergence of V_n^α to V^α. It would be interesting to know if some kind of convergence rate could be provided, so as to obtain some error bounds on the approximations. This is addressed as well in Section 2.3.

2.2.3 The average reward criterion

We consider now the control model $\mathcal{M}$ under the long-run expected average reward optimality criterion. We consider also the sequence $\{\mathcal{M}_n\}_{n\geq 1}$ of control models, defined in Section 2.2.1 above. In Section 2.1.3 we gave conditions for the solvability of the average reward problem for $\mathcal{M}$. Now we impose such conditions on the control models $\mathcal{M}_n$.

Assumption 2.2.8 *Let w be the Lyapunov function in Assumption 2.1.8. The following conditions hold for every $n \geq 1$.*

(i) With the constants $c_1 > 0$ and $b_1 \geq 0$ as in Assumption 2.1.8, and for some finite set $D_n \subseteq S_n$ we have

$$\sum_{j\in S_n} q_{ij}^n(a)w(j) \leq -c_1 w(i) + b_1 \mathbf{I}_{D_n}(i) \quad \text{for all } (i,a) \in \mathbb{K}_n,$$

with $q_n(i) \leq w(i)$ for each $i \in S_n$.

(ii) With the constant $M > 0$ taken from Assumption 2.1.9, we have

$$|r_n(i,a)| \leq Mw(i) \quad \text{for all } (i,a) \in \mathbb{K}_n.$$

(iii) The action sets $A_n(i)$ are compact, and the functions $a \mapsto q_{ij}^n(a)$ and $a \mapsto r_n(i,a)$ are continuous on $A_n(i)$ for every $i, j \in S_n$.

(iv) Taking $c_2 \in \mathbb{R}$ and $b_2 \geq 0$ from Assumption 2.1.11(iii), the following inequality holds for every $(i,a) \in \mathbb{K}_n$:

$$\sum_{j\in S_n} q_{ij}^n(a)w^2(j) \leq -c_2 w^2(i) + b_2.$$

(v) Every deterministic stationary policy in $\mathbb{F}_n$ is irreducible.

The above conditions impose that the assumptions we made on the control model $\mathcal{M}$ for the average reward criterion, namely, Assumptions 2.1.8, 2.1.9, and 2.1.11, hold "uniformly" in $n \geq 1$. Indeed, as can be seen from Assumption 2.2.8, all the involved constants are the same for every $\mathcal{M}_n$, $n \geq 1$, and for $\mathcal{M}$. In particular, each control model $\mathcal{M}_n$ is uniformly exponentially ergodic (see (2.1.6)), but it is important to mention that the above conditions do not necessarily imply that the corresponding constants R_n and γ_n do not depend on $n \geq 1$.

Now we introduce the notation for the average control model $\mathcal{M}_n$. Given a Markov policy $\varphi \in \Phi_n$ and an initial state $i \in S_n$, consider the associated probability measure $P_n^{i,\varphi}$ and expectation operator $E_n^{i,\varphi}$ that model the controlled process; those indeed exist as a consequence of Theorem 2.1.3 and Assumption 2.2.8(i). The expected average payoff is

$$J_n(i,\varphi) = \limsup_{T\to\infty} \frac{1}{T} E_n^{i,\varphi}\Big[\int_0^T r_n(x(t),a(t))dt\Big] = \limsup_{T\to\infty} \frac{1}{T} E_n^{i,\varphi}\Big[\int_0^T r_n(t,x(t),\varphi)dt\Big],$$

and the optimal average reward is

$$J_n(i) = \sup_{\varphi\in\Phi_n} J_n(i,\varphi) \quad \text{for each } i \in S_n.$$

Under our assumptions on $\mathcal{M}_n$ we have, as in (2.1.5),

$$|J_n(i,\varphi)| \leq \frac{Mb_1}{c_1} \quad \text{and} \quad |J_n(i)| \leq \frac{Mb_1}{c_1} \quad \text{for all } \varphi \in \Phi_n \text{ and } i \in S_n. \tag{2.2.6}$$

By Assumptions 2.2.8(i) and (v), for each deterministic stationary policy $f \in \mathbb{F}_n$, the Markov chain $\{x(t)\}_{t\geq 0}$ under f has a unique invariant probability measure μ_f^n on S_n, for which $\mu_f^n(w)$ is finite. Moreover, the average reward of $f \in \mathbb{F}_n$ is constant:

$$J_n(i,f) = \sum_{j\in S_n} r_n(j,f)\mu_f^n(j) =: g_n(f) \quad \text{for all } i \in S_n,$$

where we recall that $g_n(f)$ is called the gain of $f \in \mathbb{F}_n$.

Finally, under Assumption 2.2.8, an analogous of Theorem 2.1.13 holds. In particular, the optimal average reward of $\mathcal{M}_n$ is constant:

$$g_n^* = J_n(i) \quad \text{for all } i \in S_n,$$

and there exist solutions $(g,h) \in \mathbb{R} \times \mathcal{B}_w(S_n)$ to the average reward optimality equation for $\mathcal{M}_n$

$$g = \max_{a\in A_n(i)} \big\{r_n(i,a) + \sum_{j\in S_n} q_{ij}^n(a)h(j)\big\} \quad \text{for all } i \in S_n. \tag{2.2.7}$$

In this case, we have $g = g_n^*$, while h is unique up to additive constants. Moreover, optimal deterministic stationary policies are characterized as those achieving the maximum in this optimality equation.

Now we address the issue of the convergence of the optimal gains g_n^* of $\mathcal{M}_n$ to the optimal gain g^* of $\mathcal{M}$. To obtain this result, however, we must impose some additional conditions. In fact, we will propose two different sufficient conditions under which this convergence takes place.

Theorem 2.2.9 *Suppose that the control model $\mathcal{M}$ verifies Assumptions 2.1.8, 2.1.9, and 2.1.11, and that the control models $\{\mathcal{M}_n\}_{n\geq 1}$ satisfy Assumption 2.2.8. In addition, suppose that there exist solutions $(g_n^*, h_n) \in \mathbb{R}\times\mathcal{B}_w(S_n)$ to the average reward optimality equation (2.2.7) for $\mathcal{M}_n$ such that $\sup_{n\geq 1}||h_n||_w < \infty$. Under these conditions, if $\mathcal{M}_n \to \mathcal{M}$ then*

(i) The optimal gains converge: $\lim_{n\to\infty} g_n^* = g^*$.

(ii) If $f_n \in \mathbb{F}_n$ is an average reward optimal policy for $\mathcal{M}_n$, then every limit policy of $\{f_n\}_{n\geq 1}$ in $\mathbb{F}$ is average optimal for $\mathcal{M}$.

Proof. (i). We know that the sequence $\{g_n^*\}_{n\geq 1}$ is bounded (recall (2.2.6)) and, by hypothesis, the sequence $\{h_n\}_{n\geq 1}$ of solutions to the average reward optimality equation for $\mathcal{M}_n$ is also bounded in the w-norm. Therefore, by Lemma 2.2.5, there exists a subsequence (that without loss of generality we will still denote by n) and a pair $(g,h) \in \mathbb{R}\times\mathcal{B}_w(S)$ such that

$$\lim_{n\to\infty} g_n^* = g \quad \text{and} \quad \lim_{n\to\infty} h_n(i) = h(i) \quad \text{for all } i \in S.$$

(Recall that the latter expression is defined only when $n \geq n(i)$.)

Consider now a fixed $(i,a) \in \mathbb{K}$. By Definition 2.2.1(b) there exists a sequence $\{a_n\}_{n\geq n(i)}$ with $a_n \in A_n(i)$ and $a_n \to a$ as $n \to \infty$. From the average reward optimality equation for $\mathcal{M}_n$ we obtain

$$g_n^* \geq r_n(i,a_n) + \sum_{j\in S_n} q_{ij}^n(a_n)h_n(j).$$

We can use Lemma 2.2.6 to take the limit as $n \to \infty$, which yields

$$g \geq r(i,a) + \sum_{j\in S} q_{ij}(a)h(j).$$

Since this is valid for every $(i,a) \in \mathbb{K}$ we have thus established that

$$g \geq \max_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} q_{ij}(a)h(j) \right\} \quad \text{for each } i \in S.$$

Suppose now that $f_n \in \mathbb{F}_n$ is average optimal for $\mathcal{M}_n$. Again by Lemma 2.2.5, there exists some $f \in \mathbb{F}$ and a further subsequence (that we shall still denote by n) such that $f_n \to f$. For such $f_n \in \mathbb{F}_n$, the optimality equation for $\mathcal{M}_n$ reads

$$g_n^* = r_n(i,f_n) + \sum_{j\in S_n} q_{ij}^n(f_n)h_n(j) \quad \text{for each } i \in S_n. \tag{2.2.8}$$

Using Lemma 2.2.6 we deduce that

$$g = r(i,f) + \sum_{j\in S} q_{ij}(f)h(j) \quad \text{for each } i \in S.$$

Therefore, we have established that $(g, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ is a solution to the average reward optimality equation for $\mathcal{M}$, and so $g = g^*$.

Summarizing, we have shown that any convergent subsequence of the bounded sequence $\{g_n^*\}_{n\geq 1}$ converges to g^*. This implies that the whole sequence converges to g^*, that is, $\lim_{n\to\infty} g_n^* = g^*$.

(ii). To prove that any limit policy of average optimal policies $\{f_n\}_{n\geq 1}$ for $\mathcal{M}_n$ is average optimal for $\mathcal{M}$, use (2.2.8) and take the limit through some subsequence such that $f_{n'} \to f \in \mathbb{F}$ and such that $h_{n'}$ converges pointwise to obtain that f indeed attains the maximum in the average reward optimality equation for $\mathcal{M}$. □

Observe that the analogous result for discounted reward models (Theorem 2.2.7) holds without any additional condition, except for the basic assumptions on $\mathcal{M}$ and $\mathcal{M}_n$ and the convergence $\mathcal{M}_n \to \mathcal{M}$. Here, for the average reward optimality criterion, we need to impose further requirements, namely, the existence of bounded solutions h_n to the average optimality equation for the $\mathcal{M}_n$.

In connection with Remark 2.1.12, observe that if all the control models $\mathcal{M}_n$ satisfy the monotonicity and irreducibility properties given in Remark 2.1.12, together with $D_n = \{0\}$ in Assumption 2.2.8(i), then there exist solutions h_n to the average reward optimality equation for $\mathcal{M}_n$ that satisfy

$$||h_n||_w \leq \frac{2M(1 + b_1/c_1)}{c_1}$$

(recall Corollary 2.1.15), and so the condition in Theorem 2.2.9 is indeed satisfied.

We propose now another sufficient condition to obtain convergence of the optimal gains. The idea is to drop the condition $\sup_{n\geq 1} ||h_n||_w < \infty$ in Theorem 2.2.9 and to use, instead, a Lyapunov condition on some power $\delta > 2$ of the function w.

Theorem 2.2.10 *Suppose that the control model $\mathcal{M}$ verifies Assumptions 2.1.8, 2.1.9, and 2.1.11, and that the control models $\{\mathcal{M}_n\}_{n\geq 1}$ satisfy Assumption 2.2.8. In addition, suppose that there exist constants $\delta > 2$, $c_\delta > 0$ and $b_\delta > 0$ with*

$$\sum_{j\in S_n} q_{ij}^n(a) w^\delta(j) \leq -c_\delta w^\delta(i) + b_\delta \quad \textit{for all } n \geq 1 \textit{ and } (i, a) \in \mathbb{K}_n.$$

Under these conditions, if $\mathcal{M}_n \to \mathcal{M}$ then

(i) The optimal gains converge: $\lim_{n\to\infty} g_n^* = g^*$.

(ii) If $f_n \in \mathbb{F}_n$ is an average reward optimal policy for $\mathcal{M}_n$, then every limit policy of $\{f_n\}_{n\geq 1}$ in $\mathbb{F}$ is average optimal for $\mathcal{M}$.

Proof. Given $n \geq 1$, fix an arbitrary policy $f \in \mathbb{F}_n$ and an initial state $i \in S_n$. Observe that for every $k \geq 0$ and $t \geq 0$

$$\begin{aligned}\sum_{j\in S_n, j\geq k} P_n^{i,f}\{x(t)=j\}w^2(j) &\leq \frac{1}{w^{\delta-2}(k)} \sum_{j\in S_n, j\geq k} P_n^{i,f}\{x(t)=j\}w^{\delta}(j) \\ &\leq \frac{1}{w^{\delta-2}(k)} \sum_{j\in S_n} P_n^{i,f}\{x(t)=j\}w^{\delta}(j) \\ &= \frac{1}{w^{\delta-2}(k)} E_n^{i,f}[w^{\delta}(x(t))].\end{aligned}$$

By an analogous to (2.1.3) but now for the function w^δ we have

$$E_n^{i,f}[w^{\delta}(x(t))] \leq e^{-c_\delta t}w(i) + b_\delta/c_\delta,$$

and thus

$$\sum_{j\in S_n, j\geq k} P_n^{i,f}\{x(t)=j\}w^2(j) \leq \frac{1}{w^{\delta-2}(k)}\big(e^{-c_\delta t}w(i) + b_\delta/c_\delta\big).$$

Recalling that μ_f^n is the invariant probability measure of f for the control model $\mathcal{M}_n$, we can use Fatou's lemma as $t \to \infty$ to get

$$\sum_{j\in S_n, j\geq k} \mu_f^n(j)w^2(j) \leq \frac{b_\delta}{c_\delta w^{\delta-2}(k)}.$$

Summarizing, we have shown the following w^2-uniform integrability result:

$$\lim_{k\to\infty} \sup_{n\geq 1, f\in\mathbb{F}_n} \sum_{j\in S_n, j\geq k} \mu_f^n(j)w^2(j) = 0.$$

We will need also the following result. Given $(i,a) \in \mathbb{K}$ and $u \in \mathcal{B}_w(S)$, or $(i,a) \in \mathbb{K}_n$ and $u \in \mathcal{B}_w(S_n)$, we have

$$\Big|\sum_{j\in S} q_{ij}(a)u(j)\Big| \leq ||u||_w(2+b_1)w^2(i) \quad \text{and} \quad \Big|\sum_{j\in S_n} q_{ij}^n(a)w(j)\Big| \leq ||u||_w(2+b_1)w^2(i), \tag{2.2.9}$$

respectively. Indeed, if $(i,a) \in \mathbb{K}$, we have

$$\Big|\sum_{j\in S} q_{ij}(a)u(j)\Big| \leq ||u||_w\big(-2q_{ii}(a)w(i) + \sum_{j\in S} q_{ij}(a)w(j)\big) \leq ||u||_w(2w^2(i)+b_1),$$

by Assumption 2.1.8. The argument is similar for $(i,a) \in \mathbb{K}_n$.

Suppose now that $\epsilon > 0$ is given.

- Choose K_1 such that

$$\sum_{j\in S_n, j>K_1} \mu_f^n(j)w^2(j) < \epsilon \quad \text{for every } n \geq 1 \text{ and } f \in \mathbb{F}_n. \tag{2.2.10}$$

- For this K_1, choose $K_2 > K_1$ with

$$\max_{i=0,1,\ldots,K_1} \max_{a\in A(i)} \sum_{j>K_2} q_{ij}(a)w(j) < \epsilon \tag{2.2.11}$$

  (recall Lemma 2.2.3(i)).

- For these K_1 and K_2 choose N_0 with $\{0,1,\ldots,K_2\} \subseteq S_{N_0}$ such that $n \geq N_0$ implies (Lemma 2.2.3(iii))

$$\max_{i=0,1,\ldots,K_1} \max_{a\in A_n(i)} |r(i,a) - r_n(i,a)| < \epsilon \tag{2.2.12}$$

  and (Lemma 2.2.3(iv))

$$\max_{i=0,1,\ldots,K_1} \max_{a\in A_n(i)} \sum_{j\in S_n} |q_{ij}(a) - q^n_{ij}(a)|w(j) < \epsilon. \tag{2.2.13}$$

Let $n \geq N_0$ and fix arbitrary $f \in \mathbb{F}_n$. We can extend $f \in \mathbb{F}_n$ to some policy $\overline{f} \in \mathbb{F}$, where $\overline{f}(i) = f(i) \in A_n(i) \subseteq A(i)$ for $i \in S_n$, and $\overline{f}(i) \in A(i)$ is defined arbitrarily for $i \notin S_n$. The Poisson equation for $\overline{f}$ is

$$g(\overline{f}) = r(i,\overline{f}) + \sum_{j\in S} q_{ij}(\overline{f})h(j) \quad \text{for all } i \in S,$$

where $h \in \mathcal{B}_w(S)$ can be chosen to be the bias of $\overline{f}$, and thus $||h||_w \leq RM/\gamma$; recall (2.1.7) and Proposition 2.1.14. This Poisson equation is written on the states $i \in S_n$ as

$$g(\overline{f}) - r_n(i,f) = r(i,f) - r_n(i,f) + \sum_{j\in S} q_{ij}(f)h(j) \quad \text{for all } i \in S_n,$$

where we can indeed replace $\overline{f}$ with f in the righthand side. Multiply the above equations by the invariant probability measure $\mu^n_f(i)$ and sum over $i \in S_n$ to obtain

$$g(\overline{f}) - g_n(f) = \sum_{i\in S_n} \mu^n_f(i)\big(r(i,f) - r_n(i,f)\big) + \sum_{i\in S_n} \mu^n_f(i) \sum_{j\in S} q_{ij}(f)h(j). \tag{2.2.14}$$

Let us first analyze the leftmost term in the righthand side of this expression. For states $0 \leq i \leq K$ we have by (2.2.12)

$$\Big| \sum_{i=0}^{K_1} \mu^n_f(i)\big(r(i,f) - r_n(i,f)\big)\Big| \leq \sum_{j=0}^{K_1} \mu^n_f(i)\epsilon \leq \epsilon.$$

For states larger than K_1 we have, by (2.2.10),

$$\Big| \sum_{i\in S_n, i>K_1} \mu^n_f(i)\big(r(i,f) - r_n(i,f)\big)\Big| \leq 2M \sum_{i\in S_n, i>K_1} \mu^n_f(i)w(i) \leq 2M\epsilon,$$

where we use the fact that $w(i) \geq 1$.

We analyze now the rightmost term of (2.2.14). Note that it equals

$$\begin{aligned}&\sum_{i\in S_n}\mu_f^n(i)\sum_{j\in S_n}q_{ij}(f)h(j)+\sum_{i\in S_n}\mu_f^n(i)\sum_{j\notin S_n}q_{ij}(f)h(j)\\&=\sum_{i\in S_n}\mu_f^n(i)\sum_{j\in S_n}(q_{ij}(f)-q_{ij}^n(f))h(j)+\sum_{i\in S_n}\mu_f^n(i)\sum_{j\notin S_n}q_{ij}(f)h(j),\end{aligned}\tag{2.2.15}$$

where me make use of the equality $\sum_{i\in S_n}\mu_f^n(i)q_{ij}^n(f)=0$ for all $j\in S_n$ because μ_f^n is the invariant probability measure of $f\in\mathbb{F}_n$ under $\mathcal{M}_n$. Regarding the leftmost term of (2.2.15), it can be split into the sums for $0\leq i\leq K_1$ and for $i\in S_n$ with $i>K_1$. Firstly, by (2.2.13),

$$\sum_{i=0}^{K_1}\mu_f^n(i)\Big|\sum_{j\in S_n}(q_{ij}(f)-q_{ij}^n(f))h(j)\Big|\leq\frac{RM}{\gamma}\,\epsilon\sum_{i=0}^{K-1}\mu_f^n(i)\leq\frac{RM}{\gamma}\,\epsilon.$$

Secondly, applying (2.2.9)

$$\begin{aligned}\sum_{i\in S_n,i>K_1}\mu_f^n(i)\Big|\sum_{j\in S_n}(q_{ij}(f)-q_{ij}^n(f))h(j)\Big|&\leq 2||h||_w(2+b_1)\sum_{i\in S_n,i>K_1}\mu_f^n(i)w^2(i)\\&\leq\frac{2RM(2+b_1)}{\gamma}\,\epsilon,\end{aligned}$$

by (2.2.10). We proceed now with the rightmost term of (2.2.15). We have

$$\sum_{i\in S_n}\mu_f^n(i)\Big|\sum_{j\notin S_n}q_{ij}(f)h(j)\Big|\leq\frac{RM}{\gamma}\sum_{i\in S_n}\mu_f^n(i)\sum_{j\notin S_n}q_{ij}(f)w(j).$$

For states $0\leq i\leq K_1$, by (2.2.11) we have

$$\sum_{i=0}^{K_1}\mu_f^n(i)\sum_{j\notin S_n}q_{ij}(f)w(j)\leq\sum_{i=0}^{K_1}\mu_f^n(i)\sum_{j>K_2}q_{ij}(f)w(j)\leq\epsilon,$$

while for states $i\in S_n, i>K_1$, proceeding as in the proof of (2.2.9), we have

$$\sum_{i\in S_n,i>K_1}\mu_f^n(i)\sum_{j\notin S_n}q_{ij}(f)w(j)\leq(2+b_1)\sum_{i\in S_n,i>K_1}\mu_f^n(i)w^2(i)\leq(2+b_1)\epsilon.$$

Consequently, letting $\mathfrak{c}=1+2M+RM(3b_1+8)/\gamma$, we have shown that $|g(\overline{f})-g_n(f)|<\mathfrak{c}\epsilon$ for all $n\geq N_0$. Since $f\in\mathbb{F}_n$ is arbitrary it follows that

$$\lim_{n\to\infty}\sup_{f\in\mathbb{F}_n}|g(\overline{f})-g_n(f)|=0,\tag{2.2.16}$$

where $\overline{f}$ is any extension to $\mathbb{F}$ of $f \in \mathbb{F}_n$.

Now we are ready to conclude the proof. Suppose that $f^* \in \mathbb{F}$ is an average reward optimal policy for $\mathcal{M}$. We can construct a sequence $\{f_n\}_{n\geq 1}$ of policies in $\mathbb{F}_n$ such that $f_n \to f^*$ and let $\overline{f}_n$ be an arbitrary extension of f_n to $\mathbb{F}$. It should be clear that $\overline{f}_n \in \mathbb{F}$ converges to $f^* \in \mathbb{F}$. Given $\epsilon > 0$, for n large enough we have, by continuity of the gain (Lemma 2.1.17)

$$|g(\overline{f}_n) - g^*| < \epsilon,$$

and also that

$$g(\overline{f}_n) < \epsilon + g_n(f_n) \leq \epsilon + g_n^*,$$

from which $g^* - g_n^* < 2\epsilon$ follows. Conversely, if $f_n^* \in \mathbb{F}_n$ is average optimal for $\mathcal{M}_n$, then given $\epsilon > 0$ and for n large enough

$$g_n^* - \epsilon \leq g(\overline{f}_n) \leq g^*$$

for any extension $\overline{f}_n \in \mathbb{F}$ of f_n^*, and so $g_n^* - g^* < \epsilon$. This completes the proof of part (i) of this theorem, that $g_n^* \to g^*$.

For statement (ii), let $\{f_n\}_{n\geq 1}$ be average optimal policies for $\mathcal{M}_n$ and consider a subsequence (for simplicity, also denoted by n) that converges to some $f \in \mathbb{F}$. If $\overline{f}_n$ is an extension of f_n to $\mathbb{F}$, then we also have that $\overline{f}_n$ converges to $f \in \mathbb{F}$. Using (2.2.16) it follows that

$$\lim_{n\to\infty} \left[g(\overline{f}_n) - g_n^*\right] = 0,$$

and so, by part (i) and Lemma 2.1.17 again, $g(f) = \lim g(\overline{f}_n) = g^*$, thus showing that f is average optimal for $\mathcal{M}$. □

This theorem, whose proof is by far more involved than that of Theorem 2.2.9, allows to drop the condition $\sup_{n\geq 1} ||h_n||_w < \infty$ imposed in that theorem. The inconvenient of this condition in Theorem 2.2.9 is that it practically assumes that the constants R_n and γ_n in the uniform ergodicity condition for $\mathcal{M}_n$ do not depend on $n \geq 1$. Such a result is not readily available since, except for the particular case described in Remark 2.1.12, there is no explicit known relation between the coefficients R and γ and the data of the control model. On the contrary, the strengthened Lyapunov condition presented in Theorem 2.2.10 depends directly on the data (namely, the transition rates) of the control model $\mathcal{M}_n$ and, therefore, it is easy to verify it (or discard it) in practice.

2.3 Finite state and action approximations

In the previous section, we have analyzed convergence of a sequence of given control models $\mathcal{M}_n$ to a so-called original control model $\mathcal{M}$, and we have studied several properties of this convergence. Here we take another point of view: we assume that we are given the original control model $\mathcal{M}$ and we show how we can construct a sequence $\{\mathcal{M}_n\}_{n\geq 1}$ of control models that verify the hypotheses described in the previous section.

2.3.1 Definition

Consider the control model $\mathcal{M} = \{S, A, \mathbb{K}, q, r\}$ described in Section 2.1.1. We are interested in approximating numerically the optimal value and the optimal policies for $\mathcal{M}$, either for the discounted or the average reward optimality criteria.

We propose the following finite state and action truncation of the control model $\mathcal{M}$. We must assume, further, that the action sets $A(i)$ of $\mathcal{M}$ are compact for every $i \in S$; cf. Assumptions 2.1.5(i) and 2.1.11(i). For any $n \geq 1$, consider the control model $\mathcal{M}_n = \{S_n, A, \mathbb{K}_n, q_n, r_n\}$ defined as follows:

- The state space is $S_n = \{0, 1, \ldots, n\}$.
- For each $i \in S_n$, the action set $A_n(i)$ is a finite subset of $A(i)$ such that the condition in Definition 2.2.1(b) is verified (see the comment below).
- The transition rates are as follows. If $(i, a) \in \mathbb{K}_n$ and $0 \leq j < n$, let $q_{ij}^n(a) = q_{ij}(a)$, and if $j = n$ let
$$q_{in}^n(a) = \sum_{k \geq n} q_{ik}(a) = -\sum_{k=0}^{n-1} q_{ik}(a).$$
- For $(i, a) \in \mathbb{K}_n$, define the reward rate $r_n(i, a) = r(i, a)$.

Regarding the construction of the action sets, such a construction is indeed possible because the action sets $A(i)$ are compact. As an illustration, consider the family of balls with center in $A(i)$ and radius $1/n$. Define $A_n(i)$ as the set of centers of a finite subcover. Then the Hausdorff distance verifies $\rho_A(A(i), A_n(i)) \leq 1/n$, thus satisfying Definition 2.2.1(b). Observe also that the transition rates are conservative and stable, with $q_n(i) \leq q(i)$ for every $i \in S_n$. Indeed, $-q_{ii}^n(a) = -q_{ii}(a) \leq q(i)$ for $(i, a) \in \mathbb{K}_n$ with $i < n$, while for $a \in A_n(n)$
$$-q_{nn}^n(a) = \sum_{k=0}^{n-1} q_{nk}(a) \leq -q_{nn}(a) \leq q(n).$$
Finally, note that $n(0) = 1$ and that $n(i) = i$ for all $i \geq 1$.

The interpretation of $\mathcal{M}_n$ is as follows. We can say, loosely, that the truncated control model $\mathcal{M}_n$ follows the same dynamics as $\mathcal{M}$, but when it reaches a state larger than n it is "restarted" at n.

Our next lemma needs continuity of the transition and reward rates; cf. Assumptions 2.1.5(ii) and 2.1.11(ii).

Lemma 2.3.1 *Suppose that the control model $\mathcal{M}$ is such that its action sets $A(i)$ are compact for every $i \in S$, and such that the functions $a \mapsto q_{ij}(a)$ and $a \mapsto r(i, a)$ are continuous on $A(i)$ for each $i, j \in S$. Then the control models $\mathcal{M}_n$ defined above verify $\mathcal{M}_n \to \mathcal{M}$.*

Proof. It is clear that $S_n \uparrow S$ and, by construction, convergence in the Hausdorff metric of the $A_n(i)$ to $A(i)$ holds. Given $i, j \in S$ and a sequence $a_n \in A_n(i)$ such that $a_n \to a \in A(i)$, for n large enough we have $q_{ij}^n(a_n) = q_{ij}(a_n)$, which converges to $q_{ij}(a)$ by continuity of the transition rates; thus Definition 2.2.1(c) holds. A similar argument is valid for the condition on the reward rates given in Definition 2.2.1(d). □

Consequently, the control models $\mathcal{M}_n$ constructed above are finite state and action truncations of the original control model $\mathcal{M}$ and, besides, under some additional conditions on $\mathcal{M}$, they converge: $\mathcal{M}_n \to \mathcal{M}$. It remains to study if the control models $\mathcal{M}_n$ somehow inherit the assumptions so far imposed on $\mathcal{M}$, so that we can use Theorems 2.2.7, 2.2.9, and 2.2.10 to obtain convergence of the optimal values and the optimal policies.

2.3.2 Finite truncations for discounted models

Consider the control model $\mathcal{M}$ and let us focus on the discounted reward optimality criterion, with a discount rate $\alpha > 0$. Our first task is to check whether the finite state and action truncations $\mathcal{M}_n$ defined previously verify the conditions in Assumption 2.2.2, so that we can use Theorem 2.2.7.

Proposition 2.3.2 *Suppose that the control model $\mathcal{M}$ satisfies Assumptions 2.1.2, 2.1.4, and 2.1.5. Then the control models $\mathcal{M}_n$ constructed in Section 2.3.1 verify:*

(i) The optimal discounted rewards converge: for every $i \in S$, $\lim_{n\to\infty} V_n^\alpha(i) = V^\alpha(i)$.

(ii) Any limit policy of discount optimal policies for $\mathcal{M}_n$ is discount optimal for $\mathcal{M}$.

Proof. Our first step in this proof is to show that Assumptions 2.1.2, 2.1.4, and 2.1.5 on the control model $\mathcal{M}$ imply Assumption 2.2.2 for the finite state and action truncated control models $\mathcal{M}_n$.

Let us check Assumption 2.2.2(i). Recall that the Lyapunov function w, as well as the constants $c_1 > -\alpha$ and $b_1 \geq 0$, are taken from Assumption 2.1.2. Given $n \geq 1$ and $(i, a) \in \mathbb{K}_n$,

$$\begin{aligned}\sum_{j\in S_n} q_{ij}^n(a)w(j) &= \sum_{j=0}^{n-1} q_{ij}(a)w(j) + \Big(\sum_{j\geq n} q_{ij}(a)\Big)\cdot w(n)\\ &\leq \sum_{j=0}^{n-1} q_{ij}(a)w(j) + \sum_{j\geq n} q_{ij}(a)w(j)\\ &= \sum_{j\in S} q_{ij}(a)w(j) \leq -c_1 w(i) + b_1,\end{aligned}$$

where we use monotonicity of w and the fact that $q_{ij}(a) \geq 0$ for $j > n \geq i$. We have already mentioned that $q_n(i) \leq q(i) \leq w(i)$ for all $i \in S_n$. Therefore, Assumption 2.2.2(i) holds.

Clearly, by construction we have also Assumption 2.2.2(ii), while Assumption 2.2.2(iii) trivially holds because the sets $A_n(i)$ are finite. Finally, Assumption 2.2.2(iv) is derived using the same arguments as before, for part (i).

The control models $\mathcal{M}_n$ converge to $\mathcal{M}$ (recall Lemma 2.3.1) and they satisfy Assumption 2.2.2. We can therefore use Theorem 2.2.7, and the proof is complete. □

Therefore, starting from a control model $\mathcal{M}$ that satisfies suitable assumptions, we have been able to construct a sequence of control models $\mathcal{M}_n$, with finite state and action spaces, whose optimal discounted value and discount optimal policies converge to those of $\mathcal{M}$. This enables us to provide computable numerical approximations of the solution of a control model with countable state space and compact action sets.

Solving a finite discounted control problem. For completeness, we show now how we can explicitly solve a finite state and action control model, as the $\mathcal{M}_n$, under the discounted reward optimality criterion. We use the well known technique of uniformization.

Consider the finite state and action control model $\mathcal{M}_n$ for some fixed $n \geq 1$. Its discounted reward optimality equation reads

$$\alpha V_n^\alpha(i) = \max_{a \in A_n(i)} \left\{ r_n(i,a) + \sum_{j \in S_n} q_{ij}^n(a) V_n^\alpha(j) \right\} \quad \text{for } i \in S_n.$$

Let the constant $\mathbf{q}_n$ be such that $\mathbf{q}_n > q_n(i)$ for all $i \in S_n$. It suffices to choose, for instance, $\mathbf{q}_n > w(n)$. Given $(i,a) \in \mathbb{K}_n$ consider the following probability distribution on S_n:

$$p_{ij}^n(a) = \frac{q_{ij}^n(a)}{\mathbf{q}_n} + \delta_{ij} \quad \text{for } j \in S_n.$$

Clearly, the $p_{ij}^n(a)$ are nonnegative and they sum up to one. Straightforward calculations show that the discounted reward optimality equation for $\mathcal{M}_n$ can be equivalently rewritten as

$$V_n^\alpha(i) = \max_{a \in A_n(i)} \left\{ \frac{r_n(i,a)}{\alpha + \mathbf{q}_n} + \frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n} \sum_{j \in S_n} p_{ij}^n(a) V_n^\alpha(j) \right\} \quad \text{for } i \in S_n.$$

This is the discounted reward optimality equation of a discrete-time finite state and action control model with state space S_n, action sets $A_n(i)$, reward function $r_n/(\alpha+\mathbf{q}_n)$, transition probabilities $p_{ij}^n(a)$, and discount factor $\frac{\mathbf{q}_n}{\alpha+\mathbf{q}_n} < 1$. Both the continuous-time and the discrete-time problems are equivalent, meaning that their optimal discounted value function is the same, and that they have the same sets of optimal deterministic stationary policies. The so-defined discrete-time control problem can be explicitly solved using, for instance, the value iteration algorithm, which consists in the successive application of a contraction operator; see [35] for more details on this algorithm.

We can also use the policy iteration algorithm for $\mathcal{M}_n$, described next, which converges to the optimal value and an optimal policy in a finite number of steps.

Step 0. Choose an arbitrary policy $f_0 \in \mathbb{F}_n$ and set $k = 0$.

Step 1. Determine the discounted reward v_k of f_k as the unique solution of the system of linear equations

$$\alpha v_k(i) = r_n(i, f_k) + \sum_{j \in S_n} q_{ij}^n(f_k) v_k(j) \quad \text{for } j \in S_n.$$

Step 2. Determine $f_{k+1} \in \mathbb{F}_n$ as the policy attaining the maximum

$$f_{k+1}(i) \in \operatorname*{argmax}_{a \in A_n(i)} \{ r_n(i,a) + \sum_{j \in S_n} q_{ij}^n(a) v_k(j) \} \quad \text{for } i \in S_n,$$

letting $f_{k+1}(i) = f_k(i)$ if possible.

Step 3. If $f_{k+1} = f_k$ then f_k is discount optimal for $\mathcal{M}_n$ and $v_k = V_n^\alpha$. Otherwise, increase k by one and go to Step 1.

This algorithm provides a sequence $\{v_k\}_{k \geq 0}$ that is (componentwise) increasing. If the algorithm does not terminate at some step $k \geq 0$, then there exists at least one $i \in S_n$ with $v_k(i) < v_{k+1}(i)$. The policy space $\mathbb{F}_n$ being finite, it follows that the algorithm converges necessarily in a finite number of steps.

Hence, the finite state and action truncated models $\mathcal{M}_n$ can be indeed explicitly solved.

The Lipschitz continuous case. It remains to study whether it is possible to obtain a rate of the convergence of $V_n^\alpha(i)$ to $V^\alpha(i)$, for $i \in S$. We will now show that this is indeed possible, although some of our assumptions on the control model $\mathcal{M}$ must be strengthened.

First of all, we prove the following useful result. In (2.3.1) below, note that the maximum is attained as a consequence of Corollary 2.1.6.

Lemma 2.3.3 *Suppose that the control model $\mathcal{M}$ satisfies Assumptions 2.1.2, 2.1.4, and 2.1.5. Suppose also that there exists some $u \in \mathcal{B}_w(S)$ and some nonnegative function $v : S \to [0, \infty)$ such that*

$$\Big| \alpha u(i) - \max_{a \in A(i)} \{ r(i,a) + \sum_{j \in S} q_{ij}(a) u(j) \} \Big| \leq v(i) \quad \textit{for all } i \in S, \tag{2.3.1}$$

where the function v satisfies

$$\sum_{j \in S} q_{ij}(a) v(j) \leq -c_v v(i) + b_v \quad \textit{for all } (i,a) \in \mathbb{K}$$

for some constants $c_v > -\alpha$ and $b_v \geq 0$. Under these conditions, we have

$$\big| V^\alpha(i) - u(i) \big| \leq \frac{v(i)}{\alpha + c_v} + \frac{b_v}{\alpha(\alpha + c_v)} \quad \textit{for each } i \in S.$$

Proof. Given any initial state $i \in S$ and an arbitrary Markov policy $\varphi \in \Phi$, it can be proved, as in (2.1.3), that

$$E^{i,\varphi}[v(x(t))] \leq e^{-c_v t} v(i) + \frac{b_v}{c_v}\big(1 - e^{-c_v t}\big) \quad \text{for all } t \geq 0, \tag{2.3.2}$$

or $E^{i,\varphi}[v(x(t))] \leq v(i) + b_v t$ in case that $c_v = 0$.

Let $f \in \mathbb{F}$ be such that (recall Corollary 2.1.6)

$$\max_{a \in A(i)} \big\{ r(i,a) + \sum_{j \in S} q_{ij}(a) u(j) \big\} = r(i,f) + \sum_{j \in S} q_{ij}(f) u(j) \quad \text{for all } i \in S,$$

and so, as a consequence of (2.3.1),

$$\alpha u(i) - r(i,f) - \sum_{j \in S} q_{ij}(f) u(j) \leq v(i) \quad \text{for all } i \in S.$$

By Dynkin's formula, for all $i \in S$ and $t \geq 0$ we have

$$\begin{aligned} E^{i,f}[e^{-\alpha t} u(x(t))] - u(i) &= E^{i,f}\Big[\int_0^t e^{-\alpha s} \big[-\alpha u(x(s)) + \sum_{j \in S} q_{x(s)j}(f) u(j) \big] ds \Big] \\ &\geq -E^{i,f}\Big[\int_0^t e^{-\alpha s} \big[r(x(s), f) + v(x(s)) \big] ds \Big]. \end{aligned}$$

By dominated and monotone convergence, we can take the limit as $t \to \infty$ to obtain, using (2.3.2) that

$$\begin{aligned} u(i) &\leq V^\alpha(i,f) + E^{i,f}\Big[\int_0^\infty e^{-\alpha s} v(x(s)) ds \Big] \\ &\leq V^\alpha(i,f) + \frac{v(i)}{\alpha + c_v} + \frac{b_v}{\alpha(\alpha + c_v)} \\ &\leq V^\alpha(i) + \frac{v(i)}{\alpha + c_v} + \frac{b_v}{\alpha(\alpha + c_v)} \end{aligned}$$

for each $i \in S$.

On the other hand, if $f^* \in \mathbb{F}$ is a discount optimal policy for $\mathcal{M}$, using (2.3.1) we obtain

$$-v(i) \leq \alpha u(i) - r(i,f^*) - \sum_{j \in S} q_{ij}(f^*) u(j) \quad \text{for all } i \in S.$$

Using the Dynkin formula for f^*, for all $i \in S$ and $t \geq 0$

$$\begin{aligned} E^{i,f^*}[e^{-\alpha t} u(x(t))] - u(i) &= E^{i,f^*}\Big[\int_0^t e^{-\alpha s} \big[-\alpha u(x(s)) + \sum_{j \in S} q_{x(s)j}(f^*) u(j) \big] ds \Big] \\ &\leq E^{i,f^*}\Big[\int_0^t e^{-\alpha s} \big[-r(x(s), f^*) + v(x(s)) \big] ds \Big]. \end{aligned}$$

Letting $t \to \infty$ and recalling that f^* is discount optimal we obtain, proceeding as before, that

$$u(i) \geq V^{\alpha}(i) - \frac{v(i)}{\alpha + c_v} - \frac{b_v}{\alpha(\alpha + c_v)} \quad \text{for each } i \in S.$$

The stated result readily follows. □

Remark 2.3.4 *As a side result, this lemma shows that if $u \in \mathcal{B}_w(S)$ is a solution of the discounted reward optimality equation for $\mathcal{M}$, then $u = V^{\alpha}$, the optimal discounted reward function. Indeed, if u is such solution, choose $v \equiv 0$ with $b_v = 0$, and the result follows.*

The next result shows an interesting fact about Lyapunov conditions. It states, roughly, that if a Lyapunov condition holds for some function, then it is also satisfied for its powers, provided that these are less than one.

Lemma 2.3.5 *Given a family of conservative and stable transition rates $\{q_{ij}(a)\}$, suppose that the function $h : S \to [0, \infty)$ satisfies $q(i) \leq h(i)$ for all $i \in S$. If there exists a power $\gamma > 0$ and a constant $c_\gamma \geq 0$ such that*

$$\sum_{j \in S} q_{ij}(a) h^{\gamma}(j) \leq c_\gamma h^{\gamma}(i) \quad \textit{for all } (i, a) \in \mathbb{K}, \tag{2.3.3}$$

then for every power $0 < \gamma' < \gamma$

$$\sum_{j \in S} q_{ij}(a) h^{\gamma'}(j) \leq c_\gamma h^{\gamma'}(i) \quad \textit{for all } (i, a) \in \mathbb{K}.$$

Proof: Fix $(i, a) \in \mathbb{K}$ and $\eta > 0$. Rewrite (2.3.3) as

$$\frac{1}{h(i) + \eta} \sum_{j \neq i} q_{ij}(a) h^{\gamma}(j) + \Big(\frac{q_{ii}(a)}{h(i) + \eta} + 1 \Big) h^{\gamma}(i) \leq \Big(\frac{c_\gamma}{h(i) + \eta} + 1 \Big) h^{\gamma}(i). \tag{2.3.4}$$

Define now

$$p_i = \frac{q_{ii}(a)}{h(i) + \eta} + 1 \quad \text{and} \quad p_j = \frac{q_{ij}(a)}{h(i) + \eta} \quad \text{for } j \neq i.$$

These coefficients are nonnegative (we use here the fact that $q(i) \leq h(i)$) and $\sum_{j \in S} p_j = 1$. Therefore, (2.3.4) is equivalent to

$$\sum_{j \in S} p_j h^{\gamma}(j) \leq \Big(\frac{c_\gamma}{h(i) + \eta} + 1 \Big) h^{\gamma}(i).$$

Using Jensen's inequality for the concave function $x \mapsto x^{\gamma'/\gamma}$ yields

$$\sum_{j \in S} p_j h^{\gamma'}(j) \leq \Big(\frac{c_\gamma}{h(i) + \eta} + 1 \Big)^{\gamma'/\gamma} h^{\gamma'}(i)$$

or, equivalently,

$$\sum_{j\in S} q_{ij}(a)h^{\gamma'}(j) \le h^{\gamma'}(i)\left(\Big(\frac{c_\gamma}{h(i)+\eta}+1\Big)^{\gamma'/\gamma}-1\right)(h(i)+\eta).$$

Since $0<\gamma'/\gamma<1$, we have

$$\Big(\frac{c_\gamma}{h(i)+\eta}+1\Big)^{\gamma'/\gamma}-1 \le \frac{c_\gamma}{h(i)+\eta},$$

and so

$$\sum_{j\in S} q_{ij}(a)h^{\gamma'}(j) \le c_\gamma h^{\gamma'}(i),$$

which completes the proof. □

As a consequence of this lemma, we have the following. Suppose that there exists a power $\gamma>0$ such that the Lyapunov function w, taken from Assumption 2.1.2, verifies for some constants $c_\gamma\in\mathbb{R}$ and $b_\gamma\ge 0$ the inequality

$$\sum_{j\in S} q_{ij}(a)w^{\gamma}(j) \le -c_\gamma w^{\gamma}(i)+b_\gamma \quad \text{for all } (i,a)\in\mathbb{K}. \tag{2.3.5}$$

We have also $\sum q_{ij}(a)w^{\gamma}(j)\le(|c_\gamma|+b_\gamma)w^{\gamma}(i)$, and so we can use Lemma 2.3.5 to derive that, for every $0<\gamma'<\gamma$,

$$\sum_{j\in S} q_{ij}(a)w^{\gamma'}(j) \le (|c_\gamma|+b_\gamma)w^{\gamma'}(i) \quad \text{for all } (i,a)\in\mathbb{K}, \tag{2.3.6}$$

and thus $w^{\gamma'}$ also verifies a Lyapunov condition as in (2.3.5):

$$\sum_{j\in S} q_{ij}(a)w^{\gamma'}(j) \le -c_{\gamma'}w^{\gamma'}(i)+b_{\gamma'} \quad \text{for all } (i,a)\in\mathbb{K},$$

with $c_{\gamma'}=-(|c_\gamma|+b_\gamma)$ and $b_{\gamma'}=0$.

Now we present our new conditions on the control model $\mathcal{M}$. Namely, we will assume, as before, that the control model $\mathcal{M}$ satisfies Assumptions 2.1.2 and 2.1.4, while Assumption 2.1.5 will be replaced with the following stronger condition.

Assumption 2.3.6 *The control model $\mathcal{M}$ verifies the following conditions.*

(i) The action sets $A(i)$ are compact for every $i\in S$.

(ii) The functions $a\mapsto q_{ij}(a)$ and $a\mapsto r(i,a)$ are Lipschitz continuous on $A(i)$ for all $i,j\in S$, that is,

$$|q_{ij}(a)-q_{ij}(a')|\le L_{ij}d_A(a,a') \quad \text{and} \quad |r(i,a)-r(i,a')|\le L_i d_A(a,a')$$

for all $i,j\in S$ and $a,a'\in A(i)$, and some constants $L_{ij}>0$ and $L_i>0$.

(iii) There are constants $\delta > 2$, $c_\delta > -\alpha$, and $b_\delta \geq 0$ with

$$\sum_{j \in S} q_{ij}(a) w^\delta(j) \leq -c_\delta w^\delta(i) + b_\delta \quad \textit{for all } (i,a) \in \mathbb{K}.$$

Part (i) of this assumption is the same as Assumption 2.1.5(i). Observe that the continuity of the transition and reward rates imposed in Assumption 2.1.5(ii) is now strengthened to Lipschitz continuity. The condition in Assumption 2.3.6(iii) above imposes a Lyapunov inequality on some power $\delta > 2$ of the Lyapunov function w. By the previous discussion (in particular, recall Lemma 2.3.5) this condition implies Assumption 2.1.5(iii). Therefore, it is indeed true that Assumption 2.3.6 is stronger that Assumption 2.1.5. Regarding the Lyapunov condition on w^δ above, note that it has the particular feature that the coefficient c_δ is supposed to be strictly larger than $-\alpha$; cf. Assumption 2.1.4(i).

We are now ready to state our main result on the convergence rates. Recall that, starting from the original control model $\mathcal{M}$, we construct the finite state and action truncated control models $\mathcal{M}_n$, as described in Section 2.3.1. Our next result shows that if we choose a "sufficiently fine" grid of actions $A_n(i)$ for the control model $\mathcal{M}_n$, measured by the Hausdorff distance between $A_n(i)$ and $A(i)$, then we can achieve a convergence rate of order $1/w^{\delta-2}$. For the next result, recall the notation $\mathfrak{M}$ used in (2.1.4) and (2.2.1).

Theorem 2.3.7 *Suppose that the control model $\mathcal{M}$ satisfies the Assumptions 2.1.2, 2.1.4, and 2.3.6, and suppose that the action sets of the finite state and action truncated models $\{\mathcal{M}_n\}_{n \geq 1}$ are chosen such that, for some constant $D > 0$ and every $n \geq 1$ and $i \in S_n$,*

$$\rho_A(A_n(i), A(i)) \leq \frac{D w^\delta(i)}{w^{\delta-2}(n) \cdot (L_i + 2\mathfrak{M} w(n) \sum_{j=0}^{n-1} L_{ij})}.$$

Then there exists a constant $\mathfrak{c} > 0$ such that for every $n \geq 1$ and $i \in S_n$

$$|V_n^\alpha(i) - V^\alpha(i)| \leq \mathfrak{c} \, \frac{w^\delta(i)}{w^{\delta-2}(n)}.$$

Proof. Fix $n \geq 1$ and $i \in S_n$. The discounted reward optimality equation for $\mathcal{M}$ at the state $i \in S_n$ reads

$$\alpha V^\alpha(i) = \max_{a \in A(i)} \{ r(i,a) + \sum_{j \in S} q_{ij}(a) V^\alpha(j) \}. \tag{2.3.7}$$

Thus, for every $a \in A(i)$ we have

$$\alpha V^\alpha(i) \geq r(i,a) + \sum_{j \in S} q_{ij}(a) V^\alpha(j) = r(i,a) + \sum_{j \in S} q_{ij}(a)(V^\alpha(j) - V^\alpha(n)). \tag{2.3.8}$$

Observe now that, since $||V^\alpha||_w \leq \mathfrak{M}$, recall (2.1.4),

$$\Big| \sum_{j > n} q_{ij}(a)(V^\alpha(j) - V^\alpha(n)) \Big| \leq 2\mathfrak{M} \sum_{j > n} q_{ij}(a) w(j).$$

By monotonicity of w we have the following inequality

$$\sum_{j>n} q_{ij}(a)w(j) \leq \frac{1}{w^{\delta-2}(n)} \sum_{j>n} q_{ij}(a)w^{\delta-1}(j).$$

As a consequence of Assumption 2.3.6(iii) and Lemma 2.3.5, we have the following Lyapunov inequality on $w^{\delta-1}$ (cf. (2.3.6))

$$\sum_{j\in S} q_{ij}(a)w^{\delta-1}(j) \leq (|c_\delta| + b_\delta)w^{\delta-1}(i) \quad \text{for all } (i,a) \in \mathbb{K}.$$

In particular,

$$\begin{aligned}\sum_{j>n} q_{ij}(a)w^{\delta-1}(j) &\leq -q_{ii}(a)w^{\delta-1}(i) + \sum_{j\in S} q_{ij}(a)w^{\delta-1}(j) \\ &\leq q(i)w^{\delta-1}(i) + (|c_\delta| + b_\delta)w^{\delta-1}(i) \\ &\leq (1 + |c_\delta| + b_\delta)w^{\delta}(i).\end{aligned}$$

Consequently, we obtain from (2.3.8) that for every $a \in A(i)$

$$\alpha V^\alpha(i) \geq r(i,a) + \sum_{j=0}^{n-1} q_{ij}(a)(V^\alpha(j) - V^\alpha(n)) - 2\mathfrak{M}\,\frac{(1 + |c_\delta| + b_\delta)w^\delta(i)}{w^{\delta-2}(n)}.$$

Now, if $a \in A_n(i) \subseteq A(i)$ we have, by definition of the control model $\mathcal{M}_n$, that $r(i,a) = r_n(i,a)$ and that $q_{ij}(a) = q^n_{ij}(a)$ if $0 \leq j < n$, while

$$-\sum_{j=0}^{n-1} q_{ij}(a) = q^n_{in}(a).$$

We have thus shown that for every $a \in A_n(i)$

$$\alpha V^\alpha(i) \geq r_n(i,a) + \sum_{j\in S_n} q^n_{ij}(a)V^\alpha(j) - 2\mathfrak{M}\,\frac{(1 + |c_\delta| + b_\delta)w^\delta(i)}{w^{\delta-2}(n)},$$

and therefore

$$\alpha V^\alpha(i) \geq \max_{a\in A_n(i)} \left\{r_n(i,a) + \sum_{j\in S_n} q^n_{ij}(a)V^\alpha(j)\right\} - 2\mathfrak{M}\,\frac{(1 + |c_\delta| + b_\delta)w^\delta(i)}{w^{\delta-2}(n)}. \tag{2.3.9}$$

Starting again from (2.3.7), let $a^* \in A(i)$ attain the maximum in that equation, so that

$$\alpha V^\alpha(i) = r(i,a^*) + \sum_{j\in S} q_{ij}(a^*)V^\alpha(j) = r(i,a^*) + \sum_{j\in S} q_{ij}(a^*)(V^\alpha(j) - V^\alpha(n)).$$

Proceeding as in the first part of this proof, we derive

$$\alpha V^\alpha(i) \leq r(i,a^*) + \sum_{j=0}^{n-1} q_{ij}(a^*)(V^\alpha(j) - V^\alpha(n)) + 2\mathfrak{M}\,\frac{(1+|c_\delta|+b_\delta)w^\delta(i)}{w^{\delta-2}(n)}. \tag{2.3.10}$$

By the Lipschitz continuity property in Assumption 2.3.6(ii), we have that the function

$$a \mapsto r(i,a) + \sum_{j=0}^{n-1} q_{ij}(a)(V^\alpha(j) - V^\alpha(n))$$

is Lipschitz continuous on $A(i)$ with Lipschitz constant $L_i + 2\mathfrak{M}w(n)\sum_{j=0}^{n-1} L_{ij}$. Therefore, if $a_n^* \in A_n(i)$ is such that

$$d_A(a_n^*, a^*) = \min_{a' \in A_n(i)} d_A(a', a^*) \leq \rho_A(A_n(i), A(i))$$

we have, from (2.3.10),

$$\begin{aligned}\alpha V^\alpha(i) &\leq r(i,a_n^*) + \sum_{j=0}^{n-1} q_{ij}(a_n^*)(V^\alpha(j) - V^\alpha(n)) \\ &+ \big(L_i + 2\mathfrak{M}w(n)\sum_{j=0}^{n-1} L_{ij}\big)\rho_A(A_n(i), A(i)) + 2\mathfrak{M}\,\frac{(1+|c_\delta|+b_\delta)w^\delta(i)}{w^{\delta-2}(n)}.\end{aligned}$$

Recalling our hypothesis on $\rho_A(A_n(i), A(i))$ and letting

$$\overline{C} = D + 2\mathfrak{M}(1+|c_\delta|+b_\delta)$$

yields, by definition of the reward and transition rates of $\mathcal{M}_n$, that

$$\begin{aligned}\alpha V^\alpha(i) &\leq r_n(i,a_n^*) + \sum_{j\in S_n} q_{ij}^n(a_n^*)V^\alpha(j) + \overline{C}\,\frac{w^\delta(i)}{w^{\delta-2}(n)} \\ &\leq \max_{a\in A_n(i)} \big\{r_n(i,a) + \sum_{j\in S_n} q_{ij}^n(a)V^\alpha(j)\big\} + \overline{C}\,\frac{w^\delta(i)}{w^{\delta-2}(n)}.\end{aligned}$$

Combining this inequality with (2.3.9) finally establishes that

$$\Big|\alpha V^\alpha(i) - \max_{a\in A_n(i)} \big\{r_n(i,a) + \sum_{j\in S_n} q_{ij}^n(a)V^\alpha(j)\big\}\Big| \leq \overline{C}\,\frac{w^\delta(i)}{w^{\delta-2}(n)} \quad \text{for all } i \in S_n.$$

Now we are going to use Lemma 2.3.3 for the control model $\mathcal{M}_n$. By construction of $\mathcal{M}_n$, recall in particular Proposition 2.3.2, the control model $\mathcal{M}_n$ satisfies the assumptions in Lemma 2.3.3. Observe now that $\{V^\alpha(j)\}_{j\in S_n} \in \mathcal{B}_w(S_n)$ plays the role of the function u,

while the function v in Lemma 2.3.3 is now $\overline{C}\,\frac{w^\delta(i)}{w^{\delta-2}(n)}$, for $i \in S_n$. By Assumption 2.3.6(iii) and arguing as in the proof of Proposition 2.3.2, it can be shown that

$$\sum_{j\in S_n} q_{ij}^n(a)\frac{\overline{C}w^\delta(j)}{w^{\delta-2}(n)} \leq -c_\delta\frac{\overline{C}w^\delta(i)}{w^{\delta-2}(n)} + \frac{\overline{C}b_\delta}{w^{\delta-2}(n)} \quad \text{for all } (i,a) \in \mathbb{K}_n.$$

We can indeed use Lemma 2.3.3 to show that, for every $i \in S_n$

$$|V^\alpha(i) - V_n^\alpha(i)| \leq \frac{\overline{C}w^\delta(i)}{w^{\delta-2}(n)(\alpha+c_\delta)} + \frac{\overline{C}b_\delta}{w^{\delta-2}(n)\alpha(\alpha+c_\delta)}.$$

Therefore, letting

$$\mathfrak{c} = \frac{\overline{C}(\alpha+b_\delta)}{\alpha(\alpha+c_\delta)}$$

we obtain the desired result. □

This theorem shows that, if we consider a fixed initial state $i \in S$, then the convergence rate of $V_n^\alpha(i)$ to $V^\alpha(i)$ is of order $1/w^{\delta-2}(n)$. Therefore, the convergence order is related to the maximal exponent $\delta > 2$ such that a Lyapunov condition

$$\sum_{j\in S} q_{ij}(a)w^\delta(j) \leq -c_\delta w^\delta(i) + b_\delta \quad \text{for all } (i,a) \in \mathbb{K}$$

with $c_\delta > -\alpha$ holds. Clearly, the larger we can find δ with this property, the faster the convergence.

An interesting fact in Theorem 2.3.7 is that the approximation error $|V_n^\alpha(i) - V^\alpha(i)|$ can be explicitly computed because it depends on the data of the original control model: it depends on the function w and on related constants. Therefore, the finite state and action truncations provide computable approximations with explicitly computable approximation errors.

Moreover, notice that the condition on the finite action sets $A_n(i)$, expressed in terms of $\rho_A(A_n(i), A(i))$ in Theorem 2.3.7, is parametrized in the numerator by $w^\delta(i)$; assuming that the Lipschitz constants of the reward and transition rates do not vary with i, this requires to have a "dense" grid of points $A_n(i)$ in $A(i)$ for small values of the state i, whereas this grid is allowed to be "sparse" for large values of i.

2.3.3 Finite truncations for average models

Now we are interested in the expected average reward optimality criterion for the control model $\mathcal{M}$. To approximate its optimal value and policies we consider the finite state and action truncations $\{\mathcal{M}_n\}_{n\geq 1}$ defined in Section 2.3.1.

The basic assumptions for the control model $\mathcal{M}$ under the average reward criterion are Assumptions 2.1.8, 2.1.9, and 2.1.11. Our next result explores whether these conditions are inherited by the truncated control models $\mathcal{M}_n$, that is, to check whether Assumption 2.2.8 is satisfied.

Proposition 2.3.8 *If the control model $\mathcal{M}$ satisfies Assumptions 2.1.8, 2.1.9, and 2.1.11, then the truncated control models $\mathcal{M}_n$ verify Assumption 2.2.8 except perhaps item (v), and $\mathcal{M}_n \to \mathcal{M}$.*

Proof. The fact that Assumptions 2.2.8(i)–(iv) are satisfied by the $\mathcal{M}_n$ is proved as in Proposition 2.3.2 and we omit the details. Concerning Assumption 2.2.8(v) observe that it is not possible to deduce that each policy in $\mathbb{F}_n$ is irreducible for $\mathcal{M}_n$ from irreducibility of policies in $\mathbb{F}$ for $\mathcal{M}$. Indeed, since the control model $\mathcal{M}_n$ consists in "restarting" the process at state n when it leaves $\{0, 1, \ldots, n\}$, the states in S_n need not communicate. For instance, suppose that for $\mathcal{M}$ the only way to go from 0 to 1 is by visiting $n+1$. Then, for $\mathcal{M}_n$, the states 0 and 1 will not communicate. □

From this proposition we deduce that the control model $\mathcal{M}_n$ might not have a constant optimal average reward, and that there might not exist solutions to its average reward optimality equation. This is an important departure point from the discounted case for which Proposition 2.3.2 indeed establishes that the optimal discounted rewards of $\mathcal{M}_n$ converged to those of $\mathcal{M}$.

By recovering the conditions in Theorems 2.2.9 and 2.2.10 we can, however, obtain convergence. For our next result, recall that g_n^* denotes the optimal gain of the control model $\mathcal{M}_n$.

Proposition 2.3.9 *Suppose that the control model $\mathcal{M}$ satisfies Assumptions 2.1.8, 2.1.9, and 2.1.11, and assume further that the finite state and action truncated models $\mathcal{M}_n$ are such that every policy in $\mathbb{F}_n$ is irreducible. Suppose that one of the conditions below hold:*

(a) Either there exist solutions $(g_n^*, h_n) \in \mathbb{R} \times \mathcal{B}_w(S_n)$ to the average reward optimality equation of $\mathcal{M}_n$ such that $\sup_{n\geq 1} ||h_n||_w < \infty$,

(b) Or there exist constants $\delta > 2$, $c_\delta > 0$, and $b_\delta \geq 0$ such that

$$\sum_{j\in S} q_{ij}(a) w^\delta(j) \leq -c_\delta w^\delta(i) + b_\delta \quad \text{for all } (i,a) \in \mathbb{K}.$$

Under these conditions, the optimal gains g_n^* of $\mathcal{M}_n$ converge to the optimal gain g^* of $\mathcal{M}$, and any limit policy of average optimal policies for $\mathcal{M}_n$ is average optimal for $\mathcal{M}$.

Proof. Under the additional irreducibility condition, we have that the finite state and action truncated control models $\mathcal{M}_n$ satisfy Assumption 2.2.8 with, moreover, $\mathcal{M}_n \to \mathcal{M}$; recall Proposition 2.3.8. If part (a) holds, then the result follows from Theorem 2.2.9. If part (b) is true, the we can proceed as in the proof of Proposition 2.3.2 to show that the condition in Theorem 2.2.10 is satisfied by the $\mathcal{M}_n$. □

As already mentioned, for practical purposes, the condition in (b) is easier to manage than (a) for a given control model $\mathcal{M}$.

Solving a finite average control problem.

To solve explicitly a finite state and action control problem with irreducible stationary policies, one can use the policy iteration algorithm. This algorithm has been analyzed in [14] and also in [31, Section 4.2.2]. We describe it next.

Step 0. Choose an arbitrary policy $f_0 \in \mathbb{F}_n$ and set $k = 0$.

Step 1. Determine the gain $g^{(k)} \in \mathbb{R}$ of f_k and a vector $h_k \in \mathcal{B}_w(S_n)$ such that $(g^{(k)}, h_k)$ is a solution to the Poisson equation for f_k

$$g^{(k)} = r_n(i, f_k) + \sum_{j \in S_n} q_{ij}^n(f_k) h_k(j) \quad \text{for } i \in S_n.$$

Step 2. Determine $f_{k+1} \in \mathbb{F}_n$ as the policy attaining the maximum

$$f_{k+1}(i) \in \operatorname*{argmax}_{a \in A_n(i)} \left\{ r_n(i,a) + \sum_{j \in S_n} q_{ij}^n(a) h_k(j) \right\} \quad \text{for } i \in S_n,$$

letting $f_{k+1}(i) = f_k(i)$ if possible.

Step 3. If $f_{k+1} = f_k$ then f_k is average optimal for $\mathcal{M}_n$ and $g^{(k)} = g_n^*$. Otherwise, increase k by one and go to Step 1.

The sequence $\{g^{(k)}\}_{k \geq 0}$ is monotone nondecreasing and if the algorithm does not stop at step k then $g^{(k)} < g^{(k+1)}$. The policy space $\mathbb{F}_n$ being finite, the algorithm necessarily terminates by solving the average reward optimality equation for $\mathcal{M}_n$ and finding an average optimal policy.

The Lipschitz continuous case. Now we aim at obtaining rates of convergence for the convergence to the optimal gain of the original control model $\mathcal{M}$. To do so, we need to impose stronger assumptions on our control model. We begin with our next result, which is an average reward version of Lemma 2.3.3. It does not need to assume that stationary policies are irreducible.

Lemma 2.3.10 *Let the control model $\mathcal{M}$ satisfy Assumptions 2.1.8, 2.1.9, and 2.1.11(i)–(iii). Suppose that there exists a pair $(g, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ and a nonnegative function $u : S \to [0, \infty)$ such that*

$$\left| g - \max_{a \in A(i)} \left\{ r(i,a) + \sum_{j \in S} q_{ij}(a) h(j) \right\} \right| \leq u(i) \quad \textit{for all } (i,a) \in \mathbb{K},$$

where u is such that there exist constants $c_u > 0$ and $b_u \geq 0$ with

$$\sum_{j \in S} q_{ij}(a) u(j) \leq -c_u u(i) + b_u \quad \textit{for all } (i,a) \in \mathbb{K}.$$

Under these conditions,

$$|J(i) - g| \leq \frac{b_u}{c_u} \quad \textit{for all } i \in S.$$

Proof. Given arbitrary $\varphi \in \Phi$, $i \in S$, and $t \geq 0$, we have

$$r(t,i,\varphi) + \sum_{j \in S} q_{ij}(t,\varphi) h(j) \leq g + u(i).$$

Using Dynkin's formula for the function $h \in \mathcal{B}_w(S)$ we obtain

$$\begin{aligned} E^{i,\varphi}[h(x(t))] - h(i) &= E^{i,\varphi}\Big[\int_0^t \sum_{j \in S} q_{x(s)j}(s,\varphi) h(j) ds \Big] \\ &\leq gt + E^{i,\varphi}\Big[\int_0^t u(x(s)) ds \Big] - E^{i,\varphi}\Big[\int_0^t r(s,x(s),\varphi) ds \Big]. \end{aligned}$$

Dividing by $t > 0$ and rearranging terms yields

$$\frac{1}{t} E^{i,\varphi}\Big[\int_0^t r(s,x(s),\varphi) ds \Big] + \frac{1}{t}\Big(E^{i,\varphi}[h(x(t))] - h(i) \Big) \leq g + \frac{1}{t} E^{i,\varphi}\Big[\int_0^t u(x(s)) ds \Big]. \quad (2.3.11)$$

Observe that $\big| E^{i,\varphi}[h(x(t))] \big| \leq ||h||_w E^{i,\varphi}[w(x(t))] \leq ||h||_w \big(e^{-c_1 t} w(i) + b_1/c_1 \big)$, as a consequence of (2.1.3). Therefore, we have

$$\lim_{t \to \infty} \frac{1}{t}\Big(E^{i,\varphi}[h(x(t))] - h(i) \Big) = 0.$$

By (2.1.3) applied to the function u we have $E^{i,\varphi}[u(x(s))] \leq e^{-c_u s} + b_u/c_u$, and so

$$\limsup_{t \to \infty} \frac{1}{t} E^{i,\varphi}\Big[\int_0^t u(x(s)) ds \Big] \leq \frac{b_u}{c_u}.$$

Taking the lim sup as $t \to \infty$ in (2.3.11) gives $J(i,\varphi) \leq g + b_u/c_u$ for every $i \in S$ and $\varphi \in \Phi$. Therefore, for each $i \in S$ we have $J(i) \leq g + b_u/c_u$.

Choose now a policy $f \in \mathbb{F}$ such that

$$g - u(i) \leq r(i,f) + \sum_{j \in S} q_{ij}(f) h(j) \quad \text{for all } i \in S.$$

Proceeding as in the first part of this proof we can show that $J(i) \geq J(i,f) \geq g - b_u/c_u$ for each $i \in S$. The stated result follows. □

Now we give our additional condition on the control model $\mathcal{M}$. This condition is stronger than Assumption 2.1.11 and will henceforth replace it. It is inspired from Assumption 2.3.6.

Assumption 2.3.11 *The control model $\mathcal{M}$ verifies the following conditions.*

(i) The action sets $A(i)$ are compact for every $i \in S$.

(ii) The functions $a \mapsto q_{ij}(a)$ and $a \mapsto r(i,a)$ are Lipschitz continuous on $A(i)$ for all $i,j \in S$, that is,

$$|q_{ij}(a) - q_{ij}(a')| \leq L_{ij} d_A(a,a') \quad \text{and} \quad |r(i,a) - r(i,a')| \leq L_i d_A(a,a')$$

for all $i,j \in S$ and $a,a' \in A(i)$, and some constants $L_{ij} > 0$ and $L_i > 0$.

(iii) There are constants $\delta > 2$, $c_\delta > 0$, and $b_\delta \geq 0$ with

$$\sum_{j \in S} q_{ij}(a) w^\delta(j) \leq -c_\delta w^\delta(i) + b_\delta \quad \text{for all } (i,a) \in \mathbb{K}.$$

(iv) Each deterministic stationary policy in $\mathbb{F}$ is irreducible.

We can use Lemma 2.3.5 to show that item (iii) implies Assumption 2.1.11(iii), so that indeed Assumption 2.3.11 is stronger than Assumption 2.1.11.

This is our main result on the convergence rates to g^*. Since we are not assuming irreducibility of the policies in $\mathbb{F}_n$, the optimal gain g_n^* of $\mathcal{M}_n$ needs not exist, and hence our result refers to the optimal average reward $J_n(i)$ for $i \in S_n$.

Theorem 2.3.12 *Suppose that the control model $\mathcal{M}$ satisfies the Assumptions 2.1.8, 2.1.9, and 2.3.11, and suppose that the action sets of the finite state and action truncated models $\{\mathcal{M}_n\}_{n \geq 1}$ are chosen such that, for some constant $D > 0$ and every $n \geq 1$ and $i \in S_n$,*

$$\rho_A(A_n(i), A(i)) \leq \frac{D w^\delta(i)}{w^{\delta-2}(n) \cdot (L_i + 2\frac{RM}{\gamma} w(n) \sum_{j=0}^{n-1} L_{ij})}.$$

Then there exists a constant $\mathfrak{c} > 0$ such that for every $n \geq 1$

$$\max_{i \in S_n} |J_n(i) - g^*| \leq \frac{\mathfrak{c}}{w^{\delta-2}(n)}.$$

Proof. Let $(g^*, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ be a solution to the average reward optimality equation for the control model $\mathcal{M}$ with $||h||_w \leq RM/\gamma$; recall Corollary 2.1.15. Let $n \geq 1$ and $i \in S_n$ and write the average reward optimality for the control model $\mathcal{M}$ at state $i \in S_n$:

$$\begin{aligned} g^* &= \max_{a \in A(i)} \big\{ r(i,a) + \sum_{j \in S} q_{ij}(a) h(j) \big\} \\ &= \max_{a \in A(i)} \big\{ r(i,a) + \sum_{j \in S} q_{ij}(a)(h(j) - h(n)) \big\}. \end{aligned} \tag{2.3.12}$$

Observe now that for all $a \in A(i)$

$$\Big| \sum_{j > n} q_{ij}(a)(h(j) - h(n)) \Big| \leq \frac{2RM}{\gamma} \sum_{j > n} q_{ij}(a) w(j).$$

Arguing as in the proof of Theorem 2.3.7, we obtain

$$\sum_{j>n} q_{ij}(a)w(j) \leq \frac{1}{w^{\delta-2}(n)} \sum_{j>n} q_{ij}(a)w^{\delta-1}(j).$$

By Assumption 2.3.11(iii) and Lemma 2.3.5,

$$\sum_{j\in S} q_{ij}(a)w^{\delta-1}(j) \leq (c_\delta + b_\delta)w^{\delta-1}(i)$$

and so

$$\sum_{j>n} q_{ij}(a)w^{\delta-1}(j) \leq (1 + c_\delta + b_\delta)w^{\delta}(i).$$

Therefore, from (2.3.12), for every $a \in A(i)$,

$$g^* \geq r(i,a) + \sum_{j=0}^{n-1} q_{ij}(a)(h(j) - h(n)) - \frac{2RM}{\gamma w^{\delta-2}(n)}(1 + c_\delta + b_\delta)w^{\delta}(i).$$

Suppose now that $a \in A_n(i) \subseteq A(i)$. Recalling the definition of the transition and reward rates of the truncated control model, the above inequality can be written

$$g^* \geq r_n(i,a) + \sum_{j\in S_n} q^n_{ij}(a)h(j) - \frac{2RM}{\gamma w^{\delta-2}(n)}(1 + c_\delta + b_\delta)w^{\delta}(i),$$

so that

$$g^* \geq \max_{a\in A_n(i)} \left\{ r_n(i,a) + \sum_{j\in S_n} q^n_{ij}(a)h(j) \right\} - \frac{2RM}{\gamma w^{\delta-2}(n)}(1 + c_\delta + b_\delta)w^{\delta}(i). \tag{2.3.13}$$

Proceeding with the proof, let $a^* \in A(i)$ attain the maximum in (2.3.12), that is,

$$g^* = r(i,a^*) + \sum_{j\in S} q_{ij}(a^*)(h(j) - h(n)).$$

As before, we obtain

$$g^* \leq r(i,a^*) + \sum_{j=0}^{n-1} q_{ij}(a^*)(h(j) - h(n)) + \frac{2RM}{\gamma w^{\delta-2}(n)}(1 + c_\delta + b_\delta)w^{\delta}(i).$$

By Assumption 2.3.11(ii), the function

$$a \mapsto r(i,a) + \sum_{j=0}^{n-1} q_{ij}(a)(h(j) - h(n))$$

is Lipschitz continuous on $A(i)$ with Lipschitz constant $L_i + \frac{2RM}{\gamma} w(n) \sum_{j=0}^{n-1} L_{ij}$. Consequently, letting $a_n^* \in A_n(i)$ given by

$$d_A(a_n^*, a^*) = \min_{a' \in A_n(i)} d_A(a', a^*) \leq \rho_A(A_n(i), A(i)),$$

we obtain

$$g^* \leq r(i, a_n^*) + \sum_{j=0}^{n-1} q_{ij}(a_n^*)(h(j) - h(n)) + \frac{2RM}{\gamma w^{\delta-2}(n)}(1 + c_\delta + b_\delta) w^\delta(i)$$
$$+ \Big(L_i + \frac{2RM}{\gamma} w(n) \sum_{j=0}^{n-1} L_{ij} \Big) \rho_A(A_n(i), A(i)).$$

By the condition on $\rho_A(A_n(i), A(i))$, letting

$$\overline{C} = D + \frac{2RM}{\gamma}(1 + c_\delta + b_\delta)$$

we obtain

$$g^* \leq r(i, a_n^*) + \sum_{j=0}^{n-1} q_{ij}(a_n^*)(h(j) - h(n)) + \overline{C} \, \frac{w^\delta(i)}{w^{\delta-2}(n)}$$

or, recalling the definition of r_n and q^n, that

$$g^* \leq r_n(i, a_n^*) + \sum_{j \in S_n} q_{ij}^n(a_n^*) h(j) + \overline{C} \, \frac{w^\delta(i)}{w^{\delta-2}(n)}.$$

Recalling the inequality (2.3.13), we have thus established that

$$\Big| g^* - \max_{a \in A_n(i)} \big\{ r_n(i, a) + \sum_{j \in S_n} q_{ij}^n(a) h(j) \big\} \Big| \leq \overline{C} \, \frac{w^\delta(i)}{w^{\delta-2}(n)} \quad \text{for every } i \in S_n.$$

We apply now Lemma 2.3.10 to the control model $\mathcal{M}_n$ for the function

$$i \mapsto \overline{C} \, \frac{w^\delta(i)}{w^{\delta-2}(n)};$$

indeed, note that $\mathcal{M}_n$ satisfies the conditions given in that lemma (recall Proposition 2.3.8) because Lemma 2.3.10 did not assume irreducibility of stationary policies. Hence, letting $\mathfrak{c} = \overline{C} b_\delta / c_\delta$, for all $i \in S_n$ we get

$$|J_n(i) - g^*| \leq \frac{\mathfrak{c}}{w^{\delta-2}(n)}.$$

The proof is now complete. □

This result shows that the optimal average reward of the control model $\mathcal{M}_n$ converges to g^*, the optimal gain of the original control model $\mathcal{M}$, at a rate $1/w^{\delta-2}(n)$. Therefore, the convergence rate is related to the maximal exponent $\delta > 2$ for which a Lyapunov condition of the form

$$\sum_{j \in S} q_{ij}(a) w^{\delta}(j) \leq -c_{\delta} w^{\delta}(i) + b_{\delta} \quad \text{for all } (i,a) \in \mathbb{K}$$

holds, with the condition that $c_{\delta} > 0$. Interestingly, the convergence of $J_n(i)$ to g^* is uniform in $i \in S_n$ as $n \to \infty$.

Observe also that the constant $\mathfrak{c}$ in the convergence rate depends on the data of the original control model $\mathcal{M}$, except perhaps for the constants R and γ from exponential ergodicity.

2.4 Applications

Now we show some applications of the results in the previous section. We study a discounted reward problem in Section 2.4.1 and an average reward problem in Section 2.4.2.

2.4.1 A population system with catastrophes

Our next example is a generalization of the population system proposed in [17, Example 7.2]; see also [33, Section IV].

We describe the elements of the controlled population system $\mathcal{M}$. The state space is $S = \{0, 1, 2, \ldots\}$, which stands for the size of the population. The natural birth and death rates of the population are $\lambda > 0$ and $\mu > 0$, respectively.

We suppose that the decision-maker controls the immigration rate a taking values in the interval $a \in [0, a_2]$, for some $a_2 > 0$. Also, when the population size is $i \geq 1$, we assume that a catastrophe occurs at a rate $d(i,b) \geq 0$, which is is controlled by an action $b \in [b_1, b_2]$ chosen by the controller. Therefore, the action space is $A = [0, a_2] \times [b_1, b_2]$ and the action sets are

$$A(0) = [0, a_2] \times \{b_1\} \quad \text{and} \quad A(i) = [0, a_2] \times [b_1, b_2] \quad \text{for } i \geq 1.$$

Note that when the population size is 0, the controller does not take any action regarding the catastrophes: this is represented by the "void" action b_1. To have a unified notation, we define nevertheless $d(0,b) = 0$ for all $b \in [b_1, b_2]$.

If a catastrophe occurs when the size of the population is $i \geq 1$, we denote by $\gamma_i(j)$, for $1 \leq j \leq i$, the probability that j individuals perish in the catastrophe. We must have

$$\sum_{j=1}^{i} \gamma_i(j) = 1 \quad \text{for each } i > 0.$$

Let us now define the transition rates of the system. In state $i = 0$ they are given by

$$q_{01}(a,b) = a = -q_{00}(a,b) \quad \text{for all } (a,b) \in A(0),$$

while for $i > 0$ and $(a, b) \in A$ they are given by

$$q_{ij}(a,b) = \begin{cases} 0 & \text{for } j > i+1, \\ \lambda i + a & \text{for } j = i+1, \\ -(\lambda+\mu)i - a - d(i,b) & \text{for } j = i, \\ \mu i + d(i,b)\gamma_i(1) & \text{for } j = i-1, \\ d(i,b)\gamma_i(i-j) & \text{for } 0 \le j < i-1. \end{cases}$$

When the population size is $i \in S$, the controller receives a reward at a rate $p \cdot i$ for some $p > 0$. The cost rate for controlling the immigration and the catastrophe rates is $c(i, a, b)$, for $(i, a, b) \in \mathbb{K}$. We will thus consider the net reward rate

$$r(i,a,b) = p \cdot i - c(i,a,b) \quad \text{for } (i,a,b) \in \mathbb{K}.$$

The rewards earned by the controller at depreciated at the constant discount rate $\alpha > 0$.

Assumption 2.4.1 *The controlled population system $\mathcal{M}$ verifies the following conditions.*

(i) There exists constants $0 \le d_m < d_M$ such that $d_m \cdot i \le d(i,b) \le d_M \cdot i$ for all $i \in S$ and $b \in [b_1, b_2]$.

(ii) For some constant c_M we have $|c(i,a,b)| \le c_M(i+1)$ for all $(i,a,b) \in \mathbb{K}$.

(iii) The functions $d(i,b)$ and $c(i,a,b)$ are continuous in a and b for each $i \in S$.

Note that part (ii) in this assumption indeed implies that $d(0, b) = 0$ for $b \in [b_1, b_2]$. We choose the Lyapunov function w of the form $w(i) = R \cdot (i + 1)$ for $i \in S$, where the constant R satisfies

$$R \ge \max\{1, \lambda + \mu + a_2 + d_M\}.$$

Here is our first result on the control model $\mathcal{M}$.

Proposition 2.4.2 *If the discount rate $\alpha > 0$ verifies $\alpha > \lambda - \mu - d_m$ then the controlled population system $\mathcal{M}$ satisfies Assumptions 2.1.2, 2.1.4, and 2.1.5.*

Proof. By its definition, it is clear that w is a Lyapunov function on S. Also by construction, it satisfies $q(i) \le w(i)$ for each $i \in S$ because $q(i) \le (\lambda + \mu + d_M)i + a_2$ for all $i \in S$.

A direct calculation shows that, for all $(i, a, b) \in \mathbb{K}$

$$\sum_{j \in S} q_{ij}(a,b)w(j) = (\lambda - \mu)w(i) + R(\mu - \lambda + a) - Rd(i,b)\sum_{k=0}^{i-1}(i-k)\gamma_i(i-k).$$

(Note that the above sum is not defined when $i = 0$, but in this case the factor $d(i, b)$ vanishes.) Since $d(i,b)\sum_{k=0}^{i-1}(i-k)\gamma_i(i-k) \ge d_m i$ it follows that

$$\sum_{j \in S} q_{ij}(a,b)w(j) \le (\lambda - \mu - d_m)w(i) + R(\mu - \lambda + a_2 + d_m) \quad \text{for all } (i,a,b) \in \mathbb{K}.$$

Therefore, Assumption 2.1.2 holds for

$$c_1 = \mu - \lambda + d_m \quad \text{and} \quad b_1 = R|\mu - \lambda + a_2 + d_m|.$$

The condition $\alpha + c_1 > 0$ is indeed satisfied and so Assumption 2.1.4(i) holds. It is also easily seen that Assumption 2.1.4(ii) is satisfied.

Regarding Assumption 2.1.5, parts (i) and (ii) follow directly from our condition on the population system $\mathcal{M}$. Concerning part (iii), it can be shown that for all $(i, a, b) \in \mathbb{K}$, the quantity $\sum_{j\in S} q_{ij}(a, b)w^2(j)$ equals

$$2(\lambda - \mu)w^2(i) + R(2a - \lambda + 3\mu)w(i) + R^2(a - \lambda - \mu) - d(i, b)\sum_{k=0}^{i-1}[(w^2(i) - w^2(k)]\gamma_i(i - k)$$

and thus

$$\begin{aligned}\sum_{j\in S} q_{ij}(a, b)w^2(j) &\le 2(\lambda - \mu)w^2(i) + R(2a_2 - \lambda + 3\mu)w(i) + R^2(a_2 - \lambda - \mu)\\ &= w^2(i)\Big(2(\lambda - \mu) + \frac{1}{w(i)}R(2a_2 - \lambda + 3\mu) + \frac{1}{w^2(i)}R^2(a_2 - \lambda - \mu)\Big).\end{aligned}$$

Given $\epsilon > 0$, choose i_0 large enough such that $i \ge i_0$ implies

$$\frac{1}{w(i)}R(2a_2 - \lambda + 3\mu) + \frac{1}{w^2(i)}R^2(a_2 - \lambda - \mu) < \epsilon.$$

Fur such $i \ge i_0$ and $(a, b) \in A(i)$ we have

$$\sum_{j\in S} q_{ij}(a, b)w^2(j) \le \big(2(\lambda - \mu) + \epsilon\big)w^2(i),$$

while choosing the constant $b_2 \ge 0$ large enough, we have for all $0 \le i < i_0$ and $(a, b) \in A(i)$ that

$$\sum_{j\in S} q_{ij}(a, b)w^2(j) \le \big(2(\lambda - \mu) + \epsilon\big)w^2(i) + b_2,$$

thus showing the condition in Assumption 2.1.5(iii). □

As a consequence, if we consider the finite state and action truncations of $\mathcal{M}$, as defined in Section 2.3.1, we can use Proposition 2.3.2 to conclude that, for every initial state $i \in S$, the optimal discounted value $V_n^\alpha(i)$ of $\mathcal{M}_n$ converges to the optimal discounted reward $V^\alpha(i)$, and also that limit policies of discount optimal policies for $\mathcal{M}_n$ are optimal for $\mathcal{M}$. This holds for every discount rate α with $\alpha > \lambda - \mu - d_m$.

Under some additional conditions, we can obtain an explicit convergence rate.

Proposition 2.4.3 *Let the controlled population system $\mathcal{M}$ verify Assumption 2.4.1 and suppose, in addition, that the functions $c(i, a, b)$ and $d(i, b)$ are Lipschitz continuous in a and b for every fixed $i \in S$. Under these conditions,*

(i) If $\mu \geq \lambda$ then for every discount rate $\alpha > 0$ and every $\gamma > 0$ we can adequately choose the action sets for the truncated finite state and action control models $\mathcal{M}_n$ so that

$$|V_n^\alpha(i) - V^\alpha(i)| = \mathrm{O}(n^{-\gamma}) \quad \text{as } n \to \infty \text{ for all } i \in S.$$

(ii) If $\mu < \lambda$ then for every discount rate α with $\alpha > 2(\lambda - \mu)$ and every $0 < \gamma < \frac{\alpha}{\lambda-\mu} - 2$, we can adequately choose the action sets for the truncated finite state and action control models $\mathcal{M}_n$ so that

$$|V_n^\alpha(i) - V^\alpha(i)| = \mathrm{O}(n^{-\gamma}) \quad \text{as } n \to \infty \text{ for all } i \in S.$$

Proof. Fix an arbitrary power $\beta > 2$ and consider the Lyapunov function $i \mapsto w^\beta(i)$. Writing down the expression for $\sum_{j\in S} q_{ij}(a,b)w^\beta(j)$ as a series of powers of $i+1$, i.e., using expressions such as

$$(i+2)^\beta = (i+1)^\beta + \beta(i+1)^{\beta-1} + \mathrm{O}((i+1)^{\beta-2}) \quad \text{or} \quad i^\beta = (i+1)^\beta - \beta(i+1)^{\beta-1} + \mathrm{O}((i+1)^{\beta-2}),$$

it can be shown that for all $(i,a,b) \in \mathbb{K}$

$$\begin{aligned} \sum_{j\in S} q_{ij}(a,b)w^\beta(j) &\leq \beta(\lambda-\mu)w^\beta(i) + \mathrm{O}((i+1)^{\beta-1}) \\ &= w^\beta(i)\Big(\beta(\lambda-\mu) + \frac{\mathrm{O}((i+1)^{\beta-1})}{w^\beta(i)}\Big). \end{aligned}$$

Now we proceed as in the proof of Proposition 2.4.2. For every $\epsilon > 0$ there exists some $i_0 \in S$ such that the expression within parentheses is less than $\beta(\lambda-\mu)+\epsilon$ for all $i \geq i_0$ and, therefore, for such $i \geq i_0$ and $(a,b) \in A(i)$ we have

$$\sum_{j\in S} q_{ij}(a,b)w^\beta(j) \leq (\beta(\lambda-\mu)+\epsilon)w^\beta(i).$$

Finally, by choosing $b_\beta \geq 0$ large enough we obtain

$$\sum_{j\in S} q_{ij}(a,b)w^\beta(j) \leq (\beta(\lambda-\mu)+\epsilon)w^\beta(i) + b_\beta \quad \text{for all } (i,a,b) \in \mathbb{K}. \tag{2.4.1}$$

Therefore, a Lyapunov condition for w^β holds for every $\beta > 2$. We will now obtain the convergence rates by checking Assumption 2.3.6(iii).

Consider the case $\mu \geq \lambda$. Let $\alpha > 0$ be the discount rate and choose any $\gamma > 0$. Let $\beta = \gamma + 2$ and let $0 < \epsilon < \alpha$. Then we indeed have the condition in Assumption 2.3.6(iii) because $\beta(\lambda-\mu)+\epsilon < \alpha$; recall (2.4.1). Therefore, by Theorem 2.3.7, we can choose the action sets of $\mathcal{M}_n$ to achieve an $\mathrm{O}(n^{-\gamma})$ approximation error.

Consider now the case $\mu < \lambda$ and let α be a discount rate with $\alpha > 2(\lambda-\mu)$. For any $0 < \gamma < \frac{\alpha}{\lambda-\mu} - 2$, let $\beta = \gamma + 2$, so that $2 < \beta < \frac{\alpha}{\lambda-\mu}$. Choose $\epsilon > 0$ with $\epsilon < \alpha - \beta(\lambda-\mu)$. The condition in Assumption 2.3.6(iii) holds (recall (2.4.1)) and we can determine the

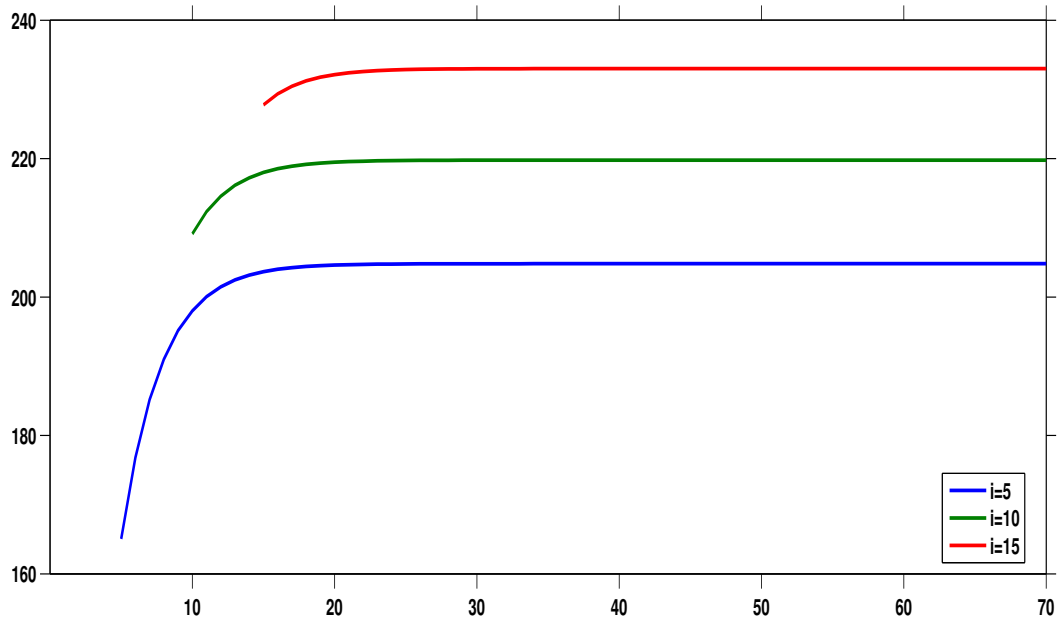


Figure 2.1: The optimal discounted rewards $V_n^\alpha(i)$ for $i = 5, 10, 15$.

action sets of $\mathcal{M}_n$ so as to obtain a convergence order of $\mathrm{O}(n^{-\gamma})$. □

Numerical experimentation. We fix the values of the parameters

$$\lambda = 3.05, \quad \mu = 3, \quad a_2 = 5, \quad b_1 = 5, \quad b_2 = 8.$$

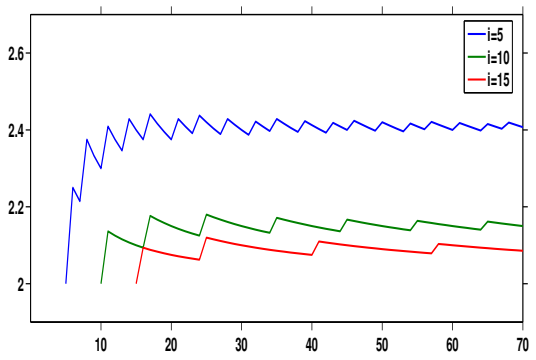
The catastrophe rate is given by $d(i, b) = ib/10$ for $i > 0$ and $b \in [5, 8]$. The distribution $\{\gamma_i(j)\}$ of the catastrophe size is a truncated geometric distribution with parameter $\gamma = 0.8$; more precisely, given $i > 0$,

$$\gamma_i(j) = \frac{\gamma^{j-1}(1-\gamma)}{1-\gamma^i} \quad \text{for } 1 \le j \le i.$$

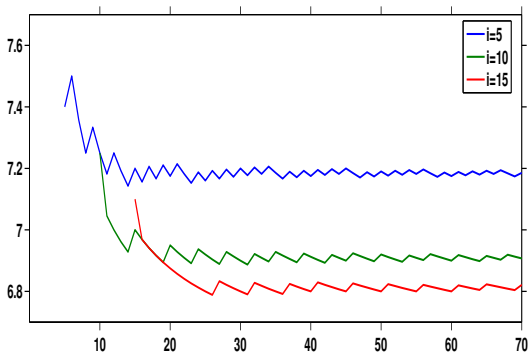
Finally, the net reward rate is

$$r(i, a, b) = \left(10 - (a-2)^2 - 0.5(b-8)^2\right) i.$$

The interpretation of the term $(a-2)^2$ is that we suppose that there is a natural immigration rate (which equals 2), and that augmenting or diminishing this natural immigration rate implies a cost for the controller. Similarly, the term $(b-8)^2$ means that there is a natural catastrophe rate (which equals 8), and the controller incurs a cost when decreasing it. The discount rate is $\alpha = 0.1$. Note that we are indeed under the conditions of Proposition 2.4.2, and so the optimal value and optimal policies of $\mathcal{M}_n$ converge to those of $\mathcal{M}$.



(a) The optimal actions $a_n^*(i)$



(b) The optimal actions $b_n^*(i)$

Figure 2.2: The optimal policies $f_n^*(i)$ for $i = 5, 10, 15$

For every $n \geq 1$, we consider the truncated control model $\mathcal{M}_n$ with state space $S_n = \{0, 1, \ldots, n\}$ and action sets

$$A_n(0) = \left\{ \frac{\ell_1 a_2}{2n} : 0 \leq \ell_1 \leq 2n \right\} \times \{b_1\}$$

and

$$A_n(i) = \left\{ \left(\frac{\ell_1 a_2}{2n} \ , \ b_1 + \frac{\ell_2}{2n}(b_2 - b_1) \right) : 0 \leq \ell_1, \ell_2 \leq 2n \right\} \quad \text{for } i > 0.$$

In particular, the actions sets $A_n(i)$ verify the Hausdorff convergence property given in Definition 2.2.1.

For every $1 \leq n \leq 70$, we solved the discounted control problem for $\mathcal{M}_n$. Given the initial states $i = 5, 10, 15$, the discount optimal rewards $V_n^\alpha(i)$ and the optimal actions actions $a_n^*(i)$ and $b_n^*(i)$ for $\mathcal{M}_n$ are displayed in Figures 2.1 and 2.2, respectively, as functions of n.

Empirically, we observe that the optimal reward and actions quickly converge, and become stable for relatively small values of n. We observe that the graph in Figure 2.1 is smoother, whereas the actions displayed in Figure 2.2 are somehow saw-shaped. This should not be surprising because $V_n^\alpha(i)$ is obtained after some kind of "averaging", while the actions shown in Figure 2.2 take values proportional to $1/2n$ for $1 \leq n \leq 70$.

2.4.2 A controlled birth-and-death process

We consider the control birth-and-death system studied in [17, Example 7.1] under the average reward optimality criterion.

The state space is $S = \{0, 1, \ldots\}$ and the state variable stands for the size of the population. The population's natural birth rate is $\lambda > 0$, while the death rate is assumed to be controlled by the decision-maker. More precisely, we consider the compact action space $A = [\mu_1, \mu_2]$, for $0 < \mu_1 < \mu_2$, with $A(i) = A$ for all $i \in S$. So, the death rate corresponds

to an action $a \in A$ chosen by the controller. In addition, we suppose that, when the population decreases, either one or two individuals die, with respective probabilities p_1 and p_2. In order to avoid extinction, we assume also that, when the population size is 0, the birth rate is given also by λ, which is then interpreted as an immigration rate.

Therefore, the transition rates of the original control model $\mathcal{M}$ are, for $i = 0$,

$$q_{01}(a) = \lambda = -q_{00}(a) \quad \text{for } a \in A,$$

whereas, for $i = 1$,

$$q_{10}(a) = a,\ q_{11}(a) = -(a + \lambda),\ q_{12}(a) = \lambda \quad \text{for } a \in A.$$

For $i \geq 2$, the transition rates of the system are given by

$$q_{ij}(a) = \begin{cases} p_2 ai & \text{for } j = i - 2, \\ p_1 ai & \text{for } j = i - 1, \\ -(a + \lambda)i & \text{for } j = i, \\ \lambda i & \text{for } j = i + 1, \\ 0 & \text{otherwise,} \end{cases}$$

for each $a \in A$, with $p_1, p_2 \geq 0$ and $p_1 + p_2 = 1$.

We suppose that the controller receives a reward $p \cdot i$ per time unit when the population size is $i \in S$, where $p > 0$ is a given constant. Moreover, we suppose that the cost rate when taking the action $a \in A$ in state $i \in S$ is $c(i, a)$. Thus, the decision-maker considers the net reward rate function

$$r(i, a) = p \cdot i - c(i, a) \quad \text{for all } (i, a) \in \mathbb{K}.$$

Next we state our assumptions on this controlled model.

Assumption 2.4.4 *The controlled birth-and-death system $\mathcal{M}$ verifies the following conditions. For some constant c_M we have $|c(i, a)| \leq c_M(i + 1)$ for all $(i, a) \in \mathbb{K}$, and the function $c(i, a)$ is continuous in $a \in A$ for every $i \in S$.*

We check our hypotheses on this control model.

Proposition 2.4.5 *If the controlled birth-and-death system verifies Assumption 2.4.4 and $\mu_1(1 + p_2) > \lambda$, then the Assumptions 2.1.8, 2.1.9, and 2.1.11 hold. If, in addition, we have*

$$\lambda < \mu_1 \quad \textit{and} \quad p_2 \leq 1/2$$

then the conditions in Remark 2.1.12 hold.

Proof. We consider the Lyapunov function $w(i) = C(i+1)$, for $i \in S$, where the constant C satisfies

$$C \geq \max\{1, \lambda + \mu_2\}.$$

With this definition, w is a Lyapunov function and it indeed satisfies $q(i) \leq w(i)$ for each $i \in S$. Direct calculations give, for every $a \in A$,

$$\sum_{j \in S} q_{0j}(a)w(j) = C\lambda \quad \text{and} \quad \sum_{j \in S} q_{1j}(a)w(j) = -C(a - \lambda)$$

and, for $i \geq 2$,

$$\begin{aligned} \sum_{j \in S} q_{ij}(a)w(j) &= -(a(1+p_2) - \lambda)w(i) + C(a(1+p_2) - \lambda) \\ &\leq -(\mu_1(1+p_2) - \lambda)w(i) + C(\mu_2(1+p_2) - \lambda). \end{aligned} \tag{2.4.2}$$

Choose now c_1 with $0 < c_1 < \mu_1(1+p_2) - \lambda$ and let $I_0 \geq 2$ be such that $i > I_0$ implies

$$\frac{C}{w(i)}(\mu_2(1+p_2) - \lambda) \leq \mu_1(1+p_2) - \lambda - c_1.$$

Then, for $i > I_0$ and $a \in A$ we have

$$\sum_{j \in S} q_{ij}(a)w(j) \leq w(i)\Big(- (\mu_1(1+p_2) - \lambda) + \frac{C}{w(i)}(\mu_2(1+p_2) - \lambda)\Big) \leq -c_1 w(i).$$

Clearly, choosing $b_1 \geq 0$ large enough, we have

$$\sum_{j \in S} q_{ij}(a)w(j) \leq -c_1 w(i) + b_1 \quad \text{for } 0 \leq i \leq I_0 \text{ and } a \in A(i).$$

Therefore, Assumption 2.1.8 holds for the finite set $D = \{0, 1, \ldots, I_0\}$.

It is trivial to check that Assumptions 2.1.9 and 2.1.11(i)–(ii) hold. Also, each deterministic stationary policy $f \in \mathbb{F}$ is irreducible because the process can travel with positive probability between any two states (by augmenting or diminishing the state by one unit). Finally, it remains to study Assumption 2.1.11(iii).

Fix a power $\beta \geq 2$. The idea is to write $q_{ij}(a)w^\beta(j)$, for $i \geq 2$ and $a \in A$, as a power series of $(i+1)$, in which we keep the terms of degree $\beta + 1$ and β. Proceeding this way, we obtain:

$$\begin{aligned} q_{i,i-2}(a)w^\beta(i-2) &= \frac{p_2 a}{C} \cdot w^{\beta+1}(i) - p_2 a(1+2\beta) \cdot w^\beta(i) + R_{-2}(i,a), \\ q_{i,i-1}(a)w^\beta(i-1) &= \frac{p_1 a}{C} \cdot w^{\beta+1}(i) - p_1 a(1+\beta) \cdot w^\beta(i) + R_{-1}(i,a), \\ q_{ii}(a)w^\beta(i) &= -\frac{a+\lambda}{C} \cdot w^{\beta+1}(i) + (a+\lambda) \cdot w^\beta(i), \\ q_{i,i+1}(a)w^\beta(i) &= \frac{\lambda}{C} \cdot w^{\beta+1}(i) + \lambda(\beta-1) \cdot w^\beta(i) + R_1(i,a), \end{aligned}$$

where the residual terms R_ℓ are all $\mathrm{O}((i+1)^{\beta-1})$ and they verify

$$\lim_{i\to\infty}\sup_{a\in A}\frac{|R_\ell(i,a)|}{w^\beta(i)}=0 \quad \text{for } \ell=-2,-1,1.$$

Summing the above equations yields

$$\begin{aligned}\sum_{j\in S} q_{ij}(a)w^\beta(j) &= -\beta(a(1+p_2)-\lambda)\cdot w^\beta(i)+\sum_\ell R_\ell(i,a)\\ &\le w^\beta(i)\left(-\beta(\mu_1(1+p_2)-\lambda)+\frac{\sum R_\ell(i,a)}{w^\beta(i)}\right).\end{aligned}$$

Therefore, proceeding as in the proof of Assumption 2.1.8, if

$$0<c_\beta<\beta\big(\mu_1(1+p_2)-\lambda\big),$$

choosing I_0 large enough such that $i>I_0$ implies

$$\frac{\sum R_\ell(i,a)}{w^\beta(i)}\le\beta\big(\mu_1(1+p_2)-\lambda\big)-c_\beta \quad \text{for all } a\in A,$$

and letting $b_\beta\ge 0$ be large enough, we obtain

$$\sum_{j\in S} q_{ij}(a)w^\beta(j)\le -c_\beta w^\beta(i)+b_\beta\,\mathbf{I}\{0\le i\le I_0\} \quad \text{for all } (i,a)\in\mathbb{K},$$

which is in fact stronger than the requirement in Assumption 2.1.11(iii).

Let us now focus on the conditions given in Remark 2.1.12. For item (a), we must check Assumption 2.1.8 but now for the set $D=\{0\}$. In this case, we assume that $\mu_1>\lambda$, which implies that $\mu_1(1+p_2)>\lambda$.

From (2.4.2) we have that, for $i\ge 2$ and $a\in A$,

$$\sum_{j\in S} q_{ij}(a)w(j)=-Ci\big(a(1+p_2)-\lambda\big)\le -Ci\big(\mu_1(1+p_2)-\lambda\big).$$

Therefore, choosing

$$0<\tilde{c}_1\le\frac{2}{3}\big(\mu_1(1+p_2)-\lambda\big)$$

we obtain, for all $i\ge 2$ and $a\in A$

$$\sum_{j\in S} q_{ij}(a)w(j)\le -\tilde{c}_1 w(i). \tag{2.4.3}$$

For $i=1$ and $a\in A$, we have $\sum_j q_{ij}(a)w(j)\le -C(\mu_1-\lambda)$. Consequently, if we choose

$$\tilde{c}_1=\frac{1}{2}(\mu_1-\lambda)\le\frac{2}{3}\big(\mu_1(1+p_2)-\lambda\big),$$

then (2.4.3) holds also for $i = 1$. Finally, since $\sum_j q_{0j} w(j) = C\lambda$, we conclude that letting

$$\tilde{c}_1 = \frac{1}{2}(\mu_1 - \lambda) \quad \text{and} \quad \tilde{b}_1 = \frac{C}{2}(\mu_1 + \lambda)$$

yields

$$\sum_{j \in S} q_{ij} w(j) \leq -\tilde{c}_1 w(i) + \tilde{b}_1 \mathbf{I}\{i = 0\} \quad \text{for all } (i, a) \in \mathbb{K},$$

and so part (a) in Remark 2.1.12 holds.

For the monotonicity conditions in part (b), after some elementary computations, it can be shown that these monotonicity conditions hold provided that $p_2 \leq 1/2$ (the critical values to obtain this bound are $i = 1$ and $k = 1$). Finally, it is obvious that part (c) in Remark 2.1.12 holds.

Consequently, under these additional conditions, the constants R and γ in the uniform exponential ergodicity conditions become

$$R = 2\big(1 + \tilde{b}_1/\tilde{c}_1\big) = 2\left(1 + \frac{C(\mu_1 + \lambda)}{\mu_1 - \lambda}\right) \quad \text{and} \quad \gamma = \tilde{c}_1 = \frac{1}{2}(\mu_1 - \lambda).$$

The proof is complete. □

Let $g^* \in \mathbb{R}$ be the optimal gain of the controlled birth-and-death process $\mathcal{M}$. Consider the finite state and action truncated control models $\mathcal{M}_n$, for $n \geq 1$, as defined in Section 2.3.1. By construction of the $\mathcal{M}_n$, deterministic stationary policies are irreducible for $\mathcal{M}_n$. So, let $g_n^* \in \mathbb{R}$ be the optimal gain for $\mathcal{M}_n$.

We can use Proposition 2.3.9 to establish the convergence $g_n^* \to g^*$. Proposition 2.3.9 gives, namely, two different sufficient conditions (a) and (b), which correspond to items (i) and (ii) in the result below, respectively. Our next result makes a direct application of the results in Proposition 2.4.5 and its proof is straightforward.

Proposition 2.4.6 *Suppose that the controlled birth-and-death process $\mathcal{M}$ satisfies Assumption 2.4.4.*

(i) If $\mu_1 > \lambda$ and $p_2 \leq 1/2$ then the condition in Proposition 2.3.9(a) holds. So, $g_n^* \to g^*$ and limit policies of average optimal policies of $\mathcal{M}_n$ are average optimal for $\mathcal{M}$.

(ii) If $\mu_1(1 + p_2) > \lambda$ then the condition in Proposition 2.3.9(b) holds. So, $g_n^* \to g^*$, and limit policies of average optimal policies of $\mathcal{M}_n$ are average optimal for $\mathcal{M}$.

Consequently, at least for this example, the Lyapunov condition on w^δ given in Proposition 2.3.9(b) is weaker than the monotonicity conditions in Proposition 2.3.9(a) because, as already noticed, $\mu_1 > \lambda$ implies $\mu_1(1 + p_2) > \lambda$. Therefore, the technique of the Lyapunov conditions on w^δ appears to be a powerful tool from the applications perspective.

The interpretation of the inequality $\mu_1(1 + p_2) > \lambda$ is as follows. The minimal death rate is μ_1, while the expected number of perished individuals is $p_1 + 2p_2 = 1 + p_2$. Hence,

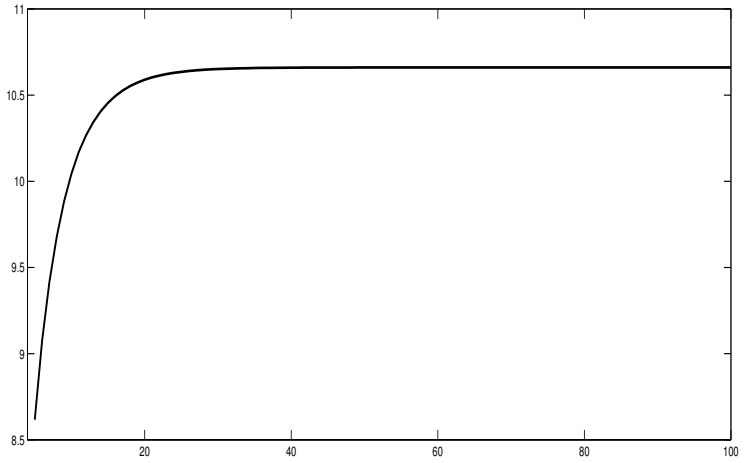


Figure 2.3: The optimal gain g_n^*.

$\mu_1(1+p_2) > \lambda$ states that the minimal death rate (taking into account the diminution of the population) is larger than the birth rate: this is just the usual ergodicity condition of a birth-and-death process.

Finally, we address the issue of the convergence rate.

Proposition 2.4.7 *Suppose that the controlled birth-and-death system satisfies Assumption 2.4.4, with $\mu_1(1+p_2) > \lambda$, and suppose also that the function $a \mapsto c(i,a)$ is Lipschitz continuous on A for each $i \in S$, with a Lipschitz constant L_i such that $L_i \leq \mathfrak{L}(i+1)$ for all $i \in S$. Under these conditions, given $\delta > 2$, if we choose the actions sets $A_n(i)$ of the truncated control model $\mathcal{M}_n$ such that, for some constant $D > 0$,*

$$\rho_A(A_n(i), A) \leq D \cdot \frac{w^\delta(i)}{w^\delta(n)} \quad \textit{for all } n \geq 1 \textit{ and } i \in S_n$$

then there is some constant $\mathfrak{c} > 0$ with

$$|g_n^* - g^*| \leq \frac{1}{n^{\delta-2}} \quad \textit{for all } n \geq 1.$$

Proof. We can use Theorem 2.3.12, which indeed applies as a consequence of our hypotheses and the proof of Proposition 2.4.5. Regarding the condition on $\rho_A(A_n(i), A)$, observe that L_i is of order i, while $\sum_j L_{ij}$ is of order i as well. Therefore, the Hausdorff distance between $A_n(i)$ and $A(i)$ must be bounded by $Dw^\delta(i)/w^\delta(n)$ to obtain the desired convergence rate. □

Therefore, under our standing conditions, we can reach any convergence rate $1/n^\beta$, for $\beta > 0$, provided that we choose sufficiently fine grids of points when discretizing the action space A.

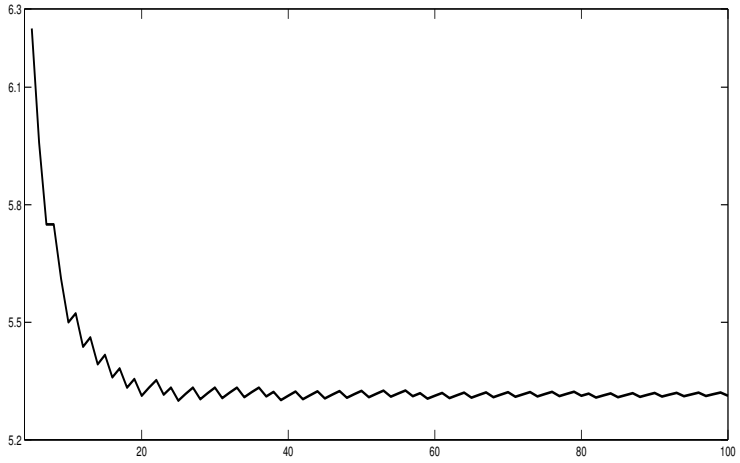


Figure 2.4: The optimal actions $f_n^*(6)$.

Numerical results. For this controlled birth-and-death system, we fix the following values of the parameters:

$$\lambda = 5, \quad \mu_1 = 4.5, \quad \mu_2 = 7, \quad p_1 = 0.75, \quad p_2 = 0.25, \quad p = 10,$$

and so, $\mu_1(1 + p_2) > \lambda$ holds but not $\mu_1 > \lambda$; recall Proposition 2.4.6. The reward rate function is

$$r(i, a) = pi - (a - \mu_2)^2 \ln(1 + \mu_2 - a) \cdot i \quad \text{for all } (i, a) \in \mathbb{K}.$$

The cost rate function $c(i, a) = (a - \mu_2)^2 \ln(1 + \mu_2 - a) \cdot i$ can be interpreted as follows. Suppose that μ_2 is the natural death rate of the population. The controller can decrease this natural death rate by using an adequate medical policy. Hence, the farther the real death rate $a \in [\mu_1, \mu_2]$ is from the natural death rate μ_2, the more expensive is the medical policy.

The finite state and action truncated control model $\mathcal{M}_n$ has state space

$$S_n = \{0, 1, \ldots, n\}$$

and action sets $A_n(i)$, for $i \in S_n$, given by

$$\left\{ \mu_1 + \frac{k}{2n}(\mu_2 - \mu_1) : 0 \leq k \leq 2n \right\}.$$

The transition and reward rates are defined as in Section 2.3.1. In particular, all the transition rates remain unchanged:

$$q_{ij}^n(a) = q_{ij}(a) \quad \text{except for} \quad q_{nn}^n(a) = -an \quad \text{for } a \in A_n(i).$$

For each $1 \leq n \leq 100$ we solve the truncated control model $\mathcal{M}_n$ by using the policy iteration algorithm. In Figure 2.3, we show the optimal gain g_n^* as a function of n. To

study the convergence of optimal policies, Figure 2.4 shows the optimal actions $f_n^*(6)$ for $\mathcal{M}_n$ as a function of n. Empirically, it is clear from Figures 2.3 and 2.4 that the sequences $\{g_n^*\}$ and $\{f_n^*(6)\}$ converge. We deduce the approximate values

$$g^* = 10.6603 \quad \text{and} \quad f^*(6) = 5.3125,$$

for the optimal gain g^* and an optimal policy f^* of $\mathcal{M}$.

Chapter 3

Approximation of Markov games

This chapter is organized as follows. In Section 3.1 we define the game models we will be dealing with, state our main assumptions, and recall some previously known results on the discounted and average payoff optimality criteria. Convergence of game models is defined in Section 3.2, in which we also state our assumptions on the sequence of converging game models. In Section 3.3 we give our main results on approximations of game models under the discounted payoff optimality criterion, while Section 3.4 addresses these issues for the average payoff criterion. Finally, in Section 3.5 we make an application to a controlled population system managed by two players, and we show some computational results using the techniques developed herein.

The results presented in Section 3.1 are already known and they are mainly borrowed from [15, 16, 31]. The rest of the material in this chapter is an original contribution and it corresponds to the publications [34] for the discounted payoff criterion and [26] for the average payoff criterion.

3.1 Basic results

The definition of the game model in this chapter and the corresponding basic results are mainly borrowed from [15, 16] and [31, Chapter 10].

3.1.1 The game model $\mathcal{G}$

We consider a two-player zero-sum continuous-time Markov game model

$$\mathcal{G} = \{S, A, B, \mathbb{K}, Q, r\},$$

where the elements of $\mathcal{G}$ are the following.

- $S = \{0, 1, 2, \ldots\}$ is the state space.
- A and B are the action spaces for players 1 and 2, respectively. We assume that A and B are Borel spaces (i.e., measurable subsets of complete and separable metric spaces). The corresponding metrics are denoted by d_A and d_B, respectively.

- For each $i \in S$, the nonempty measurable sets $A(i) \subseteq A$ and $B(i) \subseteq B$ stand for the actions available for players 1 and 2 in state $i \in S$, respectively. The family of feasible triplets is defined as

$$\mathbb{K} = \{(i, a, b) \in S \times A \times B : a \in A(i), b \in B(i)\}.$$

- The transition rate matrix of the system is $Q = [q_{ij}(a, b)]$. It is assumed that:

  (1) The function $(a, b) \mapsto q_{ij}(a, b)$ is measurable on $A(i) \times B(i)$ for all $i, j \in S$;

  (2) The transition rates are conservative, that is,

$$\sum_{j \in S} q_{ij}(a, b) = 0 \quad \text{for all } (i, a, b) \in \mathbb{K},$$

  with $q_{ij}(a, b) \geq 0$ whenever $i \neq j$;

  (3) The transition rates are stable, i.e., $q(i) := \sup_{a \in A(i), b \in B(i)}\{-q_{ii}(a, b)\} < \infty$.

- Finally, the measurable function $r : \mathbb{K} \to \mathbb{R}$ is interpreted as the reward rate function for player 1 and the cost rate function for player 2.

The game $\mathcal{G}$ is played as follows. At each time $t \geq 0$, both players observe the state of the system $x(t) = i \in S$ and then, independently and simultaneously, they choose actions $a(t) = a \in A(i)$ and $b(t) = b \in B(i)$. In a small time interval $[t, t + dt]$:

- player 1 receives a reward $r(i, a, b)dt$,
- player 2 incurs a cost $r(i, a, b)dt$,
- the system remains in state $i \in S$ with probability $1 + q_{ii}(a, b)dt$ or makes a transition to the state $j \neq i$ with probability $q_{ij}(a, b)dt$.

This procedure is carried on over all the time horizon $t \in [0, \infty)$. The optimality criteria with respect to which the players will try to make optimal decisions will be defined later. Let us mention that we are interested in the discounted and the average payoff optimality criteria.

To ensure the existence of the dynamic game model $\mathcal{G}$ we need some additional assumptions. We will use the following terminology.

Definition 3.1.1 *(a) We say that $w : S \to [1, \infty)$ is a Lyapunov function on S when w is monotone nondecreasing and, in addition, $\lim_{i \to \infty} w(i) = +\infty$.*

(b) Let $\mathcal{B}_w(S)$ denote the family of functions $u : S \to \mathbb{R}$ such that

$$||u||_w = \sup_{i \in S}\{|u(i)|/w(i)\} < \infty.$$

We have that $|| \cdot ||_w$ is a norm on $\mathcal{B}_w(S)$, under which it is a Banach space.

Next we state our conditions on the game model $\mathcal{G}$.

Assumption 3.1.2 *There exists a Lyapunov function w on S, and constants $c_1 \in \mathbb{R}$ and $d_1 \geq 0$ with*

$$\sum_{j \in S} q_{ij}(a,b) w(j) \leq -c_1 w(i) + d_1 \quad \textit{for all } (i,a,b) \in \mathbb{K}.$$

In addition, for each $i \in S$ we have $q(i) \leq w(i)$.

This assumption is usually referred to as a Lyapunov hypothesis on the transition rates of the system. In the sequel, we will explain the use of the above hypotheses.

Strategies of the players. We introduce some notation. Let $\overline{A}(i)$ and $\overline{B}(i)$ be the families of probability measures on $A(i)$ and $B(i)$, when endowed with their Borel σ-algebras $\mathbb{B}(A(i))$ and $\mathbb{B}(B(i))$, respectively. On $\overline{A}(i)$ and $\overline{B}(i)$ we will consider the topology of weak convergence.

Let

$$\pi^1 \equiv \{\pi_t^1(C|i)\}_{t \geq 0, i \in S, C \in \mathbb{B}(A(i))}$$

be such that $\pi_t^1(\cdot|i)$ is in $\overline{A}(i)$ for all $t \geq 0$ and $i \in S$, and such that $t \mapsto \pi_t^1(C|i)$ is a measurable function on $[0,\infty)$ for all $C \in \mathbb{B}(A(i))$ and $i \in S$. We say that π^1 is a *randomized Markov strategy* for player 1, and we denote by Π^1 the set of all such strategies. The family Π^2 of randomized Markov strategies

$$\pi^2 \equiv \{\pi_t^2(C|i)\}_{t \geq 0, i \in S, C \in \mathbb{B}(B(i))}$$

for player 2 is defined similarly.

We say that $\pi^1 \in \Pi^1$ is a *randomized stationary strategy* (or stationary, for short) for player 1 when $\pi_t^1(C|i)$ does not depend on $t \geq 0$. Thus, the class $\Pi^{1,s}$ of stationary strategies for player 1 can be identified with

$$\Pi^{1,s} = \prod_{i \in S} \overline{A}(i).$$

Similarly, the class of randomized stationary strategies for player 2 is $\Pi^{2,s} = \prod_{i \in S} \overline{B}(i)$.

Given a pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, $i, j \in S$, and $t \geq 0$, define

$$q_{ij}(t, \pi^1, \pi^2) = \int_{A(i)} \int_{B(i)} q_{ij}(a,b) \pi_t^2(db|i) \pi_t^1(da|i).$$

The above integral is well defined and finite because the system's transition rates are conservative and stable. In particular, they satisfy

$$-q_{ii}(t, \pi^1, \pi^2) = \sum_{j \neq i} q_{ij}(t, \pi^1, \pi^2) \leq q(i) \quad \text{for each } t \geq 0 \text{ and } i \in S.$$

Define also

$$r(t,i,\pi^1,\pi^2)=\int_{A(i)}\int_{B(i)} r(i,a,b)\pi_t^2(db|i)\pi_t^1(da|i).$$

We will also use the following notation. Given $i,j\in S$, $\varphi\in\overline{A}(i)$, and $\psi\in\overline{B}(i)$, let

$$q_{ij}(\varphi,\psi) = \int_{A(i)}\int_{B(i)} q_{ij}(a,b)\psi(db)\varphi(da), \tag{3.1.1}$$

$$r(i,\varphi,\psi) = \int_{A(i)}\int_{B(i)} r(i,a,b)\psi(db)\varphi(da), \tag{3.1.2}$$

and for stationary strategies $(\pi^1,\pi^2)\in\Pi^{1,s}\times\Pi^{2,s}$, we write

$$q_{ij}(\pi^1,\pi^2)=q_{ij}(\pi^1(\cdot|i),\pi^2(\cdot|i)) \quad\text{and}\quad r(i,\pi^1,\pi^2)=r(i,\pi^1(\cdot|i),\pi^2(\cdot|i)).$$

Our next result summarizes the main results on the existence of the state and actions process. See, e.g., Proposition 3.1 in [15] or Proposition 10.3 in [31].

Theorem 3.1.3 *Suppose that Assumption 3.1.2 is satisfied.*

(i) For every $(\pi^1,\pi^2)\in\Pi^1\times\Pi^2$ there exists a regular (nonhomogeneous) transition function

$$\{P_{ij}^{\pi^1,\pi^2}(s,t)\}_{i,j\in S,0\le s\le t}$$

with transition rates $q_{ij}(t,\pi^1,\pi^2)$, that is,

$$\lim_{h\downarrow 0}\frac{P_{ij}^{\pi^1,\pi^2}(t,t+h)-\delta_{ij}}{h}=q_{ij}(t,\pi^1,\pi^2) \quad \textit{for all } i,j\in S \textit{ and } t\ge 0.$$

Let $\Omega=\mathbb{K}^{[0,\infty)}=\{(x(t),a(t),b(t))\}_{t\ge 0}$ be endowed with the product σ-algebra $\mathcal{F}$.

(ii) Given an initial state $i\in S$ at time 0 and $(\pi^1,\pi^2)\in\Pi^1\times\Pi^2$, there exists a unique probability measure P^{i,π^1,π^2} on $(\Omega,\mathcal{F})$ such that:

- *For each $A_0\in\mathbb{B}(A(i))$ and $B_0\in\mathbb{B}(B(i))$, we have*

$$P^{i,\pi^1,\pi^2}\{x(0)=i,a(0)\in A_0,b(0)\in B_0\}=\pi_0^1(A_0|i)\cdot\pi_0^2(B_0|i).$$

- *Given arbitrary $n\ge 1$ and $0\le s_1\le s_2\le\ldots\le s_n$, and, on the other hand, given $i_k\in S$, $A_k\in\mathbb{B}(A(i_k))$, and $B_k\in\mathbb{B}(B(i_k))$, for $k=1,\ldots,n$, we have*

$$\begin{aligned}P^{i,\pi^1,\pi^2}&\{x(s_1)=i_1,a(s_1)\in A_1,b(s_1)\in B_1,\ldots,\\ &\qquad x(s_n)=i_n,a(s_n)\in A_n,b(s_n)\in B_n\}\\ &=\prod_{k=1}^n P_{i_{k-1}i_k}^{\pi^1,\pi^2}(s_{k-1},s_k)\pi_{s_k}^1(A_k|i_k)\pi_{s_k}^2(B_k|i_k),\end{aligned}$$

with the convention that $i_0=i$ and $s_0=0$.

This result ensures the existence of the dynamic game model itself. In particular, the Lyapunov condition stated in Assumption 3.1.2 is used to ensure the uniqueness and non-explosiveness of the non-homogeneous Q-process, and it guarantees that the forward and backward Kolmogorov differential equations are satisfied.

We will refer to $\{x(t)\}_{t\geq 0}$ as the state process, while $\{a(t)\}_{t\geq 0}$ and $\{b(t)\}_{t\geq 0}$ are the actions processes for players 1 and 2. The expectation operator associated to the probability measure P^{i,π^1,π^2} will be denoted by E^{i,π^1,π^2}.

3.1.2 The discounted payoff optimality criterion

In this section we analyze the discounted optimality criterion. We will suppose that the reward/cost of the players is depreciated at a constant discount rate $\alpha > 0$, and so the infinitesimal rate $r(x(t), a(t), b(t))$ at time $t \geq 0$ is brought to its present value, namely, $e^{-\alpha t} r(x(t), a(t), b(t))$. The goal of player 1 is then to maximize his total expected discounted reward, loosely,

$$E\Big[\int_0^\infty e^{-\alpha t} r(x(t), a(t), b(t))dt\Big],$$

while player 2 wants to minimize his total expected discounted cost. A formal definition will be given below.

Assumption 3.1.4 *The game model $\mathcal{G}$ satisfies the following conditions.*

(i) The discount rate α satisfies $\alpha + c_1 > 0$, with c_1 the constant in Assumption 3.1.2.

(ii) There exists a constant $M > 0$ such that $|r(i, a, b)| \leq M w(i)$ for all $(i, a, b) \in \mathbb{K}$.

Given an initial state $i \in S$ and a pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, we define the total expected discounted payoff as

$$V^\alpha(i, \pi^1, \pi^2) = E^{i,\pi_1,\pi_2}\left[\int_0^\infty e^{-\alpha t} r(x(t), a(t), b(t))dt\right]. \tag{3.1.3}$$

Thus, $V^\alpha(i, \pi^1, \pi^2)$ is the total expected discounted reward for player 1, and it is the total expected discounted cost for player 2. Using [14, Lemma 3.2] or [17, Lemma 6.3], under Assumption 3.1.2, we have that

$$E^{i,\pi^1,\pi^2}[w(x(t))] \leq e^{-c_1 t} w(i) + \frac{d_1}{c_1}(1 - e^{-c_1 t}). \tag{3.1.4}$$

We note that if $c_1 = 0$ then the righthand term of (3.1.4) is $w(i) + d_1 t$; to see this, just let $c_1 \uparrow 0$ in (3.1.4). As a consequence of Assumption 3.1.4, given $i \in S$ and $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, we have

$$|V^\alpha(i, \pi^1, \pi^2)| \leq M \int_0^\infty e^{-\alpha t} E^{i,\pi^1,\pi^2}[w(x(t))]dt \leq \frac{M w(i)}{\alpha + c_1} + \frac{d_1 M}{\alpha(\alpha + c_1)}.$$

Hence, letting $\mathfrak{M} = \frac{M(\alpha + d_1)}{\alpha(\alpha + c_1)}$, it follows that

$$\|V^\alpha(\cdot, \pi^1, \pi^2)\|_w \leq \mathfrak{M} \quad \text{for all } (\pi^1, \pi^2) \in \Pi^1 \times \Pi^2. \tag{3.1.5}$$

Remark 3.1.5 *If (π^1, π^2) is a pair of stationary strategies, by Theorem 3.1.3(ii) we have (cf. (3.1.3))*

$$V^\alpha(i, \pi^1, \pi^2) = E^{i,\pi^1,\pi^2}\left[\int_0^\infty e^{-\alpha t} r(x(t), \pi^1, \pi^2) dt\right] \quad \text{for each } i \in S.$$

Therefore, in order to obtain $V^\alpha(i, \pi^1, \pi^2)$ for a pair of stationary strategies, instead of integrating the reward state-actions process $r(x(t), a(t), b(t))$, it suffices to integrate the function $r(x(t), \pi^1, \pi^2)$ which depends only on the state process.

Given the initial state $i \in S$, the discounted lower value and upper value functions of the game model $\mathcal{G}$ are defined as

$$\begin{aligned} L^\alpha(i) &= \sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} V^\alpha(i, \pi^1, \pi^2) \\ U^\alpha(i) &= \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} V^\alpha(i, \pi^1, \pi^2), \end{aligned}$$

respectively. We note that, as a consequence of (3.1.5), we have

$$||L^\alpha||_w \leq \mathfrak{M} \quad \text{and} \quad ||U^\alpha||_w \leq \mathfrak{M}.$$

The lower value of the game is the maximal discounted reward for player 1 when using a "maximin" strategy. Indeed, for every fixed strategy $\pi^1 \in \Pi^1$, the worst scenario for player 1 is when player 2 chooses the strategy $\pi^2 \in \Pi^2$ attaining the infimum

$$\inf_{\pi^2 \in \Pi^2} V^\alpha(i, \pi^1, \pi^2).$$

Then, player 1 chooses the strategy yielding the maximal reward, that is, the one achieving the supremum in the definition of $L^\alpha(i)$. Similarly, the upper value of the game corresponds to the optimal payoff of player 2 when using a "minimax" strategy. It is easy to see that $L^\alpha(i) \leq U^\alpha(i)$ for every $i \in S$.

Definition 3.1.6 *The game $\mathcal{G}$ has a value when $L^\alpha(i) = U^\alpha(i)$ for all $i \in S$. The function $V^\alpha(i) := L^\alpha(i) = U^\alpha(i)$ is called the value function of the game $\mathcal{G}$.*

In this case, we say that $(\pi^{*1}, \pi^{*2}) \in \Pi^1 \times \Pi^2$ is a pair of discount optimal strategies when

$$V^\alpha(i, \pi^1, \pi^{*2}) \leq V^\alpha(i, \pi^{*1}, \pi^{*2}) \leq V^\alpha(i, \pi^{*1}, \pi^2)$$

for all $i \in S$ and $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$.

A direct calculation shows that if V^α is the value function of the game and $(\pi^{*1}, \pi^{*2}) \in \Pi^1 \times \Pi^2$ is a pair of discount optimal strategies, then $V^\alpha(i) = V^\alpha(i, \pi^{*1}, \pi^{*2})$ for each $i \in S$. A pair of optimal strategies is usually referred to as a noncooperative or Nash equilibrium of the game.

To ensure the existence of the value function and solutions to the corresponding Shapley equations, we need to impose further conditions on our game model.

Assumption 3.1.7 *The game model $\mathcal{G}$ satisfies the following conditions.*

(i) For each $i \in S$, the sets $A(i)$ and $B(i)$ are compact.

(ii) For all $i, j \in S$, the functions $r(i,a,b)$ and $q_{ij}(a,b)$ are continuous on $A(i) \times B(i)$.

(iii) There exist constants $c_2 \in \mathbb{R}$ and $d_2 \geq 0$ with

$$\sum_{j \in S} q_{ij}(a,b) w^2(j) \leq -c_2 w^2(i) + d_2 \quad \text{for all } (i,a,b) \in \mathbb{K}.$$

The conditions (i)–(ii) above are the usual continuity-compactness hypotheses, while (iii) is used to ensure that Dynkin's formula holds. It imposes a Lyapunov condition on the function w^2 but, this time, the constant c_2 needs not be related to the discount rate α; cf. Assumption 3.1.4(i).

The following lemma is a consequence of our assumptions, and it will be useful in the forthcoming.

Lemma 3.1.8 *Suppose that Assumption 3.1.7 holds.*

(i) Given $i \in S$ and $k > i$, we have

$$\sum_{j \geq k} q_{ij}(\varphi, \psi) w(j) \leq \frac{1}{w(k)} \Big(- c_2 w^2(i) + d_2 + q(i) w^2(i) \Big)$$

for all $\varphi \in \overline{A}(i)$ and $\psi \in \overline{B}(i)$.

(ii) For every $i \in S$ and $u \in \mathcal{B}_w(S)$, the functions

$$(\varphi, \psi) \mapsto r(i, \varphi, \psi) \quad \text{and} \quad (\varphi, \psi) \mapsto \sum_{j \in S} q_{ij}(\varphi, \psi) u(j)$$

are continuous on $\overline{A}(i) \times \overline{B}(i)$, when endowed with the product topology of the weak convergences on $\overline{A}(i)$ and $\overline{B}(i)$.

Proof. (i). First of all, observe that given $i \in S$, $k > i$, and $(a,b) \in A(i) \times B(i)$ we have

$$\begin{aligned} \sum_{j \geq k} q_{ij}(a,b) w(j) &\leq \frac{1}{w(k)} \sum_{j \geq k} q_{ij}(a,b) w^2(j) && (3.1.6) \\ &\leq \frac{1}{w(k)} \Big(\sum_{j \in S} q_{ij}(a,b) w^2(j) - q_{ii}(a,b) w^2(i) \Big) \\ &\leq \frac{1}{w(k)} \Big(- c_2 w^2(i) + d_2 + q(i) w^2(i) \Big), && (3.1.7) \end{aligned}$$

where we have used monotonicity of w in (3.1.6) and Assumption 3.1.7(iii) in (3.1.7).

(ii). From (3.1.7) we have that

$$\lim_{k\to\infty} \sup_{(a,b)\in A(i)\times B(i)} \sum_{j\geq k} q_{ij}(a,b)w(j) = 0$$

and, in particular,

$$\lim_{k\to\infty} \sup_{(a,b)\in A(i)\times B(i)} \Big| \sum_{j\geq k} q_{ij}(a,b)u(j) \Big| = 0.$$

Consequently, the series $\sum q_{ij}(a,b)u(j)$ of continuous functions converges uniformly on $A(i)\times B(i)$. It is therefore a bounded and continuous function, because $A(i)$ and $B(i)$ are compact. This establishes that $(a,b)\mapsto \sum_{j\in S} q_{ij}(a,b)u(j)$ is continuous. The continuity of $(\varphi,\psi)\mapsto \sum_{j\in S} q_{ij}(\varphi,\psi)u(j)$ follows from Theorem 3.2 in [6]. The arguments for the continuity of $(\varphi,\psi)\mapsto r(i,\varphi,\psi)$ are similar. □

The continuity of $\sum_{j\in S} q_{ij}(a,b)w(j)$ is a usual requirement in Markov game models; see [16, Assumption C.3] or [31, Assumption 10.7.b]. As seen in Lemma 3.1.8 above, this condition is in fact implied by our hypotheses.

The main result on the discounted game $\mathcal{G}$ is the following. It is borrowed from [16, 31].

Theorem 3.1.9 *Suppose that the game model $\mathcal{G}$ satisfies Assumptions 3.1.2, 3.1.4, and 3.1.7.*

(i) The game $\mathcal{G}$ has a value $V^\alpha \in \mathcal{B}_w(S)$ with $\|V^\alpha\|_w \leq \mathfrak{M}$.

(ii) The value function V^α is the unique solution u in $\mathcal{B}_w(S)$ of the equations

$$\alpha u(i) = \sup_{\varphi\in\overline{A}(i)} \inf_{\psi\in\overline{B}(i)} \Big\{ r(i,\varphi,\psi) + \sum_{j\in S} q_{ij}(\varphi,\psi)u(j) \Big\} \tag{3.1.8}$$

$$= \inf_{\psi\in\overline{B}(i)} \sup_{\varphi\in\overline{A}(i)} \Big\{ r(i,\varphi,\psi) + \sum_{j\in S} q_{ij}(\varphi,\psi)u(j) \Big\} \tag{3.1.9}$$

for all $i\in S$.

(iii) There exists a pair of optimal randomized stationary strategies.

Moreover, $(\pi^1,\pi^2)\in \Pi^{1,s}\times\Pi^{2,s}$ is a pair of optimal randomized stationary strategies if and only if $\pi^1(\cdot|i)$ and $\pi^2(\cdot|i)$ attain the supremum and the infimum in (3.1.8) and (3.1.9), respectively, for every $i\in S$. That is,

$$\alpha u(i) = \inf_{\psi\in\overline{B}(i)} \Big\{ r(i,\pi^1(\cdot|i),\psi) + \sum_{j\in S} q_{ij}(\pi^1(\cdot|i),\psi)u(j) \Big\}$$

$$= \sup_{\varphi\in\overline{A}(i)} \Big\{ r(i,\varphi,\pi^2(\cdot|i)) + \sum_{j\in S} q_{ij}(\varphi,\pi^2(\cdot|i))u(j) \Big\}$$

for all $i\in S$.

Remark 3.1.10 *The fact that there exists a pair of optimal stationary strategies implies that the value of the game satisfies*

$$V^\alpha(i) = \sup_{\pi^1\in\Pi^{1,s}} \inf_{\pi^2\in\Pi^{2,s}} V^\alpha(i,\pi^1,\pi^2) = \inf_{\pi^2\in\Pi^{2,s}} \sup_{\pi^1\in\Pi^{1,s}} V^\alpha(i,\pi^1,\pi^2)$$

for all $i \in S$. Note that we are taking the infimum and the supremum over the family of stationary strategies (cf. the definition of the lower and upper value of the game).

The equations (3.1.8)–(3.1.8) are also referred to as the Shapley equations.

3.1.3 The average payoff optimality criterion

In this section we will study the average optimality criterion, that is, when the players want to optimize their long-run expected average payoff. To study the average optimality criterion, further conditions on the game model $\mathcal{G}$ must be imposed. In particular, the Assumption 3.1.2, which ensures the existence of the state and actions process, is replaced with the following stronger condition.

Assumption 3.1.11 *There exists a Lyapunov function w on S, constants $c_1 > 0$ and $d_1 \geq 0$, and a finite set $D \subset S$ such that*

$$\sum_{j\in S} q_{ij}(a,b)w(j) \leq -c_1 w(i) + d_1 \mathbf{I}_D(i) \quad \text{for all } (i,a,b) \in \mathbb{K}.$$

Moreover, for all $i \in S$ we have $q(i) \leq w(i)$.

We note that we use the same notation as in Assumption 3.1.2 for the constants c_1, d_1 in the Lyapunov condition. This will not cause confusion because it will always be clear from the context whether we are in the discounted or the average payoff case.

Since Assumption 3.1.11 implies Assumption 3.1.2, Theorem 3.1.3 applies and there indeed exist state and actions processes $\{(x(t),a(t),b(t))\}_{t\geq 0}$ for the game model $\mathcal{G}$. The expectation operator associated to P^{i,π^1,π^2} will be, as before, denoted by E^{i,π^1,π^2}.

Under Assumption 3.1.11, given an initial state $i \in S$ and a pair of strategies $(\pi^1,\pi^2) \in \Pi^1 \times \Pi^2$, by [17, Lemma 6.3], cf. (3.1.4), we have

$$E^{i,\pi^1,\pi^2}[w(x(t))] \leq e^{-c_1 t} w(i) + \frac{d_1}{c_1}(1 - e^{-c_1 t}) \quad \text{for all } t \geq 0. \tag{3.1.10}$$

The long-run expected average payoff (or average payoff) of the players, when starting from the initial state $i \in S$, and using the strategies $\pi^1 \in \Pi^1$ and $\pi^2 \in \Pi^2$, is defined as

$$J(i,\pi^1,\pi^2) = \limsup_{T\to\infty} \frac{1}{T} E^{i,\pi^1,\pi^2}\left[\int_0^T r(x(t),a(t),b(t))dt\right].$$

To ensure finiteness of the average payoff of the players, we use Assumption 3.1.4(ii), which is stated here for ease of reference.

Assumption 3.1.12 *There exists a constant $M > 0$ such that $|r(i,a,b)| \leq Mw(i)$ for all $(i,a,b) \in \mathbb{K}$.*

Under Assumption 3.1.12 and as a consequence of (3.1.10), the long-run average payoff is finite. Moreover,

$$|J(i,\pi^1,\pi^2)| \leq \frac{Md_1}{c_1} \quad \text{for all } i \in S \text{ and } (\pi^1,\pi^2) \in \Pi^1 \times \Pi^2.$$

Definition 3.1.13 *The long-run average lower and upper value functions of the game $\mathcal{G}$ are respectively defined as*

$$L(i) = \sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} J(i,\pi^1,\pi^2) \quad \textit{and} \quad U(i) = \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} J(i,\pi^1,\pi^2)$$

for each $i \in S$. If $L(i) = U(i) =: V(i)$ for all $i \in S$ then we say that the game $\mathcal{G}$ has a value, and V is the value function of the game. If the game $\mathcal{G}$ has a value then we say that $(\pi^1_*, \pi^2_*) \in \Pi^1 \times \Pi^2$ is a pair of average optimal strategies if

$$V(i) = \inf_{\pi^2 \in \Pi^2} J(i,\pi^1_*,\pi^2) = \sup_{\pi^1 \in \Pi^1} J(i,\pi^1,\pi^2_*) \quad \textit{for each } i \in S.$$

As a consequence of Assumption 3.1.12, the lower and upper value functions of the game are finite, namely,

$$|L(i)| \leq \frac{Md_1}{c_1} \quad \text{and} \quad |U(i)| \leq \frac{Md_1}{c_1}, \quad \text{for all } i \in S. \tag{3.1.11}$$

The next assumption uses the following terminology. We say that a pair of stationary strategies $(\pi^1,\pi^2) \in \Pi^{1,s} \times \Pi^{2,s}$ is irreducible if, given arbitrary distinct states $i, j \in S$, there exist states $i = i_0, i_1, \ldots, i_n = j$ such that $q_{i_{k-1},i_k}(\pi^1,\pi^2) > 0$ for all $k = 1, \ldots, n$. Equivalently, the homogeneous Markov chain $\{x(t)\}_{t\geq 0}$ is irreducible under P^{i,π^1,π^2} for every initial state $i \in S$.

Assumption 3.1.14 *The game model $\mathcal{G}$ satisfies the following conditions.*

(i) For each $i \in S$, the sets $A(i)$ and $B(i)$ are compact.

(ii) For all $i, j \in S$, the functions $r(i,a,b)$ and $q_{ij}(a,b)$ are continuous on $A(i) \times B(i)$.

(iii) There exist constants $c_2 \in \mathbb{R}$ and $d_2 \geq 0$ with

$$\sum_{j \in S} q_{ij}(a,b)w^2(j) \leq -c_2 w^2(i) + d_2 \quad \textit{for all } (i,a,b) \in \mathbb{K}.$$

(iv) Each pair of strategies in $\Pi^{1,s} \times \Pi^{2,s}$ is irreducible.

We note that Assumptions 3.1.14(i)–(iii) are the same as Assumption 3.1.7. For ease of reference, however, we prefer to state them again here.

Under Assumptions 3.1.11 and 3.1.14(iv) we have that, for each $(\pi^1, \pi^2) \in \Pi^{1,s} \times \Pi^{2,s}$, the Markov chain $\{x(t)\}_{t\geq 0}$ (under the probability measure P^{i,π^1,π^2}) has a unique invariant probability measure μ_{π^1,π^2} on S that does not depend on the initial state and which, in addition, satisfies $\mu_{\pi^1,\pi^2}(w) < \infty$; see, e.g., [31, Theorem 2.5]. In particular, the average payoff of $(\pi^1, \pi^2) \in \Pi^{1,s} \times \Pi^{2,s}$ does not depend on the initial state $i \in S$ and

$$J(i, \pi^1, \pi^2) = \mu_{\pi^1,\pi^2}(r(\cdot, \pi^1, \pi^2)) \quad \text{for all } i \in S.$$

Furthermore, if Assumption 3.1.14 is satisfied then the game model $\mathcal{G}$ is uniformly exponentially ergodic on $\Pi^{1,s} \times \Pi^{2,s}$; see, e.g., [31, Assumption 10.10]. This means that there exist positive constants R and γ such that for each $u \in \mathcal{B}_w(S)$, $t \geq 0$, and $i \in S$,

$$\sup_{(\pi^1,\pi^2)\in\Pi^{1,s}\times\Pi^{2,s}} \left|E^{i,\pi^1,\pi^2}[u(x(t))] - \mu_{\pi^1,\pi^2}(u)\right| \leq Re^{-\gamma t}||u||_w w(i). \tag{3.1.12}$$

Remark 3.1.15 *In general, it is not possible to have an explicit expression for the value of the constants R and γ in (3.1.12) above. There exists, however, a particular case in which these constants are actually known. For a reference, see [14, 27] or [30, Theorem 2.8]. Suppose that Assumptions 3.1.11 and 3.1.14 hold and, in addition, assume that*

(a) In Assumption 3.1.11 we have $D = \{0\}$.

(b) For every $(\pi^1, \pi^2) \in \Pi^{1,s} \times \Pi^{2,s}$, the stochastic process $\{x(t)\}_{t\geq 0}$ is stochastically ordered in its initial value:

$$\sum_{j=k}^{\infty} q_{ij}(\pi^1, \pi^2) \leq \sum_{j=k}^{\infty} q_{i+1,j}(\pi^1, \pi^2)$$

for all $i, k \in S$ with $k \neq i + 1$.

(c) For each $(\pi^1, \pi^2) \in \Pi^{1,s} \times \Pi^{2,s}$ and every $0 < i < j$, the process $\{x(t)\}_{t\geq 0}$ can travel with positive probability from i to $\{j, j + 1, \ldots\}$ without passing through 0. Equivalently, there exist nonzero states $i = k_0, k_1, \ldots, k_n$, with $k_n \geq j$, such that $q_{k_{s-1}k_s}(\pi^1, \pi^2) > 0$ for all $s = 1, \ldots, n$.

Under these conditions, the game model $\mathcal{G}$ is uniformly exponentially ergodic on $\Pi^{1,s} \times \Pi^{2,s}$ and the value of the constants in (3.1.12) is

$$R = 2(1 + d_1/c_1) \quad \text{and} \quad \gamma = c_1.$$

We introduce some more terminology. We say that a pair $(g, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ is a solution of the average optimality equations of the game model $\mathcal{G}$ if, for all $i \in S$,

$$\begin{aligned} g &= \sup_{\varphi\in\overline{A}(i)} \inf_{\psi\in\overline{B}(i)} \Big\{r(i, \varphi, \psi) + \sum_{j\in S} q_{ij}(\varphi, \psi)h(j)\Big\} \\ &= \inf_{\psi\in\overline{B}(i)} \sup_{\varphi\in\overline{A}(i)} \Big\{r(i, \varphi, \psi) + \sum_{j\in S} q_{ij}(\varphi, \psi)h(j)\Big\}. \end{aligned}$$

The main result on the existence of the value of the game and on characterization of optimal strategies follows.

Theorem 3.1.16 *Suppose that the game model $\mathcal{G}$ satisfies Assumptions 3.1.11, 3.1.12, and 3.1.14. Then the following statements hold.*

(i) *The game has a value $g^* \in \mathbb{R}$ that does not depend on the initial state, that is, $V(i) = g^*$ for all $i \in S$.*

(ii) *There exist solutions to the average optimality equations.*

If $(g, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ is a solution to the average optimality equations then $g = g^*$, the value of the game, and h is unique up to additive constants.

There exists a solution (g^*, h) of the average optimality equations with $||h||_w \leq RM/\gamma$ (recall (3.1.12)).

(iii) *There exists a pair of optimal stationary strategies. A pair of stationary strategies $(\pi^1, \pi^2) \in \Pi^{1,s} \times \Pi^{2,s}$ is average optimal if and only if*

$$\begin{aligned} g^* &= \inf_{\psi \in \overline{B}(i)} \left\{ r(i, \pi^1(\cdot|i), \psi) + \sum_{j \in S} q_{ij}(\pi^1(\cdot|i), \psi) h(j) \right\} \\ &= \sup_{\varphi \in \overline{A}(i)} \left\{ r(i, \varphi, \pi^2(\cdot|i)) + \sum_{j \in S} q_{ij}(\varphi, \pi^2(\cdot|i)) h(j) \right\} \end{aligned}$$

for each $i \in S$, where (g^*, h) is any solution of the average optimality equations.

3.2 Convergence of game models

3.2.1 Definition

In the forthcoming we consider a sequence of game models

$$\mathcal{G}_n = \{S_n, A, B, \mathbb{K}_n, Q_n, r_n\} \quad \text{for } n \geq 1.$$

These game models will be interpreted as approximations of the original game model $\mathcal{G}$. The elements of these game models satisfy the following conditions.

- The state space S_n is a (finite or infinite) subset of S.
- The action spaces are A and B, as for the game model $\mathcal{G}$.
- The set of available actions in state $i \in S_n$ are the nonempty measurable sets $A_n(i) \subseteq A(i)$ and $B_n(i) \subseteq B(i)$ for players 1 and 2, respectively. Let

$$\mathbb{K}_n = \{(i, a, b) \in S_n \times A \times B : a \in A_n(i), b \in B_n(i)\} \subseteq \mathbb{K}.$$

- The transition rate matrix is given by $Q_n = [q_{ij}^n(a,b)]$ for $i,j \in S_n$ and $(a,b) \in A_n(i) \times B_n(i)$. We assume that $(a,b) \mapsto q_{ij}^n(a,b)$ is measurable on $A_n(i) \times B_n(i)$ for all $i,j \in S_n$. The transition rates are assumed to be conservative and stable, that is,

$$\sum_{j \in S_n} q_{ij}^n(a,b) = 0 \quad \text{and} \quad \sup_{a \in A_n(i), b \in B_n(i)} \{-q_{ii}^n(a,b)\} =: q_n(i) < \infty$$

  for $(i,a,b) \in \mathbb{K}_n$, with the condition that $q_{ij}^n(a,b) \geq 0$ for $i \neq j$.

- The reward/cost rate function is $r_n : \mathbb{K}_n \to \mathbb{R}$, assumed to be measurable.

As for $\mathcal{G}$, the game model $\mathcal{G}_n$ is a two-person zero-sum continuous-time Markov game. We will be interested in the discounted and the average payoff optimality criteria for the game models $\mathcal{G}_n$.

Next we introduce some notation. In the game model $\mathcal{G}_n$, the family of randomized Markov strategies for players 1 and 2 are denoted by Π_n^1 and Π_n^2, respectively. They are defined similarly to the corresponding strategies for the game model $\mathcal{G}$. The family of stationary strategies is

$$\Pi_n^{1,s} = \prod_{i \in S_n} \overline{A}_n(i) \quad \text{and} \quad \Pi_n^{2,s} = \prod_{i \in S_n} \overline{B}_n(i),$$

where $\overline{A}_n(i)$ and $\overline{B}_n(i)$ denote the family of probability measures on $A_n(i)$ and $B_n(i)$, respectively. We will consider the topology of weak convergence on $\overline{A}_n(i)$ and $\overline{B}_n(i)$.

With w a Lyapunov function in S, let $\mathcal{B}_w(S_n)$ be the Banach space of functions $u : S_n \to \mathbb{R}$ with finite w-norm

$$||u||_w = \sup_{i \in S_n} \{|u(i)|/w(i)\}.$$

(We note that we use the same notation $||u||_w$ for $u : S \to \mathbb{R}$ and $u : S_n \to \mathbb{R}$.) Notations such as

$$q_{ij}^n(\varphi, \psi) \quad \text{and} \quad r_n(i, \varphi, \psi)$$

for $i,j \in S_n$, $\varphi \in \overline{A}_n(i)$, and $\psi \in \overline{B}_n(i)$ are given the obvious definitions; see (3.1.1)–(3.1.2).

Consider the game models $\mathcal{G}$ and $\mathcal{G}_n$ studied so far. We propose a definition of the game models $\mathcal{G}_n$ converging to the original game model $\mathcal{G}$. In this definition, we make use of the Hausdorff metric (recall its definition given in Section 1.4).

Definition 3.2.1 *We say that $\mathcal{G}_n \to \mathcal{G}$ as $n \to \infty$ when the following conditions hold:*

(a) The sequence of states $\{S_n\}_{n \geq 1}$ verifies

$$S_1 \subseteq S_2 \subseteq S_3 \subseteq \ldots \quad \text{and} \quad \bigcup_{n=1}^{\infty} S_n = S.$$

This will be denoted by $S_n \uparrow S$. We define $n(i) = \min\{n \geq 1 : i \in S_n\}$ for each $i \in S$, and so $i \in S_n$ if and only if $n \geq n(i)$.

(b) For each $i \in S$, the sequences of action sets $A_n(i)$ and $B_n(i)$ verify

$$\lim_{n\to\infty} \rho_A(A_n(i), A(i)) = 0 \quad \textit{and} \quad \lim_{n\to\infty} \rho_B(B_n(i), B(i)) = 0.$$

We will write $\rho_n(i) = \rho_A(A_n(i), A(i)) \vee \rho_B(B_n(i), B(i))$.

(In the sequel, we will assume that the sets $A_n(i), B_n(i), A(i), B(i)$ are closed and so we will properly say that $A_n(i)$ and $B_n(i)$ converge to $A(i)$ and $B(i)$ as $n \to \infty$, respectively, in the Hausdorff metric.)

For every $i \in S$, given sequences $\{a_n\}_{n\geq n(i)}$ and $\{b_n\}_{n\geq n(i)}$, with $a_n \in A_n(i)$ and $b_n \in B_n(i)$, such that $a_n \to a$ and $b_n \to b$ for some $a \in A(i)$ and $b \in B(i)$, we have:

(c) $\lim_{n\to\infty} q^n_{ij}(a_n, b_n) = q_{ij}(a, b)$ for all $j \in S$, and

(d) $\lim_{n\to\infty} r_n(i, a_n, b_n) = r(i, a, b)$.

Observe that expressions such as $\rho_A(A_n(i), A(i))$ or $q^n_{ij}(a_n, b_n)$ are defined only for large enough n (namely, $n \geq n(i)$ in the former case, and $n \geq n(i) \vee n(j)$ in the latter). This is not made explicit in the notation since we are dealing with the limit as $n \to \infty$.

Our next lemma gives some equivalent statements of Definition 3.2.1.

Lemma 3.2.2 *Suppose that Assumptions 3.1.7(i)–(ii) hold.*

(i) The condition in Definition 3.2.1(c) can be replaced with the following statement. Given $i, j \in S$ and $\epsilon > 0$ there exists $n_0 \geq n(i) \vee n(j)$ such that for all $n \geq n_0$

$$\sup_{(a,b)\in A_n(i)\times B_n(i)} |q^n_{ij}(a, b) - q_{ij}(a, b)| \leq \epsilon.$$

(ii) The condition in Definition 3.2.1(d) can be replaced with the following statement. Given $i \in S$ and $\epsilon > 0$ there exists $n_0 \geq n(i)$ such that for all $n \geq n_0$

$$\sup_{(a,b)\in A_n(i)\times B_n(i)} |r_n(i, a, b) - r(i, a, b)| \leq \varepsilon.$$

Proof. (i). First we prove that if Definition 3.2.1 holds, then (i) also holds. We proceed by contradiction. If (i) does not hold then there is some $i, j \in S$ and $\epsilon > 0$ such that, for infinitely many $n \geq n(i) \vee n(j)$, there exist $a_n \in A_n(i)$ and $b_n \in B_n(i)$ with

$$|q^n_{ij}(a_n, b_n) - q_{ij}(a_n, b_n)| > \varepsilon. \tag{3.2.1}$$

For such n, since $a_n \in A_n(i) \subseteq A(i)$, there exists a subsequence $\{n'\}$ and $a \in A(i)$ such that $a_{n'} \to a$. Similarly, for some subsequence, still denoted by $\{n'\}$, we have $b_{n'} \to b$ for some $b \in B(i)$. Next, define $\tilde{a}_n \in A_n(i)$ and $\tilde{b}_n \in B_n(i)$, for $n \geq n(i)$, as follows.

- If n belongs to the subsequence $\{n'\}$ then let $\tilde{a}_n = a_n$ and $\tilde{b}_n = b_n$;

- Otherwise, let $\tilde{a}_n \in A_n(i)$ and $\tilde{b}_n \in B_n(i)$ be such that

$$d_A(\tilde{a}_n, a) \leq \inf_{x \in A_n(i)} d_A(x, a) + \frac{1}{n} \leq \rho_A(A_n(i), A(i)) + \frac{1}{n}$$
$$d_B(\tilde{b}_n, b) \leq \inf_{y \in B_n(i)} d_B(y, b) + \frac{1}{n} \leq \rho_B(B_n(i), B(i)) + \frac{1}{n}.$$

We have thus constructed sequences $\tilde{a}_n \in A_n(i)$ and $\tilde{b}_n \in B_n(i)$, for $n \geq n(i)$, such that $\tilde{a}_n \to a$ and $\tilde{b}_n \to b$. Consequently, by (c), for n large enough we have

$$|q_{ij}^n(\tilde{a}_n, \tilde{b}_n) - q_{ij}(a, b)| \leq \frac{\epsilon}{2}.$$

In particular, recalling (3.2.1), along the subsequence $\{n'\}$ we have

$$|q_{ij}(a_{n'}, b_{n'}) - q_{ij}(a, b)| > \frac{\varepsilon}{2}.$$

This contradicts the continuity of the transition rate function.

Conversely, let us now prove that Definition 3.2.1(a), (b), and (d), together with (i), imply (c). Fix $i, j \in S$, and let $a_n \in A_n(i)$ and $b_n \in B_n(i)$ be such that $a_n \to a \in A(i)$ and $b_n \to b \in B(i)$. By the condition (i), given $\epsilon > 0$, for n large enough we have

$$|q_{ij}^n(a_n, b_n) - q_{ij}(a_n, b_n)| \leq \frac{\epsilon}{2}.$$

But now continuity of the function $(a, b) \mapsto q_{ij}(a, b)$ implies that for n large enough we also have

$$|q_{ij}(a_n, b_n) - q_{ij}(a, b)| \leq \frac{\epsilon}{2}.$$

This yields

$$|q_{ij}^n(a_n, b_n) - q_{ij}(a, b)| \leq \epsilon$$

and so $\lim_n q_{ij}^n(a_n, b_n) = q_{ij}(a, b)$. This completes the proof that (c) $\Leftrightarrow$ (i).

To prove statement (ii) we can proceed similarly. □

Given a sequence of functions $u_n : S_n \to \mathbb{R}$, for $n \geq 1$, we say that $\{u_n\}$ converges pointwise to some function $u : S \to \mathbb{R}$ when

$$\lim_{n \to \infty} u_n(i) = u(i) \quad \text{for all } i \in S.$$

Note that, for fixed $i \in S$, $u_n(i)$ is well defined only when $n \geq n(i)$. Since the above definition is concerned with the limit as $n \to \infty$, the requirement $n \geq n(i)$ will not be explicit in the notation.

We introduce some more terminology. Given a pair of randomized stationary strategies $(\pi_n^1, \pi_n^2) \in \Pi_n^{1,s} \times \Pi_n^{2,s}$ for the game model $\mathcal{G}_n$, for $n \geq 1$, we say that the randomized

stationary strategies $(\pi^1, \pi^2) \in \Pi_s^1 \times \Pi_s^2$ are a limit strategy of $\{(\pi_n^1, \pi_n^2)\}_{n\geq 1}$ if there exists a subsequence $\{n'\}$ such that

$$\pi_{n'}^1(\cdot|i) \xrightarrow{d} \pi^1(\cdot|i) \quad \text{and} \quad \pi_{n'}^2(\cdot|i) \xrightarrow{d} \pi^2(\cdot|i)$$

for all $i \in S$. Under the assumption that the action sets $A(i)$ and $B(i)$ are compact, every such sequence $\{(\pi_n^1, \pi_n^2)\}$ indeed has a limit strategy because $\pi_n^1 \in \overline{A}(i)$ and $\pi_n^2 \in \overline{B}(i)$, which are compact metric spaces with the Wasserstein metric (recall Section 1.4).

3.2.2 The discounted payoff case

Suppose for the moment that we are interested in analyzing the discounted payoff optimality criterion for the game model $\mathcal{G}$. For each game model $\mathcal{G}_n$ we will analyze as well the discounted payoff criterion.

The discount rate $\alpha > 0$ is the same for all the game models $\mathcal{G}_n$ and $\mathcal{G}$. Supposing that the Assumptions 3.1.2, 3.1.4, and 3.1.7 are satisfied, next we state our hypotheses on the sequence $\{\mathcal{G}_n\}_{n\geq 1}$.

Assumption 3.2.3 *The following statements hold for every $n \geq 1$.*

(i) For all $(i, a, b) \in \mathbb{K}_n$

$$\sum_{j\in S_n} q_{ij}^n(a, b)w(j) \leq -c_1 w(i) + d_1,$$

where the Lyapunov function w and the constants $c_1 > -\alpha$ and $d_1 \geq 0$ come from Assumption 3.1.2. For each $i \in S_n$, we have $q_n(i) \leq w(i)$.

(ii) For the constant $M > 0$ in Assumption 3.1.4(ii) we have

$$|r_n(i, a, b)| \leq Mw(i) \quad \text{for all } (i, a, b) \in \mathbb{K}_n.$$

(iii) For each $i \in S_n$, the sets $A_n(i) \subseteq A(i)$ and $B_n(i) \subseteq B(i)$ are compact, while for all $i, j \in S_n$, the functions $r_n(i, a, b)$ and $q_{ij}^n(a, b)$ are continuous on $A_n(i) \times B_n(i)$.

(iv) With $c_2 \in \mathbb{R}$ and $d_2 \geq 0$ as in Assumption 3.1.7(iii), we have

$$\sum_{j\in S_n} q_{ij}^n(a, b)w^2(j) \leq -c_2 w^2(i) + d_2 \quad \text{for all } (i, a, b) \in \mathbb{K}_n.$$

We can say, roughly, that Assumption 3.2.3 consists in supposing that Assumptions 3.1.2, 3.1.4, and 3.1.7 hold "uniformly" in $n \geq 1$.

We can apply Theorem 3.1.3 to $\mathcal{G}_n$ and, therefore, there indeed exists a stochastic process $\{(x(t), a(t), b(t))\}_{t\geq 0}$ taking values in $\mathbb{K}_n$ that models the state and actions processes for the game model $\mathcal{G}_n$. In particular, the corresponding expectation operator will

be denoted by E_n^{i,π^1,π^2}. Given $i \in S_n$ and $(\pi^1, \pi^2) \in \Pi_n^1 \times \Pi_n^2$, define the total expected discounted payoff for the game model $\mathcal{G}_n$ as

$$V_n^\alpha(i, \pi^1, \pi^2) = E_n^{i,\pi^1,\pi^2}\Big[\int_0^\infty e^{-\alpha t} r_n(x(t), a(t), b(t))dt\Big].$$

We also have (cf. (3.1.5)),

$$\|V_n^\alpha(\cdot, \pi^1, \pi^2)\| \leq \mathfrak{M} \quad \text{for all } \pi^1 \in \Pi_n^1 \text{ and } \pi^2 \in \Pi_n^2. \tag{3.2.2}$$

The lower and upper value functions of the game L_n^α and U_n^α in $\mathcal{B}_w(S_n)$, and the value function V_n^α (provided it exists) are given the usual definitions. We have a result similar to Lemma 3.1.8, which is stated without proof.

Lemma 3.2.4 *Suppose that Assumption 3.2.3 holds and fix $n \geq 1$.*

(i) Given $i \in S_n$ and $k > i$, we have

$$\sum_{j \geq k, j \in S_n} q_{ij}^n(\varphi, \psi) w(j) \leq \frac{1}{w(k)}\Big(- c_2 w^2(i) + d_2 + q(i) w^2(i)\Big)$$

for all $\varphi \in \overline{A}_n(i)$ and $\psi \in \overline{B}_n(i)$.

(ii) For every $i \in S_n$ and $u \in \mathcal{B}_w(S_n)$, the functions

$$(\varphi, \psi) \mapsto r_n(i, \varphi, \psi) \quad \text{and} \quad (\varphi, \psi) \mapsto \sum_{j \in S_n} q_{ij}^n(\varphi, \psi) u(j)$$

are continuous on $\overline{A}_n(i) \times \overline{B}_n(i)$.

Our next lemma states a useful continuity result.

Lemma 3.2.5 *Suppose that the game models $\mathcal{G}$ and $\{\mathcal{G}_n\}_{n \geq 1}$ satisfy Assumptions 3.1.2, 3.1.4, 3.1.7, and 3.2.3, and also that $\mathcal{G}_n \to \mathcal{G}$. Suppose that the sequence $v_n \in \mathcal{B}_w(S_n)$, for $n \geq 1$, converges pointwise to $v \in \mathcal{B}_w(S)$, and that*

$$\sup_{n \geq 1} ||v_n||_w = \mathfrak{m} < \infty.$$

For fixed $i \in S$, assume also that $\varphi_n \in \overline{A}_n(i)$ and $\psi_n \in \overline{B}_n(i)$, for $n \geq n(i)$, are such that

$$\varphi_n \xrightarrow{d} \varphi \quad \text{and} \quad \psi_n \xrightarrow{d} \psi \quad \text{as } n \to \infty$$

for some $\varphi \in \overline{A}(i)$ and $\psi \in \overline{B}(i)$. Under these conditions,

$$\lim_{n \to \infty} \big[r_n(i, \varphi_n, \psi_n) + \sum_{j \in S_n} q_{ij}^n(\varphi_n, \psi_n) v_n(j)\big] = r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) v(j).$$

Proof. Let us first analyze the term $r_n(i,\varphi_n,\psi_n)$. By Lemma 3.2.2(ii), given $\epsilon>0$ there exists $n_0\geq n(i)$ such that $n\geq n_0$ implies

$$|r_n(i,a,b)-r(i,a,b)|\leq\frac{\epsilon}{2}\quad\text{for all }(a,b)\in A_n(i)\times B_n(i).$$

In particular, we have

$$|r_n(i,\varphi_n,\psi_n)-r(i,\varphi_n,\psi_n)|\leq\int_{A_n(i)}\int_{B_n(i)}|r_n(i,a,b)-r(i,a,b)|\psi_n(db)\varphi_n(da)\leq\frac{\epsilon}{2}\tag{3.2.3}$$

for $n\geq n_0$. On the other hand, by Lemma 3.1.8(ii) we have that $r(i,\varphi_n,\psi_n)$ converges to $r(i,\varphi,\psi)$ as $n\to\infty$. Consequently, there is some $n_1\geq n(i)$ such that $n\geq n_1$ gives

$$\big|r(i,\varphi_n,\psi_n)-r(i,\varphi,\psi)\big|\leq\frac{\epsilon}{2}.\tag{3.2.4}$$

From (3.2.3) and (3.2.4) we have that $|r_n(i,\varphi_n,\psi_n)-r(i,\varphi,\psi)|\leq\epsilon$ for $n\geq n_0\vee n_1$. Therefore,

$$\lim_{n\to\infty}r_n(i,\varphi_n,\psi_n)=r(i,\varphi,\psi).$$

We proceed with the proof. As a consequence of Lemmas 3.1.8(i) and 3.2.4(i) we deduce that, given $\epsilon>0$, there exists some $k>i$ such that $\sum_{j\geq k}q_{ij}(\varphi,\psi)w(j)\leq\epsilon$ and such that, for all $n\geq n(i)$,

$$\sum_{j\geq k,j\in S_n}q_{ij}^n(\varphi_n,\psi_n)w(j)\leq\epsilon.$$

Therefore, since $\|v_n\|_w\leq\mathfrak{m}$ implies $\|v\|_w\leq\mathfrak{m}$, we have

$$\Big|\sum_{j\geq k}q_{ij}(\varphi,\psi)v(j)\Big|\leq\mathfrak{m}\epsilon$$

and, for all $n\geq n(i)$,

$$\Big|\sum_{j\geq k,j\in S_n}q_{ij}^n(\varphi_n,\psi_n)v_n(j)\Big|\leq\mathfrak{m}\epsilon.$$

Consequently, if $n\geq n(i)$ is such that, in addition, $\{0,1,\ldots,k-1\}\subseteq S_n$, we have

$$\Big|\sum_{j\in S_n}q_{ij}^n(\varphi_n,\psi_n)v_n(j)-\sum_{j\in S}q_{ij}(\varphi,\psi)v(j)\Big|\leq\sum_{j=0}^{k-1}\big|q_{ij}^n(\varphi_n,\psi_n)v_n(j)-q_{ij}(\varphi,\psi)v(j)\big|+2\mathfrak{m}\epsilon.$$

The left-hand term of the last expression can be made arbitrarily small by choosing n large enough. Indeed, as made at the beginning of this proof, we can prove that

$$q_{ij}^n(\varphi_n,\psi_n)\to q_{ij}(\varphi,\psi)$$

which, together with the fact that $v_n(j)\to v(j)$ for all $j\in S$, yields the stated result because, once $i\in S$ and $\epsilon>0$ are given, the state k remains fixed and does not depend on n. The proof is now complete. □

Each discounted game model $\mathcal{G}_n$ has a value function that can be characterized by the corresponding Shapley equations; cf. Theorem 3.1.9.

Theorem 3.2.6 *Suppose that Assumption 3.2.3 is satisfied. Then the following statements hold for each $n \geq 1$.*

(i) The game $\mathcal{G}_n$ has a value $V_n^\alpha \in \mathcal{B}_w(S_n)$ with $\|V_n^\alpha\|_w \leq \mathfrak{M}$.

(ii) The value function V_n^α is the unique solution u in $\mathcal{B}_w(S_n)$ of the equations

$$\alpha u(i) = \sup_{\varphi \in \overline{A}_n(i)} \inf_{\psi \in \overline{B}_n(i)} \Big\{ r_n(i,\varphi,\psi) + \sum_{j \in S_n} q_{ij}^n(\varphi,\psi) u(j) \Big\} \tag{3.2.5}$$

$$= \inf_{\psi \in \overline{B}_n(i)} \sup_{\varphi \in \overline{A}_n(i)} \Big\{ r_n(i,\varphi,\psi) + \sum_{j \in S_n} q_{ij}^n(\varphi,\psi) u(j) \Big\} \tag{3.2.6}$$

for each $i \in S_n$.

(iii) There exists a pair of optimal randomized stationary strategies for the game model $\mathcal{G}_n$. Moreover, $(\pi^1, \pi^2) \in \Pi_n^{1,s} \times \Pi_n^{2,s}$ is a pair of optimal randomized stationary strategies if and only if $\pi^1(\cdot|i)$ and $\pi^2(\cdot|i)$ attain the supremum and the infimum in (3.2.5) and (3.2.6), respectively, for every $i \in S_n$.

3.2.3 The average payoff case

Suppose now that we are interested in the average optimality criterion for the game model $\mathcal{G}$. In this case, we will also consider the long-run average payoff criterion for the approximating game models $\mathcal{G}_n$, for $n \geq 1$. The hypotheses we made on the average game model $\mathcal{G}$ were Assumptions 3.1.11, 3.1.12, and 3.1.14. We will impose the following conditions on the game models $\mathcal{G}_n$.

Assumption 3.2.7 *The following statements hold for every $n \geq 1$.*

(i) For all $(i,a,b) \in \mathbb{K}_n$

$$\sum_{j \in S_n} q_{ij}^n(a,b) w(j) \leq -c_1 w(i) + d_1 \mathbf{I}_{D_n}(i),$$

where the Lyapunov function w and the constants $c_1 > 0$ and $d_1 \geq 0$ come from Assumption 3.1.11, and $D_n \subseteq S_n$ is a finite set. For each $i \in S_n$, we have $q_n(i) \leq w(i)$.

(ii) For the constant $M > 0$ in Assumption 3.1.12 we have

$$|r_n(i,a,b)| \leq M w(i) \quad \text{for all } (i,a,b) \in \mathbb{K}_n.$$

(iii) For each $i \in S_n$, the sets $A_n(i) \subseteq A(i)$ and $B_n(i) \subseteq B(i)$ are compact, while for all $i,j \in S_n$, the functions $r_n(i,a,b)$ and $q_{ij}^n(a,b)$ are continuous on $A_n(i) \times B_n(i)$.

(iv) With $c_2 \in \mathbb{R}$ and $d_2 \geq 0$ as in Assumption 3.1.14(iii), we have

$$\sum_{j \in S_n} q_{ij}^n(a,b) w^2(j) \leq -c_2 w^2(i) + d_2 \quad \textit{for all } (i,a,b) \in \mathbb{K}_n.$$

(v) Each pair of strategies in $\Pi_n^{1,s} \times \Pi_n^{2,s}$ is irreducible.

Once again, these conditions consist in assuming the the hypotheses for the game model $\mathcal{G}$ hold "uniformly" in $n \geq 1$ for the game models $\mathcal{G}_n$. It is worth noting that Assumption 3.2.7 implies Assumption 3.2.3. In particular, Lemmas 3.2.4 and 3.2.5 remain valid under Assumption 3.2.7.

For the game model $\mathcal{G}_n$, the long-run expected average payoff of the players, when starting from the initial state $i \in S_n$, and using the strategies $\pi^1 \in \Pi_n^1$ and $\pi^2 \in \Pi_n^2$, is defined as

$$J_n(i,\pi^1,\pi^2) = \limsup_{T \to \infty} \frac{1}{T} E_n^{i,\pi^1,\pi^2} \left[\int_0^T r_n(x(t),a(t),b(t)) dt \right].$$

The long-run average lower and upper value functions of the game $\mathcal{G}_n$ are respectively defined as

$$L_n(i) = \sup_{\pi^1 \in \Pi_n^1} \inf_{\pi^2 \in \Pi_n^2} J_n(i,\pi^1,\pi^2) \;\text{ and }\; U(i) = \inf_{\pi^2 \in \Pi_n^2} \sup_{\pi^1 \in \Pi_n^1} J_n(i,\pi^1,\pi^2)$$

for each $i \in S$. If $L_n(i) = U_n(i) =: V_n(i)$ for all $i \in S$ then we say that the game $\mathcal{G}_n$ has a value, and V_n is the value function of the game.

As a consequence of Assumptions 3.2.7(i)–(ii), the lower and upper value functions of the game are finite and they have the same bounds as the value functions of $\mathcal{G}$, recall (3.1.11),

$$|L_n(i)| \leq \frac{M d_1}{c_1} \quad \text{and} \quad |U_n(i)| \leq \frac{M d_1}{c_1}, \quad \text{for all } i \in S_n.$$

The next theorem is derived directly from Theorem 3.1.16. It just states that every game model $\mathcal{G}_n$ has a constant value function that can be characterized by means of the corresponding optimality equations.

Theorem 3.2.8 *Suppose that Assumption 3.2.7 holds and fix $n \geq 1$.*

(i) *The average game $\mathcal{G}_n$ has a constant value $g_n^* \in \mathbb{R}$, with $|g_n^*| \leq M d_1/c_1$.*

(ii) *There exist solutions $(g,h) \in \mathbb{R} \times \mathcal{B}_w(S_n)$ to the average optimality equations for $\mathcal{G}_n$*

$$\begin{aligned} g &= \sup_{\varphi \in \overline{A}_n(i)} \inf_{\psi \in \overline{B}_n(i)} \Big\{ r_n(i,\varphi,\psi) + \sum_{j \in S_n} q_{ij}^n(\varphi,\psi) h(j) \Big\} \\ &= \inf_{\psi \in \overline{B}_n(i)} \sup_{\varphi \in \overline{A}_n(i)} \Big\{ r_n(i,\varphi,\psi) + \sum_{j \in S_n} q_{ij}^n(\varphi,\psi) h(j) \Big\} \end{aligned}$$

for $i \in S_n$.

If $(g,h) \in \mathbb{R} \times \mathcal{B}_w(S_n)$ is a solution to the average optimality equations then $g = g_n^*$, the value of the game, and h is unique up to additive constants.

(iii) *There exists a pair of optimal stationary strategies. A pair of stationary strategies* $(\pi^1, \pi^2) \in \Pi_n^{1,s} \times \Pi_n^{2,s}$ *is average optimal if and only if they attain the supremum and the infimum in the average optimality equations.*

We note that, in Theorem 3.1.16, we mentioned the existence of a function $h \in \mathcal{B}_w(S)$, solution of the average optimality equations, such that $||h||_w \leq RM/\gamma$, where the positive constants R and γ come from the uniform exponential ergodicity property; recall (3.1.12). For the approximating game models $\mathcal{G}_n$, we cannot say that the constants R_n, γ_n for the game model $\mathcal{G}_n$ (which indeed is uniformly exponentially ergodic) do not depend on $n \geq 1$. However, if each game model $\mathcal{G}_n$ satisfies the conditions given in Remark 3.1.15, then we deduce the existence of solutions $h_n \in \mathcal{B}_w(S_n)$ with $||h_n||_w \leq RM/\gamma$, the same bound for $\mathcal{G}$ and every $\mathcal{G}_n$.

3.3 Approximation results for discounted games

3.3.1 Convergence results: the general case

Now we prove our main result on the convergence of discounted game models.

Theorem 3.3.1 *Suppose that the game models* $\mathcal{G}$ *and* $\{\mathcal{G}_n\}_{n\geq 1}$ *satisfy Assumptions 3.1.2, 3.1.4, 3.1.7, and 3.2.3. If* $\mathcal{G}_n \to \mathcal{G}$ *then the following statements are satisfied.*

(i) For all $i \in S$, $\lim_{n\to\infty} V_n^\alpha(i) = V^\alpha(i)$.

(ii) If (π_n^1, π_n^2) *is a pair of optimal randomized stationary strategies for the game model* $\mathcal{G}_n$, *then any limit strategy* $(\pi^1, \pi^2) \in \Pi_s^1 \times \Pi_s^2$ *is a pair of optimal randomized stationary strategies for the game model* $\mathcal{G}$.

Proof. (i). Recall that the sequence $\{V_n^\alpha\}$ of the values of the games $\mathcal{G}_n$ verifies

$$|V_n^\alpha(i)| \leq \mathfrak{M}w(i) \quad \text{for all } n \geq 1 \text{ and } i \in S_n;$$

see (3.2.2). Therefore, by using a diagonal argument, we deduce the existence of $u \in \mathcal{B}_w(S)$ and a subsequence $\{k_n\}$ such that

$$\lim_{n\to\infty} V_{k_n}^\alpha(i) = u(i) \quad \text{for all } i \in S.$$

Fix $i \in S$ and, for n such that $k_n \geq n(i)$, consider the function on $\overline{A}_{k_n}(i) \times \overline{B}_{k_n}(i)$

$$(\varphi, \psi) \mapsto r_{k_n}(i, \varphi, \psi) + \sum_{j\in S_{k_n}} q_{ij}^{k_n}(\varphi, \psi) V_{k_n}^\alpha(j).$$

This function is continuous as a consequence of Lemma 3.2.4. Therefore,

$$\varphi \mapsto \inf_{\psi\in\overline{B}_{k_n}(i)} \Big\{ r_{k_n}(i, \varphi, \psi) + \sum_{j\in S_{k_n}} q_{ij}^{k_n}(\varphi, \psi) V_{k_n}^\alpha(j) \Big\}$$

is upper semi-continuous on the compact set $\overline{A}_{k_n}(i)$ and, hence, it has a maximum which is reached at some $\varphi_n \in \overline{A}_{k_n}(i)$. There exists a further subsequence $\{k_{n'}\}$ such that $\varphi_{n'} \xrightarrow{d} \varphi_0$ for some $\varphi_0 \in \overline{A}(i)$. Without loss of generality, and to simplify the notation, we will suppose that the whole sequence $\{\varphi_n\}$ converges to φ_0.

Fix now arbitrary $\psi \in \overline{B}(i)$. For each n there exist

$$x_1, \ldots, x_t \in B(i) \quad \text{and} \quad \beta_1, \ldots, \beta_t \in [0,1]$$

with $\sum \beta_j = 1$ such that $d_W(\psi, \hat{\psi}_n) \leq 1/n$, with $\hat{\psi}_n = \sum \beta_j \delta_{x_j}$. Let $y_j \in B_{k_n}(i)$ be such that $d_B(y_j, x_j) = \min_{y \in B_{k_n}(i)} d_B(y, x_j)$ for each $j = 1, \ldots, t$, and define

$$\tilde{\psi}_n = \sum_{j=1}^{t} \beta_j \delta_{y_j} \in \overline{B}_{k_n}(i).$$

If f is a bounded L-Lipschitz continuous function on $B(i)$ then we have

$$\Big| \int f d\tilde{\psi}_n - \int f d\psi \Big| \leq \Big| \int f d\tilde{\psi}_n - \int f d\hat{\psi}_n \Big| + \Big| \int f d\hat{\psi}_n - \int f d\psi \Big|.$$

We note that

$$\begin{aligned} \Big| \int f d\tilde{\psi}_n - \int f d\hat{\psi}_n \Big| &= \Big| \sum_{j=1}^{t} \beta_j [f(y_j) - f(x_j)] \Big| \leq L \sum_{j=1}^{t} \beta_j d_B(y_j, x_j) \\ &\leq L \rho_B(B_{k_n}(i), B(i)), \end{aligned}$$

which converges to 0 as $n \to \infty$. On the other hand, since $\hat{\psi}_n \xrightarrow{d} \psi$ we have $\int f d\hat{\psi}_n \to \int f d\psi$. So, we have shown that

$$\lim_{n \to \infty} \int f d\tilde{\psi}_n = \int f d\psi$$

for all bounded and Lipschitz-continuous functions on $B(i)$. This implies that $\tilde{\psi}_n \xrightarrow{d} \psi$. Summarizing, given arbitrary $\psi \in \overline{B}(i)$ we have constructed $\tilde{\psi}_n \in \overline{B}_{k_n}(i)$ such that $\{\tilde{\psi}_n\}$ converges weakly to ψ.

By Theorem 3.2.6(ii), the value $V^{\alpha}_{k_n}$ of the game $\mathcal{G}_{k_n}$ verifies

$$\begin{aligned} \alpha V^{\alpha}_{k_n}(i) &= \sup_{\varphi \in \overline{A}_{k_n}(i)} \inf_{\psi \in \overline{B}_{k_n}(i)} \Big\{ r_{k_n}(i, \varphi, \psi) + \sum_{j \in S_{k_n}} q^{k_n}_{ij}(\varphi, \psi) V^{\alpha}_{k_n}(j) \Big\} \\ &= \inf_{\psi \in \overline{B}_{k_n}(i)} \Big\{ r_{k_n}(i, \varphi_n, \psi) + \sum_{j \in S_{k_n}} q^{k_n}_{ij}(\varphi_n, \psi) V^{\alpha}_{k_n}(j) \Big\} \\ &\leq r_{k_n}(i, \varphi_n, \tilde{\psi}_n) + \sum_{j \in S_{k_n}} q^{k_n}_{ij}(\varphi_n, \tilde{\psi}_n) V^{\alpha}_{k_n}(j). \end{aligned}$$

Taking the limit as $n \to \infty$ and recalling Lemma 3.2.5, we obtain

$$\alpha u(i) \leq r(i, \varphi_0, \psi) + \sum_{j \in S} q_{ij}(\varphi_0, \psi) u(j).$$

But $\psi \in \overline{B}(i)$ being arbitrary, we conclude that

$$\alpha u(i) \leq \inf_{\psi \in \overline{B}(i)} \Big\{ r(i, \varphi_0, \psi) + \sum_{j \in S} q_{ij}(\varphi_0, \psi) u(j) \Big\},$$

and so,

$$\alpha u(i) \leq \sup_{\varphi \in \overline{A}(i)} \inf_{\psi \in \overline{B}(i)} \Big\{ r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) u(j) \Big\}.$$

Arguing similarly, we can show that

$$\alpha u(i) \geq \inf_{\psi \in \overline{B}(i)} \sup_{\varphi \in \overline{A}(i)} \Big\{ r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) u(j) \Big\}.$$

Combining these two inequalities, we conclude that

$$\begin{aligned} \alpha u(i) &= \sup_{\varphi \in \overline{A}(i)} \inf_{\psi \in \overline{B}(i)} \Big\{ r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) u(j) \Big\} \\ &= \inf_{\psi \in \overline{B}(i)} \sup_{\varphi \in \overline{A}(i)} \Big\{ r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) u(j) \Big\} \end{aligned}$$

for each $i \in S$. By Theorem 3.1.9(ii), this implies that u equals V^α, the value of the game $\mathcal{G}$.

So far, we have shown that if u is any limit point of $\{V_n^\alpha\}$ then, necessarily, $u = V^\alpha$. But this implies that $\lim_{n \to \infty} V_n^\alpha(i) = V^\alpha(i)$ for all $i \in S$. The proof of (i) is now complete.

(ii). Suppose that $(\pi^1, \pi^2) \in \Pi_s^1 \times \Pi_s^2$ is a limit strategy through the subsequence $\{n'\}$. Fix $i \in S$ and write

$$\varphi_n^* = \pi_n^1(\cdot|i), \quad \psi_n^* = \pi_n^2(\cdot|i), \quad \varphi^* = \pi^1(\cdot|i), \quad \psi^* = \pi^2(\cdot|i)$$

for $n \geq n(i)$. Then we have

$$\varphi_{n'}^* \xrightarrow{d} \varphi^* \quad \text{and} \quad \psi_{n'}^* \xrightarrow{d} \psi^*.$$

For $n \geq n(i)$, we know that $\varphi_{n'}^*$ and $\psi_{n'}^*$ attain the supremum and the infimum in the Shapley equation for $\mathcal{G}_{n'}$ for the state i; recall Theorem 3.2.6(iii). Therefore,

$$\begin{aligned} \alpha V_{n'}^\alpha(i) &= \sup_{\varphi \in \overline{A}_{n'}(i)} \inf_{\psi \in \overline{B}_{n'}(i)} \Big\{ r_{n'}(i, \varphi, \psi) + \sum_{j \in S_{n'}} q_{ij}^{n'}(\varphi, \psi) V_{n'}^\alpha(j) \Big\} \\ &= \inf_{\psi \in \overline{B}_{n'}(i)} \Big\{ r_{n'}(i, \varphi_{n'}^*, \psi) + \sum_{j \in S_{n'}} q_{ij}^{n'}(\varphi_{n'}^*, \psi) V_{n'}^\alpha(j) \Big\}. \end{aligned} \tag{3.3.1}$$

Proceeding as in the proof of part (i), we can show that for every $\psi \in \overline{B}(i)$ there exists a sequence $\psi_{n'} \in \overline{B}_{n'}(i)$ such that $\psi_{n'} \xrightarrow{d} \psi$, and so, by (3.3.1),

$$\alpha V^{\alpha}_{n'}(i) \;\leq\; r_{n'}(i, \varphi^*_{n'}, \psi_{n'}) + \sum_{j \in S_{n'}} q^{n'}_{ij}(\varphi^*_{n'}, \psi_{n'}) V^{\alpha}_{n'}(j).$$

Taking the limit as $n' \to \infty$ and recalling that V^{α}_n converges pointwise to V^{α} (part (i) of this theorem), gives

$$\alpha V^{\alpha}(i) \leq r(i, \varphi^*, \psi) + \sum_{j \in S} q_{ij}(\varphi^*, \psi) V^{\alpha}(j).$$

Since $\psi \in \overline{B}(i)$ is arbitrary, we have

$$\alpha V^{\alpha}(i) \leq \inf_{\psi \in \overline{B}(i)} \Big\{ r(i, \varphi^*, \psi) + \sum_{j \in S} q_{ij}(\varphi^*, \psi) V^{\alpha}(j) \Big\}.$$

But from the Shapley equation for $\mathcal{G}$ we know that

$$\alpha V^{\alpha}(i) = \sup_{\varphi \in \overline{A}(i)} \inf_{\psi \in \overline{B}(i)} \Big\{ r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) V^{\alpha}(j) \Big\}.$$

Hence, φ^* attains the supremum in the Shapley equation for $i \in S$.

Similarly, ψ^* attains the infimum in the Shapley equation for $\mathcal{G}$ and $i \in S$ and, by Theorem 3.1.9(iii), this implies that $(\pi^1, \pi^2) \in \Pi^1_s \times \Pi^2_s$ is indeed a pair of optimal strategies for $\mathcal{G}$. □

Theorem 3.3.1 above proposes a general convergence result but it is not ready yet for numerical applications. Next, we show how to construct finite state and actions game models $\mathcal{G}_n$ starting from the original game model $\mathcal{G}$.

3.3.2 Convergence results: finite approximations

Given a game model $\mathcal{G}$ satisfying Assumptions 3.1.2, 3.1.4, and 3.1.7 we now show how to construct a sequence of game models $\{\mathcal{G}_n\}_{n \geq 1}$ for which Assumption 3.2.3 holds. For each $n \geq 1$, the elements of the game model $\mathcal{G}_n$ are the following.

- The state space is $S_n = \{0, 1, \ldots, n\}$.
- For $i \in S_n$, let $A_n(i)$ and $B_n(i)$ be finite subsets of $A(i)$ and $B(i)$, respectively, such that
$$\rho_A(A_n(i), A(i)) \to 0 \quad \text{and} \quad \rho_B(B_n(i), B(i)) \to 0$$
as $n \to \infty$.

- Given $i \in S_n$ and $0 \leq j < n$, define $q_{ij}^n(a,b) = q_{ij}(a,b)$, and let

$$q_{in}^n(a,b) = \sum_{j \geq n} q_{ij}(a,b) = -\sum_{j=0}^{n-1} q_{ij}(a,b)$$

  for $(a,b) \in A_n(i) \times B_n(i)$.

- The reward/cost rate is $r_n(i,a,b) = r(i,a,b)$ for $(a,b) \in A_n(i) \times B_n(i)$.

We note that construction of $A_n(i)$ and $B_n(i)$ with the property of convergence in the Hausdorff metric is indeed possible. For instance, for each $n \geq 1$, consider the open cover of $A(i)$ given by the open balls centered in $a \in A(i)$ with radius $1/n$, and let $A_n(i)$ be the centers of a finite subcover. Then $\rho_A(A_n(i), A(i)) \leq 1/n$.

Theorem 3.3.2 *If the game model $\mathcal{G}$ satisfies Assumptions 3.1.2, 3.1.4, and 3.1.7 then the sequence $\{\mathcal{G}_n\}_{n \geq 1}$ defined above satisfies Assumption 3.2.3 and, moreover, $\mathcal{G}_n \to \mathcal{G}$.*

Proof: First of all, we observe that the transition rates of $\mathcal{G}_n$ are conservative:

$$\sum_{j \in S_n} q_{ij}^n(a,b) = \sum_{j \in S} q_{ij}(a,b) = 0 \quad \text{for all } (i,a,b) \in \mathbb{K}_n,$$

and stable. Indeed, $-q_{ii}^n(a,b) = -q_{ii}(a,b) \leq w(i)$ for $i < n$ and

$$-q_{nn}^n(a,b) = -\sum_{j \geq n} q_{nj}(a,b) \leq -q_{nn}(a,b) \leq w(n).$$

Concerning Assumption 3.2.3(i), observe that for all $(i,a,b) \in \mathbb{K}_n$

$$\begin{aligned} \sum_{j \in S_n} q_{ij}^n(a,b) w(j) &= \sum_{j \in S_n} q_{ij}(a,b) w(j) + \sum_{j > n} q_{ij}(a,b) w(n) \\ &\leq \sum_{j \in S_n} q_{ij}(a,b) w(j) + \sum_{j > n} q_{ij}(a,b) w(j) \\ &= \sum_{j \in S} q_{ij}(a,b) w(j) \leq -c_1 w(i) + d_1, \end{aligned}$$

where we make use of the monotonicity of w. The fact that $q_n(i) \leq w(i)$ has been established along with the stability of the transition rates of $\mathcal{G}_n$. So, Assumption 3.2.3(i) indeed holds.

Clearly, Assumptions 3.2.3(ii)–(iii) are also satisfied, while Assumption 3.2.3(iv) is proved similarly to Assumption 3.2.3(i).

It remains to check that $\mathcal{G}_n \to \mathcal{G}$. Items (a) and (b) in Definition 3.2.1 hold by construction of $\mathcal{G}_n$. Finally, given $i, j \in S$, if $(a_n, b_n) \in A_n(i) \times B_n(i)$ are such that $a_n \to a \in A(i)$ and $b_n \to b \in B(i)$ then

$$r_n(i, a_n, b_n) = r(i, a_n, b_n) \quad \text{and} \quad q_{ij}^n(a_n, b_n) = q_{ij}(a_n, b_n)$$

for $n > i \vee j$, and so Definitions 3.2.1(c)–(d) hold by continuity of the transition and reward/cost rates of $\mathcal{G}$. □

As a consequence of Theorem 3.3.1, the value functions of the finite state and actions games $\mathcal{G}_n$ converge to the value function of $\mathcal{G}$, and any limit strategy of optimal stationary strategies for $\mathcal{G}_n$ is optimal for $\mathcal{G}$.

Next, we address the issue of the rate of convergence of $V_n^\alpha(i)$ to $V^\alpha(i)$. To establish such convergence rates, we need to strengthen our hypotheses. Namely, Assumption 3.1.7 will be replaced with the following stronger condition.

Assumption 3.3.3 *The game model $\mathcal{G}$ satisfies the following conditions.*

(i) For each $i \in S$, the sets $A(i)$ and $B(i)$ are compact.

(ii) For each $i, j \in S$, the functions $(a,b) \mapsto r(i,a,b)$ and $(a,b) \mapsto q_{ij}(a,b)$ are L_i- and L_{ij}-Lipschitz continuous on $A(i) \times B(i)$, i.e.,

$$\begin{aligned} |r(i,a,b) - r(i,a',b')| &\leq L_i\big(d_A(a,a') + d_B(b,b')\big) \\ |q_{ij}(a,b) - q_{ij}(a',b')| &\leq L_{ij}\big(d_A(a,a') + d_B(b,b')\big) \end{aligned}$$

for all $a, a' \in A(i)$ and $b, b' \in B(i)$, and some $L_i > 0$ and $L_{ij} > 0$.

(iii) With w the Lyapunov function in Assumption 3.1.2, there exist constants $\delta > 2$, $c_\delta > -\alpha$, and $d_\delta \geq 0$ with

$$\sum_{j \in S} q_{ij}(a,b) w^\delta(j) \leq -c_\delta w^\delta(i) + d_\delta \quad \text{for all } (i,a,b) \in \mathbb{K}. \tag{3.3.2}$$

We have that Assumption 3.3.3(iii) is indeed stronger than Assumption 3.1.7(iii). To this end we use the following transcription of Lemma 2.3.5.

Lemma 3.3.4 *Suppose that the function $h : S \to [0,\infty)$ satisfies $q(i) \leq h(i)$ for all $i \in S$. If there exists a power $\gamma > 0$ and a constant $c_\gamma \geq 0$ such that*

$$\sum_{j \in S} q_{ij}(a,b) h^\gamma(j) \leq c_\gamma h^\gamma(i) \quad \text{for all } (i,a,b) \in \mathbb{K}, \tag{3.3.3}$$

then for every power $0 < \gamma' < \gamma$

$$\sum_{j \in S} q_{ij}(a,b) h^{\gamma'}(j) \leq c_\gamma h^{\gamma'}(i) \quad \text{for all } (i,a,b) \in \mathbb{K}.$$

Consequently, if there exists a power $\gamma > 0$ such that the Lyapunov function w verifies

$$\sum_{j \in S} q_{ij}(a,b) w^\gamma(j) \leq -c_\gamma w^\gamma(i) + d_\gamma \quad \text{for all } (i,a,b) \in \mathbb{K}$$

and some constants $c_\gamma \in \mathbb{R}$ and $d_\gamma \geq 0$, then for every $0 < \gamma' < \gamma$

$$\sum_{j \in S} q_{ij}(a,b) w^{\gamma'}(j) \leq (|c_\gamma| + d_\gamma) w^{\gamma'}(i) \quad \text{for all } (i,a,b) \in \mathbb{K}$$

(indeed, just note that $\sum q_{ij}(a,b) w^{\gamma}(j) \leq (|c_\gamma| + d_\gamma) w^{\gamma}(i)$ and use Lemma 3.3.4). In particular, if Assumption 3.3.3(iii) holds then necessarily Assumption 3.1.7(iii) is satisfied. Moreover, the following inequality, which will be used in the sequel, easily follows as well:

$$\sum_{j \in S} q_{ij}(a,b) w^{\delta-1}(j) \leq (|c_\delta| + d_\delta) w^{\delta-1}(i) \quad \text{for all } (i,a,b) \in \mathbb{K}. \tag{3.3.4}$$

Lemma 3.3.5 *Consider a fixed $n \geq 1$ and suppose that the game model $\mathcal{G}_n$ satisfies Assumption 3.2.3. Suppose that there exists a function $u \in \mathcal{B}_w(S_n)$ such that, for all $i \in S_n$,*

$$\left| \alpha u(i) - \sup_{\varphi \in \overline{A}_n(i)} \inf_{\psi \in \overline{B}_n(i)} \{ r_n(i,\varphi,\psi) + \sum_{j \in S_n} q^n_{ij}(\varphi,\psi) u(j) \} \right| \leq h(i)$$

for some $h(i) \geq 0$. Assume, in addition, that there exist constants $c_h > -\alpha$ and $d_h \geq 0$ such that

$$\sum_{j \in S_n} q^n_{ij}(a,b) h(j) \leq -c_h h(i) + d_h \quad \textit{for all } (i,a,b) \in \mathbb{K}_n.$$

Under these conditions,

$$|V^\alpha_n(i) - u(i)| \leq \frac{h(i)}{\alpha + c_h} + \frac{d_h}{\alpha(\alpha + c_h)} \quad \textit{for each } i \in S_n.$$

Proof: First of all, we note that for every $(\pi^1, \pi^2) \in \Pi^1_n \times \Pi^2_n$, $t \geq 0$, and $i \in S_n$ we have

$$E^{i,\pi^1,\pi^2}_n [h(x(t))] \leq e^{-c_h t} h(i) + \frac{d_h}{c_h}(1 - e^{-c_h t}) \quad \text{if } c_h \neq 0$$

or $E^{i,\pi^1,\pi^2}_n [h(x(t))] \leq h(i) + d_h t$ when $c_h = 0$ (the proof of these inequalities is similar to that of (3.1.4)). Therefore, in either case,

$$E^{i,\pi^1,\pi^2}_n \left[\int_0^\infty e^{-\alpha t} h(x(t)) \right] \leq \frac{h(i)}{\alpha + c_h} + \frac{d_h}{\alpha(\alpha + c_h)} \quad \text{for all } i \in S_n. \tag{3.3.5}$$

For every $i \in S_n$, there exists some $\varphi \in \overline{A}_n(i)$ such that for all $\psi \in \overline{B}_n(i)$

$$\alpha u(i) - r_n(i,\varphi,\psi) - \sum_{j \in S_n} q^n_{ij}(\varphi,\psi) u(j) \leq h(i).$$

Consequently, there exists a stationary policy $\pi^1 \in \Pi^{1,s}_n$ such that for every stationary $\pi^2 \in \Pi^{2,s}_n$

$$\alpha u(i) - r_n(i,\pi^1,\pi^2) - \sum_{j \in S_n} q^n_{ij}(\pi^1,\pi^2) u(j) \leq h(i) \quad \text{for all } i \in S_n.$$

Using Dynkin's formula gives, for every $i \in S_n$ and $t \geq 0$,

$$\begin{aligned} E_n^{i,\pi^1,\pi^2}[e^{-\alpha t}u(x(t))] - u(i) &= E_n^{i,\pi^1,\pi^2}\Big[\int_0^t e^{-\alpha s}[-\alpha u(x(s)) + \sum_{j\in S_n} q^n_{x(s)j}(\pi^1,\pi^2)u(j)]ds\Big] \\ &\geq -E_n^{i,\pi^1,\pi^2}\Big[\int_0^t e^{-\alpha s}[r_n(x(s),\pi^1,\pi^2) + h(x(s))]ds\Big]. \end{aligned}$$

Now we let $t \to \infty$ in this inequality. Recalling (3.1.4) and Remark 3.1.5 for the game model $\mathcal{G}_n$, and using dominated and monotone convergence, we obtain

$$\begin{aligned} u(i) &\leq V_n^\alpha(i,\pi^1,\pi^2) + E_n^{i,\pi^1,\pi^2}\Big[\int_0^\infty e^{-\alpha s}h(x(s))ds\Big] \\ &\leq V_n^\alpha(i,\pi^1,\pi^2) + \frac{h(i)}{\alpha + c_h} + \frac{d_h}{\alpha(\alpha+c_h)}, \end{aligned}$$

where we have used (3.3.5). Since this inequality holds for some π^1 and all π^2, we obtain

$$\begin{aligned} u(i) &\leq \sup_{\pi^1\in\Pi_n^{1,s}} \inf_{\pi^2\in\Pi_n^{2,s}} \{V_n^\alpha(i,\pi^1,\pi^2)\} + \frac{h(i)}{\alpha + c_h} + \frac{d_h}{\alpha(\alpha+c_h)} \\ &= V_n^\alpha(i) + \frac{h(i)}{\alpha + c_h} + \frac{d_h}{\alpha(\alpha+c_h)} \end{aligned} \tag{3.3.6}$$

for all $i \in S_n$ (use Remark 3.1.10 for the game model $\mathcal{G}_n$).

Observe now that, the sets $A_n(i)$ and $B_n(i)$ being finite, we have

$$\begin{aligned} &\inf_{\psi\in\overline{B}_n(i)} \sup_{\varphi\in\overline{A}_n(i)} \{r_n(i,\varphi,\psi) + \sum_{j\in S_n} q^n_{ij}(\varphi,\psi)u(j)\} \\ &= \sup_{\varphi\in\overline{A}_n(i)} \inf_{\psi\in\overline{B}_n(i)} \{r_n(i,\varphi,\psi) + \sum_{j\in S_n} q^n_{ij}(\varphi,\psi)u(j)\}; \end{aligned}$$

see Theorem 1 in [12]. So, using a symmetric argument with the inequality

$$-h(i) \leq \alpha u(i) - \inf_{\psi\in\overline{B}_n(i)} \sup_{\varphi\in\overline{A}_n(i)} \{r_n(i,\varphi,\psi) + \sum_{j\in S_n} q^n_{ij}(\varphi,\psi)u(j)\}$$

gives the existence of $\pi^2 \in \Pi_n^{2,s}$ such that for all $\pi^1 \in \Pi_n^{1,s}$

$$V_n^\alpha(i,\pi^1,\pi^2) \leq u(i) + \frac{h(i)}{\alpha + c_h} + \frac{d_h}{\alpha(\alpha+c_h)} \quad \text{for all } i \in S_n,$$

and, therefore,

$$V_n^\alpha(i) \leq u(i) + \frac{h(i)}{\alpha + c_h} + \frac{d_h}{\alpha(\alpha+c_h)} \quad \text{for all } i \in S_n.$$

Together with (3.3.6), this proves the stated result. □

Finally, we state our main result on the convergence rates to the value of the game. In this theorem, the game models $\mathcal{G}_n$ are constructed, starting from $\mathcal{G}$, as described at the beginning of this section. It uses the notation $\rho_n(i)$ introduced in Definition 3.2.1.

Theorem 3.3.6 *Suppose that the game model $\mathcal{G}$ satisfies Assumptions 3.1.2, 3.1.4 and 3.3.3. Let $\{\mathcal{G}_n\}_{n\geq 1}$ be the sequence of finite state and actions truncations of $\mathcal{G}$, and suppose that the action sets for $\mathcal{G}_n$ are chosen in such a way that, for all $n \geq 1$ and $i \in S_n$, and for some constant $D > 0$*

$$\rho_n(i) \leq \frac{Dw^{\delta}(i)}{w^{\delta-2}(n+1)(L_i + 2\mathfrak{M}w(n)\sum_{j=0}^{n-1} L_{ij})}.$$

Under these conditions, there exists a constant $\mathfrak{c} > 0$ such that, for every $n \geq 1$ and $i \in S_n$,

$$|V_n^{\alpha}(i) - V^{\alpha}(i)| \leq \mathfrak{c}\,\frac{w^{\delta}(i)}{w^{\delta-2}(n+1)}.$$

Proof: Fix $n \geq 1$ and $i \in S_n$. We have

$$\begin{aligned} \alpha V^{\alpha}(i) &= \sup_{\varphi\in\overline{A}(i)} \inf_{\psi\in\overline{B}(i)} \{r(i,\varphi,\psi) + \sum_{j\in S} q_{ij}(\varphi,\psi)V^{\alpha}(j)\} \\ &\leq \sup_{\varphi\in\overline{A}(i)} \inf_{\psi\in\overline{B}_n(i)} \{r(i,\varphi,\psi) + \sum_{j\in S} q_{ij}(\varphi,\psi)V^{\alpha}(j)\}. \end{aligned} \tag{3.3.7}$$

Note that for every $(\varphi,\psi) \in \overline{A}(i) \times \overline{B}(i)$

$$\sum_{j\in S} q_{ij}(\varphi,\psi)V^{\alpha}(j) = \sum_{j=0}^{n-1} q_{ij}(\varphi,\psi)(V^{\alpha}(j) - V^{\alpha}(n)) + \sum_{j>n} q_{ij}(\varphi,\psi)(V^{\alpha}(j) - V^{\alpha}(n)),$$

and recalling that $||V^{\alpha}||_w \leq \mathfrak{M}$,

$$\Big|\sum_{j>n} q_{ij}(\varphi,\psi)(V^{\alpha}(j) - V^{\alpha}(n))\Big| \leq 2\mathfrak{M}\sum_{j>n} q_{ij}(\varphi,\psi)w(j).$$

Observe now that proceeding as in the proof of Lemma 3.1.8(i) and recalling (3.3.4), we can show that

$$\begin{aligned} \sum_{j>n} q_{ij}(\varphi,\psi)w(j) &\leq \frac{1}{w^{\delta-2}(n+1)} \cdot \Big((|c_{\delta}| + d_{\delta})w^{\delta-1}(i) + q(i)w^{\delta-1}(i)\Big) \\ &\leq (|c_{\delta}| + d_{\delta} + 1)\frac{w^{\delta}(i)}{w^{\delta-2}(n+1)}. \end{aligned} \tag{3.3.8}$$

Therefore, combining (3.3.7) and (3.3.8), we obtain

$$\alpha V^{\alpha}(i) \leq \sup_{\varphi\in\overline{A}(i)} \inf_{\psi\in\overline{B}_n(i)} \Big\{r(i,\varphi,\psi) + \sum_{j=0}^{n-1} q_{ij}(\varphi,\psi)(V^{\alpha}(j) - V^{\alpha}(n))\Big\} + \overline{C}\frac{w^{\delta}(i)}{w^{\delta-2}(n+1)},$$

with

$$\overline{C} = 2\mathfrak{M}(|c_{\delta}| + d_{\delta} + 1).$$

By upper semicontinuity, the above supremum is indeed attained. Consequently, there exists $\varphi \in \overline{A}(i)$ such that, for every $\psi \in \overline{B}_n(i)$, we have

$$\alpha V^\alpha(i) \leq r(i,\varphi,\psi) + \sum_{j=0}^{n-1} q_{ij}(\varphi,\psi)(V^\alpha(j) - V^\alpha(n)) + \overline{C}\frac{w^\delta(i)}{w^{\delta-2}(n+1)}. \tag{3.3.9}$$

Given arbitrary $\epsilon > 0$, there exist a finite set $\{x_1, \ldots, x_k\} \subseteq A(i)$ and $\beta_1, \ldots, \beta_k \geq 0$, with $\beta_1 + \ldots + \beta_k = 1$, such that

$$d_W(\varphi, \sum_{j=1}^{k} \beta_j \delta_{x_j}) \leq \epsilon.$$

For every x_j, let $\hat{x}_j \in A_n(i)$ be such that

$$d_A(x_j, \hat{x}_j) = \min_{y \in A_n(i)} d_A(x_j, y) \leq \rho_A(A_n(i), A(i)).$$

It is easy to see (recall (1.4.5)) that

$$d_W(\sum_{j=1}^{k} \beta_j \delta_{x_j}, \sum_{j=1}^{k} \beta_j \delta_{\hat{x}_j}) \leq \sum_{j=1}^{k} \beta_j d_A(x_j, \hat{x}_j) \leq \rho_A(A_n(i), A(i)),$$

and so, letting $\hat{\varphi} = \sum_{j=1}^{k} \beta_j \delta_{\hat{x}_j} \in \overline{A}_n(i)$,

$$d_W(\varphi, \hat{\varphi}) \leq \epsilon + \rho_A(A_n(i), A(i)).$$

Summarizing, for $\varphi \in \overline{A}(i)$ we have found a probability measure in $\hat{\varphi} \in \overline{A}_n(i)$ which is "close" to φ in the Wasserstein metric. By the Lipschitz continuity Assumption 3.3.3, observe that the function on $A(i) \times B(i)$ given by

$$(a,b) \mapsto r(i,a,b) + \sum_{j=0}^{n-1} q_{ij}(a,b)(V^\alpha(j) - V^\alpha(n))$$

is Lipschitz continuous, with Lipschitz constant $L_i + 2\mathfrak{M}w(n)\sum_{j=0}^{n-1} L_{ij}$. Consequently, the same applies to

$$a \mapsto \int_{B_n(i)} \Big[r(i,a,b) + \sum_{j=0}^{n-1} q_{ij}(a,b)(V^\alpha(j) - V^\alpha(n)) \Big] \psi(db).$$

Use now (1.4.5) to derive that

$$\Big| r(i,\varphi,\psi) + \sum_{j=0}^{n-1} q_{ij}(\varphi,\psi)(V^\alpha(j) - V^\alpha(n)) - r(i,\hat{\varphi},\psi) + \sum_{j=0}^{n-1} q_{ij}(\hat{\varphi},\psi)(V^\alpha(j) - V^\alpha(n)) \Big|$$

$$\leq \Big(L_i + 2\mathfrak{M}w(n) \sum_{j=0}^{n-1} L_{ij} \Big) \cdot d_W(\varphi,\hat{\varphi}) \leq \Big(L_i + 2\mathfrak{M}w(n) \sum_{j=0}^{n-1} L_{ij} \Big) \cdot (\epsilon + \rho_A(A_n(i), A(i))).$$

Therefore, recalling (3.3.9), this yields that $\alpha V^\alpha(i)$ is less than or equal to

$$r(i,\hat{\varphi},\psi)+\sum_{j=0}^{n-1} q_{ij}(\hat{\varphi},\psi)(V^\alpha(j)-V^\alpha(n))$$
$$+\overline{C}\frac{w^\delta(i)}{w^{\delta-2}(n+1)}+\Big(L_i+2\mathfrak{M}w(n)\sum_{j=0}^{n-1}L_{ij}\Big)\cdot(\epsilon+\rho_A(A_n(i),A(i))).$$

Since this holds for all $\psi\in\overline{B}_n(i)$ and the particular $\hat{\varphi}\in\overline{A}_n(i)$ constructed above, we deduce that

$$\begin{aligned}\alpha V^\alpha(i) &\le \sup_{\varphi\in\overline{A}_n(i)}\inf_{\psi\in\overline{B}_n(i)}\Big\{r(i,\varphi,\psi)+\sum_{j=0}^{n-1}q_{ij}(\varphi,\psi)(V^\alpha(j)-V^\alpha(n))\Big\}+\overline{C}\frac{w^\delta(i)}{w^{\delta-2}(n+1)}\\ &\quad+\Big(L_i+2\mathfrak{M}w(n)\sum_{j=0}^{n-1}L_{ij}\Big)\cdot(\epsilon+\rho_A(A_n(i),A(i))).\end{aligned}$$

But $\epsilon>0$ being arbitrary and recalling our hypothesis on $\rho_A(A_n(i),A(i))\le\rho_n(i)$, we derive that

$$\begin{aligned}\alpha V^\alpha(i) &\le \sup_{\varphi\in\overline{A}_n(i)}\inf_{\psi\in\overline{B}_n(i)}\{r(i,\varphi,\psi)+\sum_{j=0}^{n-1}q_{ij}(\varphi,\psi)(V^\alpha(j)-V^\alpha(n))\}+\frac{(\overline{C}+D)w^\delta(i)}{w^{\delta-2}(n+1)}.\\ &= \sup_{\varphi\in\overline{A}_n(i)}\inf_{\psi\in\overline{B}_n(i)}\{r_n(i,\varphi,\psi)+\sum_{j\in S_n}q^n_{ij}(\varphi,\psi)V^\alpha(j)\}+\frac{(\overline{C}+D)w^\delta(i)}{w^{\delta-2}(n+1)},\end{aligned}$$

where the last equality is derived from the definition of the reward and transition rates of the game model $\mathcal{G}_n$.

Using a symmetric argument, we can show that

$$\alpha V^\alpha(i)\ge\inf_{\psi\in\overline{B}_n(i)}\sup_{\varphi\in\overline{A}_n(i)}\{r_n(i,\varphi,\psi)+\sum_{j\in S_n}q^n_{ij}(\varphi,\psi)V^\alpha(j)\}-\frac{(\overline{C}+D)w^\delta(i)}{w^{\delta-2}(n+1)}.$$

As in the proof of Theorem 3.3.2 we can show that the inequality in Assumption 3.3.3(iii) is satisfied by the transition rates of $\mathcal{G}_n$ with the same constants c_δ and d_δ. Therefore, by Lemma 3.3.5, we conclude that, for every $i\in S_n$,

$$|V^\alpha_n(i)-V^\alpha(i)|\le\frac{(\overline{C}+D)w^\delta(i)}{(\alpha+c_\delta)w^{\delta-2}(n+1)}+\frac{(\overline{C}+D)d_\delta}{\alpha(\alpha+c_\delta)w^{\delta-2}(n+1)}.$$

Recalling the definition of the constants $\overline{C}$ and $\mathfrak{M}$, and letting

$$\mathfrak{c}=\frac{(2M(\alpha+d_1)(|c_\delta|+d_\delta+1)+D\alpha(\alpha+c_1))(d_\delta+\alpha)}{\alpha^2(\alpha+c_1)(\alpha+c_\delta)}$$

we have

$$|V_n^\alpha(i) - V^\alpha(i)| \leq \mathfrak{c}\, \frac{w^\delta(i)}{w^{\delta-2}(n+1)}$$

for all $n \geq 1$ and $i \in S_n$. □

The above theorem shows that, if a Lyapunov condition holds for the function w^δ (with $\delta > 2$) then, by making a suitable choice of the finite action sets $A_n(i)$ and $B_n(i)$, the error when approximating $V^\alpha(i)$ with $V_n^\alpha(i)$ is of order $1/w^{\delta-2}(n+1)$. Moreover, we have an explicit expression for the multiplicative constant $\mathfrak{c}$ that depends on the initial data (and related constants) of the game model $\mathcal{G}$.

3.3.3 Solving numerically a finite discounted game

Our previous results show that the value function V^α of the original game model $\mathcal{G}$ can be approximated by the value V_n^α of the finite state and actions game models $\mathcal{G}_n$. But, from the numerical perspective, it remains to explain how a finite state and actions game model can be solved explicitly.

Consider the finite state and actions game $\mathcal{G}_n$ defined at the beginning of Section 3.3.2. Let $\mathbf{q}_n > 0$ be such that

$$\mathbf{q}_n > -q_{ii}^n(a,b) \quad \text{for all } (i,a,b) \in \mathbb{K}_n \tag{3.3.10}$$

(it suffices to let $\mathbf{q}_n > w(n)$). For $u = \{u(i)\}_{i \in S_n} \in \mathbb{R}^{n+1}$ define the operator $T_n u \in \mathbb{R}^{n+1}$ as

$$\begin{aligned} T_n u(i) &= \max_{\varphi \in \overline{A}_n(i)} \min_{\psi \in \overline{B}_n(i)} \Big\{ r_n(i,\varphi,\psi) + \sum_{j \in S_n} q_{ij}^n(\varphi,\psi) u(j) \Big\} \\ &= \min_{\psi \in \overline{B}_n(i)} \max_{\varphi \in \overline{A}_n(i)} \Big\{ r_n(i,\varphi,\psi) + \sum_{j \in S_n} q_{ij}^n(\varphi,\psi) u(j) \Big\} \end{aligned}$$

for $i \in S_n$. Define also $\widetilde{T}_n u \in \mathbb{R}^{n+1}$ as

$$\widetilde{T}_n u(i) = \max_{\varphi \in \overline{A}_n(i)} \min_{\psi \in \overline{B}_n(i)} \Big\{ \frac{r_n(i,\varphi,\psi)}{\alpha + \mathbf{q}_n} + \frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n} \sum_{j \in S_n} \Big(\frac{q_{ij}^n(\varphi,\psi)}{\mathbf{q}_n} + \delta_{ij} \Big) u(j) \Big\} \tag{3.3.11}$$

$$= \min_{\psi \in \overline{B}_n(i)} \max_{\varphi \in \overline{A}_n(i)} \Big\{ \frac{r_n(i,\varphi,\psi)}{\alpha + \mathbf{q}_n} + \frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n} \sum_{j \in S_n} \Big(\frac{q_{ij}^n(\varphi,\psi)}{\mathbf{q}_n} + \delta_{ij} \Big) u(j) \Big\} \tag{3.3.12}$$

for $i \in S_n$ (cf. Section 8 in [16]). It is easily seen that the equation $\alpha u = T_n u$ is equivalent to the fixed point equation $u = \widetilde{T}_n u$. Therefore, as a consequence of Theorem 3.2.6, the value V_n^α of the game $\mathcal{G}_n$ is the unique fixed point of the operator $\widetilde{T}_n$. Moreover, by a standard calculation, it follows that $\widetilde{T}_n$ is a contraction operator on $\mathbb{R}^{n+1}$ with modulus $\mathbf{q}_n/(\alpha + \mathbf{q}_n) < 1$ when considering the supremum norm; that is,

$$\|\widetilde{T}_n u - \widetilde{T}_n v\| \leq \frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n} \|u - v\| \quad \text{for all } u, v \in \mathbb{R}^{n+1}.$$

Hence, the iterative procedure (which is a sort of value iteration algorithm):

1. Fix arbitrary $u_0 \in \mathbb{R}^{n+1}$,
2. For $k \geq 1$, let $u_k = \widetilde{T}_n u_{k-1}$,

converges geometrically to V_n^α in the supremum norm. Concerning the computation of the iterate $\widetilde{T}_n u$ for a given $u \in \mathbb{R}^{n+1}$, we can apply our next lemma, which uses the following notation. Given a positive integer N, define Δ_N as the set of nonnegative $\lambda_1, \ldots, \lambda_N$ such that $\lambda_1 + \ldots + \lambda_N = 1$.

Lemma 3.3.7 *Given the real-valued matrix $C = \{C_{s,t}\}_{1 \leq s \leq I, 1 \leq t \leq J}$, define*

$$V^* = \max_{\lambda \in \Delta_I} \min_{1 \leq t \leq J} \sum_{1 \leq s \leq I} \lambda_s C_{s,t} = \min_{\mu \in \Delta_J} \max_{1 \leq s \leq I} \sum_{1 \leq t \leq J} \mu_t C_{s,t}.$$

Let $\mathbf{c} \geq 0$ be such that all the elements of the matrix D, with $D_{s,t} = C_{s,t} + \mathbf{c}$, are strictly positive. Consider the linear programming problem

$$\min \ \mathbf{1}'\mathbf{x} \quad \textit{subject to} \quad D'\mathbf{x} \geq \mathbf{1}, \ \mathbf{x} \geq \mathbf{0},$$

and let $\mathbf{x}^* \in \mathbb{R}^I$ be an optimal solution. Then $V^* = \frac{1}{\mathbf{1}'\mathbf{x}^*} - \mathbf{c}$.

Proof: We have

$$V^* + \mathbf{c} = \max_{\lambda \in \Delta_I} \min_{1 \leq t \leq J} \sum_{1 \leq s \leq I} \lambda_s D_{s,t} =: \tilde{V}$$

and suppose that $D_{s,t} \geq \epsilon > 0$ for all s and t. Observe that $\tilde{V}$ equals the optimum of the linear programming problem: maximize v subject to

$$v \leq \sum_{1 \leq s \leq I} \lambda_s D_{s,t} \quad \text{for all } 1 \leq t \leq J,$$

with $v \geq \epsilon$ and $\lambda \in \Delta_I$. Letting $x_s = \lambda_s / v$ for $s = 1, \ldots, I$, it follows that $1/\tilde{V}$ equals the optimum of:

$$\min \ \mathbf{1}'\mathbf{x} \quad \text{subject to} \quad D'\mathbf{x} \geq \mathbf{1}, \ \mathbf{x} \geq \mathbf{0}, \ \mathbf{1}'\mathbf{x} \leq 1/\epsilon,$$

where the last constraint is redundant. □

Therefore, once u_{k-1} is known, we can effectively compute u_k by solving the linear programming problem described in Lemma 3.3.7. Namely, given $i \in S_n$ and for all $a_s \in A_n(i)$ with $1 \leq s \leq I$, and all $b_t \in B_n(i)$ with $1 \leq t \leq J$, define

$$C_{s,t} = \frac{r_n(i, a_s, b_t)}{\alpha + \mathbf{q}_n} + \frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n} \sum_{j \in S_n} \Big(\frac{q_{ij}^n(a_s, b_t)}{\mathbf{q}_n} + \delta_{ij} \Big) u_{k-1}(j)$$

and then use Lemma 3.3.7 to determine u_k.

Regarding a stopping criterion for the above algorithm, we have the following result. In the next lemma, the norm $||\cdot||$ refers to the supremum norm on $\mathbb{R}^{n+1}$.

Lemma 3.3.8 *Given the finite state and actions game model $\mathcal{G}_n$, consider the sequence of iterates $\{u_k\}_{k\geq 0}$, where $u_0 \in \mathbb{R}^{n+1}$ is arbitrary and, for $k \geq 1$, $u_k = \widetilde{T}_n u_{k-1}$. Fix $\epsilon > 0$ and let $k \geq 1$ be such that $||u_{k-1} - u_k|| \leq \epsilon\alpha/\mathbf{q}_n$. The following statements hold.*

(i) $||u_k - V_n^\alpha|| \leq \epsilon$.

(ii) The strategy $\pi_*^1 \in \Pi_n^{1,s}$ such that, for all $i \in S_n$, $\pi_*^1(\cdot|i)$ attains the maximum in (3.3.11) for the iteration $u_{k+1} = \widetilde{T}_n u_k$ is 2ϵ-optimal for player 1, meaning that

$$V_n^\alpha(i) - 2\epsilon \leq \inf_{\pi^2 \in \Pi_n^2} V_n^\alpha(i, \pi_*^1, \pi^2) \quad \textit{for all } i \in S_n.$$

(iii) The strategy $\pi_*^2 \in \Pi_n^{2,s}$ such that, for all $i \in S_n$, $\pi_*^2(\cdot|i)$ attains the minimum in (3.3.12) for the iteration $u_{k+1} = \widetilde{T}_n u_k$ is 2ϵ-optimal for player 2, meaning that

$$V_n^\alpha(i) + 2\epsilon \geq \sup_{\pi^1 \in \Pi_n^1} V_n^\alpha(i, \pi^1, \pi_*^2) \quad \textit{for all } i \in S_n.$$

Proof. (i). We have

$$||u_k - V_n^\alpha|| \leq ||u_k - u_{k+1}|| + ||u_{k+1} - V_n^\alpha|| \leq \frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n}\Big(||u_{k-1} - u_k|| + ||u_k - V_n^\alpha||\Big)$$

because V_n^α is the fixed point of $\widetilde{T}_n$, and so

$$||u_k - V_n^\alpha|| \leq \frac{\mathbf{q}_n}{\alpha}||u_{k-1} - u_k|| \leq \epsilon.$$

(ii). For $u : S_n \to \mathbb{R}$, consider the operator

$$\widetilde{U}u(i) = \min_{b \in \overline{B}_n(i)} \Big\{ \frac{r_n(i, \pi_*^1(\cdot|i), b)}{\alpha + \mathbf{q}_n} + \frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n} \sum_{j \in S_n} \Big(\frac{q_{ij}^n(\pi_*^1(\cdot|i), b)}{\mathbf{q}_n} + \delta_{ij} \Big) u(j) \Big\} \quad \text{for } i \in S_n,$$

which is a contraction on $\mathbb{R}^{n+1}$ with modulus $\frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n}$, and let W be its unique fixed point. The fixed point equation

$$W(i) = \min_{b \in \overline{B}_n(i)} \Big\{ \frac{r_n(i, \pi_*^1(\cdot|i), b)}{\alpha + \mathbf{q}_n} + \frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n} \sum_{j \in S_n} \Big(\frac{q_{ij}^n(\pi_*^1(\cdot|i), b)}{\mathbf{q}_n} + \delta_{ij} \Big) W(j) \Big\} \quad \text{for } i \in S_n,$$

corresponds to the discounted cost optimality equation of a continuous-time control problem (for player 2) when the strategy of player 1 is π_*^1; see [31, Section 3.3]. Therefore, $W(i) = \inf_{\pi^2 \in \Pi_n^2} V_n^\alpha(i, \pi_*^1, \pi^2)$ for all $i \in S_n$.

Observe now that

$$||W - V_n^\alpha|| \leq ||W - u_k|| + ||u_k - V_n^\alpha||. \tag{3.3.13}$$

Now, on the one hand,

$$\begin{aligned} ||W - u_k|| &\leq ||W - \widetilde{T}_n u_k|| + ||\widetilde{T}_n u_k - u_k|| \\ &= ||\widetilde{U}W - \widetilde{U}u_k|| + ||\widetilde{T}_n u_k - u_k|| \leq \frac{\mathbf{q}_n}{\alpha + \mathbf{q}_n}\Big(||W - u_k|| + ||u_{k-1} - u_k||\Big) \end{aligned}$$

because $\widetilde{T}_n u_k = \widetilde{U} u_k$, and so

$$||W - u_k|| \leq \frac{\mathbf{q}_n}{\alpha}||u_{k-1} - u_k||.$$

On the other hand, as established in part (i), $||u_k - V_n^\alpha|| \leq \frac{\mathbf{q}_n}{\alpha}||u_{k-1} - u_k||$. From (3.3.13), we obtain

$$||W - V_n^\alpha|| \leq \frac{2\mathbf{q}_n}{\alpha}||u_{k-1} - u_k|| \leq 2\epsilon,$$

and the result follows. The proof of (iii) is similar. □

As a consequence of this lemma, we can explicitly obtain an approximation of the value and nearly optimal strategies for both players for the game model $\mathcal{G}_n$.

3.4 Approximation results for average games

In this section we study the approximation problem for an average payoff Markov game; recall Sections 3.1.3 and 3.2.3.

3.4.1 Convergence results: the general case

In our next theorem we prove the main result on the convergence of the average value of the game models $\mathcal{G}_n$ to the average value g^* of $\mathcal{G}$. We also analyze the optimality of the limiting strategies. We note that, in addition to the conditions imposed so far on the game models $\mathcal{G}$ and $\mathcal{G}_n$ we need a supplementary hypothesis. Namely, we suppose that the functions $h_n \in \mathcal{B}_w(S_n)$ in the solution of the average optimality equations for $\mathcal{G}_n$, recall Theorem 3.2.8, can be chosen in such a way that $\sup_{n\geq 1} ||h_n||_w < \infty$. In connection with this, see Remark 3.4.2 below.

Theorem 3.4.1 *Suppose that the game model $\mathcal{G}$ satisfies Assumptions 3.1.11, 3.1.12, and 3.1.14, and that the game models $\mathcal{G}_n$ satisfy Assumption 3.2.7. In addition, assume that there exist $h_n \in \mathcal{B}_w(S_n)$, solutions to the average optimality equations of $\mathcal{G}_n$ in Theorem 3.2.8, such that $\sup_{n\geq 1} ||h_n||_w$ is finite. Under these conditions, if $\mathcal{G}_n \to \mathcal{G}$ then*

(i) The average value of $\mathcal{G}_n$ converges to the average value of $\mathcal{G}$, i.e., $\lim_{n\to\infty} g_n^* = g^*$.

(ii) If $(\pi_n^1, \pi_n^2) \in \Pi_s^{1,n} \times \Pi_s^{2,n}$ is a pair of average optimal strategies for $\mathcal{G}_n$ for every $n \geq 1$, then any limiting strategy $(\pi^1, \pi^2) \in \Pi_s^1 \times \Pi_s^2$ is average optimal for $\mathcal{G}$.

Proof. (i). Since the sequence of average values $\{g_n^*\}_{n\geq 1}$ is bounded (recall Theorem 3.2.8(i)), and by the hypothesis on the sequence $\{h_n\}_{n\geq 1}$, it follows that there exists some subsequence $\{n'\}$ along which $\{g_n^*\}$ and $\{h_n\}$ converge. Without loss of generality, we shall assume that the whole sequences converge. Hence, there exists a pair $(g,h)\in\mathbb{R}\times\mathcal{B}_w(S)$ such that

$$\lim_{n\to\infty} g_n^* = g \quad \text{and} \quad \lim_{n\to\infty} h_n(i) = h(i) \ \text{ for all } i\in S.$$

Fix $i\in S$ and consider $n\geq n(i)$. From the average optimality equations for the game model $\mathcal{G}_n$ in state $i\in S_n$ we obtain

$$g_n^* = \sup_{\varphi\in\overline{A}_n(i)}\inf_{\psi\in\overline{B}_n(i)}\Big\{r_n(i,\varphi,\psi)+\sum_{j\in S_n} q_{ij}^n(\varphi,\psi)h_n(j)\Big\}.$$

The function

$$(\varphi,\psi)\mapsto r_n(i,\varphi,\psi)+\sum_{j\in S_n} q_{ij}^n(\varphi,\psi)h_n(j)$$

being continuous (Lemma 3.2.4), it follows that

$$\varphi\mapsto \inf_{\psi\in\overline{B}_n(i)}\Big\{r_n(i,\varphi,\psi)+\sum_{j\in S_n} q_{ij}^n(\varphi,\psi)h_n(j)\Big\}$$

is upper semicontinuous, and so it reaches a maximum on the compact set $\overline{A}_n(i)$ at the point, say, $\varphi_n^*\in\overline{A}_n(i)$, that is,

$$g_n^* = \inf_{\psi\in\overline{B}_n(i)}\Big\{r_n(i,\varphi_n^*,\psi)+\sum_{j\in S_n} q_{ij}^n(\varphi_n^*,\psi)h_n(j)\Big\}. \tag{3.4.1}$$

The sequence $\{\varphi_n^*\}\subseteq\overline{A}(i)$ has some convergent subsequence. We shall assume that the whole sequence converges: that is, for some $\varphi\in\overline{A}(i)$ we have $\varphi_n^*\xrightarrow{d}\varphi$.

Fix now arbitrary $\psi\in\overline{B}(i)$ and $n\geq n(i)$. For every such n there exist some points $x_1,\ldots,x_t\in B(i)$ and positive $\beta_1,\ldots,\beta_t$ with $\sum\beta_k=1$, such that

$$d_W\big(\psi,\sum_{k=1}^t\beta_k\delta_{x_k}\big)<1/n.$$

We will write $\hat{\psi}_n=\sum_{k=1}^t\beta_k\delta_{x_k}$. For each x_k let $y_k\in B_n(i)$ be such that

$$d_B(y_k,x_k)=\min_{y\in B_n(i)} d_B(y,x_k)\leq\rho_B(B_n(i),B(i)).$$

Consider the probability measure $\tilde{\psi}_n=\sum_{k=1}^t\beta_k\delta_{y_k}$. A straightforward calculation yields that $d_W(\hat{\psi}_n,\tilde{\psi}_n)\leq\rho_B(B_n(i),B(i))$. Consequently, we have that $\tilde{\psi}_n\in\overline{B}_n(i)$ verifies $\tilde{\psi}_n\xrightarrow{d}\psi$ because

$$d_W(\hat{\psi}_n,\psi)\leq\rho_n(i)+1/n\to 0.$$

Now, from (3.4.1), for every $n \geq n(i)$ we have

$$g_n^* \leq r_n(i, \varphi_n^*, \hat{\psi}_n) + \sum_{j \in S_n} q_{ij}^n(\varphi_n^*, \hat{\psi}_n) h_n(j).$$

Take the limit as $n \to \infty$ and use Lemma 3.2.5 to conclude that

$$g \leq r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) h(j).$$

Since $\psi \in \overline{B}(i)$ is arbitrary, we obtain that

$$g \leq \inf_{\psi \in \overline{B}(i)} \big\{ r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) h(j) \big\},$$

and so

$$g \leq \sup_{\varphi \in \overline{A}(i)} \inf_{\psi \in \overline{B}(i)} \big\{ r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) h(j) \big\}.$$

Proceeding with a symmetric argument we can establish that

$$g \geq \inf_{\psi \in \overline{B}(i)} \sup_{\varphi \in \overline{A}(i)} \big\{ r(i, \varphi, \psi) + \sum_{j \in S} q_{ij}(\varphi, \psi) h(j) \big\}.$$

Since $i \in S$ is arbitrary, we conclude that the pair $(g, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ is a solution to the average optimality equations for the game model $\mathcal{G}$. Hence, we must have $g = g^*$.

Summarizing, we have proved that any convergent sequence of the bounded sequence $\{g_n^*\}_{n \geq 1}$ converges to g^*. Necessarily, we must have $\lim g_n^* = g^*$.
(ii). Suppose that $(\pi_n^1, \pi_n^2) \in \Pi_s^{1,n} \times \Pi_s^{2,n}$ are optimal stationary strategies for the players, and suppose that (π^1, π^2) is a limiting strategy through the subsequence $\{n'\}$. As in the proof of statement (i) of this theorem, we can show that

$$g^* \leq \inf_{\psi \in \overline{B}(i)} \big\{ r(i, \pi^1(\cdot|i), \psi) + \sum_{j \in S} q_{ij}(\pi^1(\cdot|i), \psi) h(j) \big\} \quad \text{for all } i \in S,$$

from which optimality of strategy π^1 for player 1 follows. We proceed similarly for player 2. The proof is now complete. □

Remark 3.4.2 *This convergence theorem has been proved under the additional hypothesis that* $\sup_n ||h_n||_w < \infty$. *We note that if each game model* $\mathcal{G}_n$ *satisfies the conditions given in Remark 3.1.15 then the condition* $\sup_n ||h_n||_w < \infty$ *is fulfilled. See also Theorem 3.1.16(ii).*

Remark 3.4.3 *It is also worth noting that, in Theorem 3.4.1, we have not used Assumption 3.1.14(iv) on the irreducibility of stationary strategies for* $\mathcal{G}$. *In fact, if we drop Assumption 3.1.14(iv) then we cannot apply directly Theorem 3.1.16 to the game model* $\mathcal{G}$. *Interestingly, Theorem 3.4.1 is in fact a proof of the existence to solutions of the average optimality equations for* $\mathcal{G}$ *(and, hence, the existence of a constant average value function), based on suitable properties of a family of (simpler) game models* $\mathcal{G}_n$.

3.4.2 Convergence results: finite approximations

Starting from the countable state space, general action spaces game model $\mathcal{G}$, we consider the sequence of finite state and actions game models $\mathcal{G}_n$ that were defined in Section 3.3.2. For ease of reference, we give again the definition of the game model $\mathcal{G}_n$. For each $n \geq 1$ we define the game model $\mathcal{G}_n$ with elements $\{S_n, A, B, \mathbb{K}_n, Q_n, r\}$:

- The state space of $\mathcal{G}_n$ is the finite set $S_n = \{0, 1, \ldots, n\}$.
- The action spaces are A and B, as for the game model $\mathcal{G}$.
- The available actions for players 1 and 2 are arbitrary finite sets $A_n(i) \subseteq A(i)$ and $B_n(i) \subseteq B(i)$, respectively. For every $i \in S_n$, they verify

$$\rho_n(i) = \rho_A(A_n(i), A(i)) \vee \rho_B(B_n(i), B(i)) \to 0$$

  as $n \to \infty$. Let $\mathbb{K}_n = \{(i, a, b) \in S_n \times A \times B : a \in A_n(i), b \in B_n(i)\}$.
- Given $(i, a, b) \in \mathbb{K}_n$ and $j \in S_n$, the transition rates Q_n are

$$q_{ij}^n(a, b) = \begin{cases} q_{ij}(a, b) & \text{when } j \neq n \\ \sum_{k \geq n} q_{ik}(a, b) & \text{when } j = n. \end{cases}$$

  These transition rates are conservative and stable, with $-q_{ii}^n(a, b) \leq -q_{ii}(a, b) \leq q(i)$ for $(i, a, b) \in \mathbb{K}_n$.
- The payoff rate function is the restriction of r to $\mathbb{K}_n$, that we will also denote by r.

The dynamics of the game model $\mathcal{G}_n$ is roughly as follows: the game model $\mathcal{G}_n$ evolves according to the same dynamics as the game $\mathcal{G}$ and, whenever the system reaches a state strictly larger than n, it is restarted at state n.

Lemma 3.4.4 *If the game model $\mathcal{G}$ satisfies Assumptions 3.1.11, 3.1.12, and 3.1.14(i)–(iii) then the sequence $\{\mathcal{G}_n\}_{n \geq 1}$ satisfies Assumptions 3.2.7(i)–(iv) and, besides, $\mathcal{G}_n \to \mathcal{G}$.*

Proof. The proof is easy and similar to that of Theorem 3.3.2. □

It is important to mention that the irreducibility of stationary strategies for $\mathcal{G}_n$ cannot be deduced from the irreducibility of stationary policies for $\mathcal{G}$. That is why we have not included Assumption 3.1.14(iv) in the hypotheses of Lemma 3.4.4. Therefore, Assumption 3.2.7(v) needs not hold, we cannot use Theorem 3.2.8, and $\mathcal{G}_n$ might not have a value.

Moreover, we can neither use Theorem 3.4.1 on the convergence of the value functions of $\mathcal{G}_n$ to that of $\mathcal{G}$. So, in the context of a Markov game with the average payoff optimality criterion, the finite state and actions truncated game models $\mathcal{G}_n$ might not be used as approximations of $\mathcal{G}$. This is an important departure point from the results in the discounted payoff setting in which, under mild hypotheses, the truncated game models $\mathcal{G}_n$

could be used to approximate the discounted game $\mathcal{G}$. So, we will need to impose additional assumptions to obtain convergence.

The next result will be useful in the forthcoming. We note that this lemma does not suppose Assumption 3.1.14(iv) on the irreducibility of stationary strategies for $\mathcal{G}$. Hence, the game model $\mathcal{G}$ might not have a value function.

Lemma 3.4.5 *Let $\mathcal{G}$ satisfy Assumptions 3.1.11, 3.1.12, and 3.1.14(i)–(iii). Suppose that there exist $g \in \mathbb{R}$, $h \in \mathcal{B}_w(S)$, and $u : S \to [0,\infty)$ such that, for all $i \in S$,*

$$g - u(i) \leq \sup_{\varphi \in \overline{A}(i)} \inf_{\psi \in \overline{B}(i)} \{r(i,\varphi,\psi) + \sum_{j \in S} q_{ij}(\varphi,\psi)h(j)\} \tag{3.4.2}$$

$$\leq \inf_{\psi \in \overline{B}(i)} \sup_{\varphi \in \overline{A}(i)} \{r(i,\varphi,\psi) + \sum_{j \in S} q_{ij}(\varphi,\psi)h(j)\} \leq g + u(i), \tag{3.4.3}$$

and such that, for some constants $c_u > 0$ and $d_u \geq 0$, the function u satisfies

$$\sum_{j \in S} q_{ij}(a,b)u(j) \leq -c_u u(i) + d_u \quad \textit{for all } (i,a,b) \in \mathbb{K}.$$

Then $g - d_u/c_u \leq L(i) \leq U(i) \leq g + d_u/c_u$ for all $i \in S$.

Proof. By Lemma 3.1.8 and Assumptions 3.1.14(i)–(iii), given $i \in S$, the function $\varphi \mapsto r(i,\varphi,\psi) + \sum_{j \in S} q_{ij}(\varphi,\psi)h(j)$ is continuous on $\overline{A}(i)$ for all $\psi \in \overline{B}(i)$. Hence,

$$\varphi \mapsto \inf_{\psi \in \overline{B}(i)} \{r(i,\varphi,\psi) + \sum_{j \in S} q_{ij}(\varphi,\psi)h(j)\}$$

is upper semicontinuous and it reaches a maximum on the compact set $\overline{A}(i)$. As a consequence of (3.4.2), there exists some $\varphi_i^* \in \overline{A}(i)$ such that for all $\psi \in \overline{B}(i)$

$$g - u(i) \leq r(i,\varphi_i^*,\psi) + \sum_{j \in S} q_{ij}(\varphi_i^*,\psi)h(j).$$

Consider the stationary strategy $\pi_*^1 \in \Pi^{1,s}$ such that $\pi_*^1(\cdot|i) = \varphi_i^*$ for all $i \in S$. Given an arbitrary Markov strategy $\pi^2 \in \Pi^2$,

$$g - u(i) \leq r(t,i,\pi_*^1,\pi^2) + \sum_{j \in S} q_{ij}(t,\pi_*^1,\pi^2)h(j)$$

for all $i \in S$ and $t \geq 0$. Using Kolmogorov's backward equations for a nonstationary Markov chain [17, Proposition C.4],

$$\begin{aligned} &E^{i,\pi_*^1,\pi^2}[h(x(t))] - h(i) \\ &\geq -\int_0^t E^{i,\pi_*^1,\pi^2}\big[r(s,x(s),\pi_*^1,\pi^2) + u(x(s))\big]ds + gt. \end{aligned} \tag{3.4.4}$$

By arguments similar to those used to derive (3.1.10) we have

$$E^{i,\pi^1_*,\pi^2}[u(x(s))] \le e^{-c_u s}u(i) + \frac{d_u}{c_u}(1-e^{-c_u s}) \quad \text{for all } s \ge 0,$$

and so dividing by t in (3.4.4) and taking the lim sup as $t \to \infty$ yields $g \le J(i,\pi^1_*,\pi^2) + \frac{d_u}{c_u}$. Consequently, since $\pi^2 \in \Pi^2$ is arbitrary

$$g - \frac{d_u}{c_u} \le \inf_{\pi^2\in\Pi^2} J(i,\pi^1_*,\pi^2) \le \sup_{\pi^1\in\Pi^1}\inf_{\pi^2\in\Pi^2} J(i,\pi^1,\pi^2) = L(i).$$

Proceeding similarly with (3.4.3) we obtain $U(i) \le g + \frac{d_u}{c_u}$ for each $i \in S$. The stated result follows. □

As already mentioned, we need to strengthen our hypotheses to obtain convergence. Namely, we will replace Assumption 3.1.14 with the following stronger condition.

Assumption 3.4.6 *(i) With w the Lyapunov function in Assumption 3.1.11, there exist $\delta > 2$, $c_\delta > 0$, and $d_\delta \ge 0$ such that*

$$\sum_{j\in S} q_{ij}(a,b)w^\delta(j) \le -c_\delta w^\delta(i) + d_\delta \quad \text{for all } (i,a,b) \in \mathbb{K}.$$

(ii) The action sets $A(i)$ and $B(i)$ are compact for every $i \in S$. For all $i,j \in S$ there are positive constants L_i and L_{ij} with

$$\begin{aligned} |r(i,a,b) - r(i,a',b')| &\le L_i\big(d_A(a,a') + d_B(b,b')\big) \\ |q_{ij}(a,b) - q_{ij}(a',b')| &\le L_{ij}\big(d_A(a,a') + d_B(b,b')\big) \end{aligned}$$

for all $a,a' \in A(i)$ and $b,b' \in B(i)$; i.e., the functions $r(i,\cdot,\cdot)$ and $q_{ij}(\cdot,\cdot)$ are Lipschitz continuous.

(iii) Each pair of strategies in $\Pi^{1,s} \times \Pi^{2,s}$ is irreducible.

In this assumption, we impose a Lyapunov condition on w^δ for some $\delta > 2$, in which the coefficient c_δ is positive (cf. Assumption 3.3.3(iii)). Also, we impose Lipschitz continuity of the reward and transition rates. Note that Assumption 3.4.6(i)–(ii) is analogous to Assumption 3.3.3 expect for the additional condition on c_δ. Note that Assumption 3.4.6(i) indeed implies Assumption 3.1.14(iii), as consequence of Lemma 3.3.4.

Next we state our main result on the convergence of the value functions.

Theorem 3.4.7 *Suppose that the game model $\mathcal{G}$ satisfies Assumptions 3.1.11, 3.1.12, and 3.4.6. Fix $D > 0$ and suppose that the action sets $A_n(i)$ and $B_n(i)$ of $\mathcal{G}_n$, for $n \ge 1$, are chosen so that*

$$\rho_n(i) \le \frac{Dw^\delta(i)}{w^{\delta-2}(n+1)\Big(L_i + w(n)\frac{2RM}{\gamma}\sum_{j=0}^{n-1} L_{ij}\Big)} \quad \text{for } i \in S_n. \tag{3.4.5}$$

Then there exists a constant $\mathfrak{c} > 0$ *that depends neither on* $n \geq 1$ *nor on* $i \in S_n$ *such that, for all* $i \in S_n$ *and* $n \geq 1$,

$$g^* - \frac{\mathfrak{c}}{w^{\delta-2}(n+1)} \leq L_n(i) \leq U_n(i) \leq g^* + \frac{\mathfrak{c}}{w^{\delta-2}(n+1)}.$$

Consequently,

$$\max_{i \in S_n} |L_n(i) - g^*| \quad \textit{and} \quad \max_{i \in S_n} |U_n(i) - g^*|$$

are both $\mathrm{O}(w^{-(\delta-2)}(n+1))$ *as* $n \to \infty$.

Proof. Fix $n \geq 1$ and $i \in S_n$. Let $(g^*, h) \in \mathbb{R} \times \mathcal{B}_w(S)$ be a solution of the average optimality equations for the game model $\mathcal{G}$ such that $||h||_w \leq RM/\gamma$. From the average optimality equation in Theorem 3.1.16 we deduce that there exists $\varphi^* \in \overline{A}(i)$ such that, for all $\psi \in \overline{B}(i)$,

$$g^* \leq r(i, \varphi^*, \psi) + \sum_{j \in S} q_{ij}(\varphi^*, \psi) h(j). \tag{3.4.6}$$

Observe that $\sum_{j \in S} q_{ij}(\varphi^*, \psi) h(j)$ equals

$$\begin{aligned}
&\sum_{j=0}^{n-1} q_{ij}(\varphi^*, \psi)(h(j) - h(n)) + \sum_{j>n} q_{ij}(\varphi^*, \psi)(h(j) - h(n)) \\
&\leq \sum_{j=0}^{n-1} q_{ij}(\varphi^*, \psi)(h(j) - h(n)) + 2||h||_w \sum_{j>n} q_{ij}(\varphi^*, \psi) w(j),
\end{aligned}$$

where, proceeding as in Lemma 3.1.8 we obtain

$$\begin{aligned}
\sum_{j>n} q_{ij}(\varphi^*, \psi) w(j) &\leq \frac{1}{w^{\delta-2}(n+1)} \sum_{j>n} q_{ij}(\varphi^*, \psi) w^{\delta-1}(j) \\
&\leq \frac{1}{w^{\delta-2}(n+1)} (1 + d_\delta) w^\delta(i).
\end{aligned}$$

On the other hand, the probability measures with finite support being dense in $\overline{A}(i)$ (see Theorem 8.9.4 in [7]), given $\epsilon > 0$ there exist $x_1, \ldots, x_k \in A(i)$ and positive $\beta_1, \ldots, \beta_k$ with $\sum \beta_t = 1$ such that $d_W(\varphi^*, \sum \beta_t \delta_{x_t}) < \epsilon$. For each x_t let $y_t \in A_n(i)$ be such that

$$d_A(x_t, y_t) = \min_{a \in A_n(i)} d_A(x_t, a) \leq \rho_A(A_n(i), A(i))$$

and define $\widetilde{\varphi}_n = \sum \beta_t \delta_{y_t} \in \overline{A}_n(i)$. It is easily seen that

$$d_W(\sum \beta_t \delta_{x_t}, \widetilde{\varphi}_n) \leq \rho_A(A_n(i), A(i)) \leq \rho_n(i)$$

and thus $d_W(\varphi^*, \widetilde{\varphi}_n) < \epsilon + \rho_n(i)$. By Lipschitz continuity in Assumption 3.4.6(ii), it follows that $\sum_{j\in S} q_{ij}(\varphi^*, \psi)h(j)$ is less than or equal to

$$\sum_{j=0}^{n-1} q_{ij}(\widetilde{\varphi}_n, \psi)(h(j) - h(n)) + 2||h||_w w(n) d_W(\varphi^*, \widetilde{\varphi}_n) \sum_{j=0}^{n-1} L_{ij}$$
$$+ 2||h||_w (1 + d_\delta) \frac{w^\delta(i)}{w^{\delta-2}(n+1)}.$$

On the other hand, we have

$$r(i, \varphi^*, \psi) \le r(i, \widetilde{\varphi}_n, \psi) + L_i d_W(\varphi^*, \widetilde{\varphi}_n).$$

Summarizing, for $\widetilde{\varphi}_n \in \overline{A}_n(i)$ constructed above and for all $\psi \in \overline{B}_n(i) \subseteq \overline{B}(i)$ we have (see (3.4.6)) that g^* is less than or equal to

$$r(i, \widetilde{\varphi}_n, \psi) + \sum_{j=0}^{n-1} q_{ij}(\widetilde{\varphi}_n, \psi)(h(j) - h(n)) + d_W(\varphi^*, \widetilde{\varphi}_n)\Big(L_i +$$
$$+ 2||h||_w w(n) \sum_{j=0}^{n-1} L_{ij}\Big) + 2||h||_w (1 + d_\delta) \frac{w^\delta(i)}{w^{\delta-2}(n+1)}.$$

Recalling the bound on $d_W(\varphi^*, \widetilde{\varphi}_n)$ and the definition of the transition rates of $\mathcal{G}_n$, it follows that

$$g^* \le r(i, \widetilde{\varphi}_n, \psi) + \sum_{j\in S_n} q_{ij}^n(\widetilde{\varphi}_n, \psi)h(j) + (\epsilon + \rho_n(i)) \cdot$$
$$\cdot \Big(L_i + 2||h||_w w(n) \sum_{j=0}^{n-1} L_{ij}\Big) + 2||h||_w (1 + d_\delta) \frac{w^\delta(i)}{w^{\delta-2}(n+1)}.$$

Since $\psi \in \overline{B}_n(i)$ is arbitrary, recalling the bounds on $||h||_w$ and $\rho_n(i)$, and since $\epsilon > 0$ is arbitrary as well, it follows that

$$g^* \le \sup_{\varphi\in\overline{A}_n(i)} \inf_{\psi\in\overline{B}_n(i)} \big\{r(i, \varphi, \psi) + \sum_{j\in S_n} q_{ij}^n(\varphi, \psi)h(j)\big\} + \frac{\mathfrak{C} w^\delta(i)}{w^{\delta-2}(n+1)}$$

for all $i \in S_n$, where $\mathfrak{C} = D + 2RM(1 + d_\delta)/\gamma$. The symmetric inequality, that is,

$$g^* \ge \inf_{\psi\in\overline{B}_n(i)} \sup_{\varphi\in\overline{A}_n(i)} \big\{r(i, \varphi, \psi) + \sum_{j\in S_n} q_{ij}^n(\varphi, \psi)h(j)\big\} - \frac{\mathfrak{C} w^\delta(i)}{w^{\delta-2}(n+1)}$$

for $i \in S_n$, is proved similarly. Now we use Lemma 3.4.5 for the game model $\mathcal{G}_n$ to derive the stated result for $\mathfrak{c} = \mathfrak{C}/c_\delta$. □

Therefore, by making a suitable choice of the sets $A_n(i)$ and $B_n(i)$, we obtain convergence, uniformly in $i \in S_n$, of the lower and upper value functions L_n and U_n to the value of $\mathcal{G}$ at a rate $1/w^{\delta-2}(n+1)$. Note that, for fixed $n \geq 1$, the bound on $\rho_n(i)$ in (3.4.5) grows (loosely) with $w^\delta(i)$ for $i \in S_n$. This means that "fine" choices of the action sets of $\mathcal{G}_n$ must be made for "small" states, but "coarser" actions sets are allowed for "large" states.

It is also worth mentioning that Theorem 3.4.7 above does not assume the existence of the value of the games $\mathcal{G}_n$, and yet it provides a result on the convergence of the lower and upper value functions. If the game $\mathcal{G}_n$ has a value $g_n^* \in \mathbb{R}$ then the result of Theorem 3.4.7 becomes, obviously,

$$|g_n^* - g^*| \leq \frac{\mathfrak{c}}{w^{\delta-2}(n+1)} \quad \text{for all } n \geq 1.$$

3.4.3 Solving numerically a finite average game

It remains to show how to solve numerically the game model $\mathcal{G}_n$. Under an additional irreducibility assumption, we can use the following "policy iteration" procedure.

Theorem 3.4.8 *Suppose that the finite state and actions game models $\{\mathcal{G}_n\}_{n\geq 1}$ satisfy Assumption 3.2.7. For each fixed $n \geq 1$ consider the following iterative procedure.*

Step 0. *Choose arbitrary $\pi_0^1 \in \Pi_n^{1,s}$. Set $k = 0$ and go to Step 1.*

Step 1. *Find solutions $g_k \in \mathbb{R}$ and $h_k : S_n \to \mathbb{R}$ of the average cost optimality equation*

$$g_k = \min_{b \in B_n(i)} \Big\{ r_n(i, \pi_k^1, b) + \sum_{j \in S_n} q_{ij}^n(\pi_k^1, b) h_k(j) \Big\} \quad \text{for } i \in S_n.$$

Step 2. *For each $i \in S_n$, find $\varphi_i \in \overline{A}_n(i)$ attaining the maximum*

$$\max_{\varphi \in \overline{A}_n(i)} \min_{b \in B_n(i)} \Big\{ r_n(i, \varphi, b) + \sum_{j \in S_n} q_{ij}^n(\varphi, b) h_k(j) \Big\}.$$

Define $\pi_{k+1}^1 \in \Pi_n^{1,s}$ by means of $\pi_{k+1}^1(\cdot|i) = \varphi_i$ for $i \in S_n$. Increase k by one and go to Step 1.

The sequence $\{g_k\}_{k\geq 0}$ is monotone nondecreasing and it converges to the value g_n^* of the game $\mathcal{G}_n$.

Remark 3.4.9 *In Step 1, we solve the average cost optimality equation of a control problem for player 2 when the strategy of player 1 is π_k^1. It can be solved, in a finite number of steps, with the usual policy iteration algorithm (see chapters 3 and 4 in [31]). But also it is well known that solving this optimality equation reduces to a linear programming problem; see [35, Section 8.8] for the discrete-time analogue. Regarding Step 2, this maximin problem is equivalent to a linear programming problem (recall Lemma 3.3.7 and see also [22, Section 7.11]) and, therefore, it can be solved explicitly.*

Hence, the "policy iteration" algorithm in Theorem 3.4.8 reduces to solving, iteratively, linear programming problems. This makes the algorithm computationally tractable.

Proof. Fix $k \geq 0$ and observe that, by construction of π^1_{k+1},

$$\begin{aligned} &\min_{b \in B_n(i)} \left\{ r_n(i, \pi^1_{k+1}, b) + \sum_{j \in S_n} q^n_{ij}(\pi^1_{k+1}, b) h_k(j) \right\} \\ &\geq \min_{b \in B_n(i)} \left\{ r_n(i, \pi^1_k, b) + \sum_{j \in S_n} q^n_{ij}(\pi^1_k, b) h_k(j) \right\} = g_k \end{aligned} \tag{3.4.7}$$

for all $i \in S_n$. By a standard dynamic programming argument (see, e.g., [31, Lemma 3.10]) it follows that

$$g_{k+1} = \inf_{\pi^2 \in \Pi^2_n} J_n(i, \pi^1_{k+1}, \pi^2) \geq g_k,$$

since g_{k+1} is the optimal average reward of player 2 when the strategy of player 1 is fixed and equal to π^1_{k+1}. On the other hand, it also follows that

$$g^*_n = \sup_{\pi^1 \in \Pi^1_n} \inf_{\pi^2 \in \Pi^2_n} J_n(i, \pi^1, \pi^2) \geq g_k.$$

Consequently, the sequence $\{g_k\}_{k \geq 0}$ is monotone nondecreasing and its limit, denoted by $\overline{g}$, satisfies $\overline{g} \leq g^*_n$.

Let $\pi^2_k \in \Pi^{2,s}_n$ be a (nonrandomized) strategy attaining the minimum in the definition of g_k. We can choose h_k as the bias of (π^1_k, π^2_k) and, by uniform exponential ergodicity of $\mathcal{G}_n$ and the results in [31, Section 3.4], we can choose h_k such that $\sup_{k \geq 0} ||h_k||_w$ is finite. There exists a subsequence $\{k'\}$ such that the sequences $h_{k'}$ and $h_{k'+1}$ converge to some functions $h^{(1)}$ and $h^{(2)}$ on S_n, respectively, and such that, in addition,

$$\pi^1_{k'+1}(\cdot|i) \xrightarrow{d} \pi^1(\cdot|i) \quad \text{and} \quad \pi^2_{k'+1}(\cdot|i) \xrightarrow{d} \pi^2(\cdot|i) \quad \text{for all } i \in S_n,$$

for some $(\pi^1, \pi^2) \in \Pi^{1,s}_n \times \Pi^{2,s}_n$, as $k' \to \infty$. We have that

$$g_{k'+1} = r_n(i, \pi^1_{k'+1}, \pi^2_{k'+1}) + \sum_{j \in S_n} q^n_{ij}(\pi^1_{k'+1}, \pi^2_{k'+1}) h_{k'+1}(j)$$

and also, by (3.4.7), that

$$g_{k'} \leq r_n(i, \pi^1_{k'+1}, \pi^2_{k'+1}) + \sum_{j \in S_n} q^n_{ij}(\pi^1_{k'+1}, \pi^2_{k'+1}) h_{k'}(j)$$

for all $i \in S_n$. Therefore, for all $i \in S_n$,

$$\sum_{j \in S_n} q^n_{ij}(\pi^1_{k'+1}, \pi^2_{k'+1})(h_{k'+1}(j) - h_{k'}(j)) \leq g_{k'+1} - g_{k'}.$$

Taking the limit as $k' \to \infty$, it follows that

$$\sum_{j \in S_n} q^n_{ij}(\pi^1, \pi^2)(h^{(2)}(j) - h^{(1)}(j)) \leq 0 \quad \text{for all } i \in S_n.$$

The function $h^{(2)} - h^{(1)}$ being subharmonic for the irreducible strategy (π^1, π^2), it is constant; see [31, Proposition 2.6]. Now,

$$\begin{aligned}
&\max_{\varphi \in \overline{A}_n(i)} \min_{b \in B_n(i)} \Big\{ r_n(i, \varphi, b) + \sum_{j \in S_n} q_{ij}^n(\varphi, b) h_{k'}(j) \Big\} \qquad (3.4.8) \\
&= \min_{b \in B_n(i)} \Big\{ r_n(i, \pi^1_{k'+1}, b) + \sum_{j \in S_n} q_{ij}^n(\pi^1_{k'+1}, b) h_{k'}(j) \Big\} \\
&\leq r_n(i, \pi^1_{k'+1}, \pi^2_{k'+1}) + \sum_{j \in S_n} q_{ij}^n(\pi^1_{k'+1}, \pi^2_{k'+1}) h_{k'}(j) \\
&= g_{k+1} + \sum_{j \in S_n} q_{ij}^n(\pi^1_{k'+1}, \pi^2_{k'+1})(h_{k'} - h_{k'+1})(j) \quad \text{for all } i \in S_n.
\end{aligned}$$

Taking the limit as $k' \to \infty$ and recalling that $h_{k'} - h_{k'+1}$ converges to a constant function, we obtain

$$\begin{aligned}
&\max_{\varphi \in \overline{A}_n(i)} \min_{b \in B_n(i)} \Big\{ r_n(i, \varphi, b) + \sum_{j \in S_n} q_{ij}^n(\varphi, b) h^{(1)}(j) \Big\} \\
&= \min_{\psi \in \overline{B}_n(i)} \max_{a \in A_n(i)} \Big\{ r_n(i, a, \psi) + \sum_{j \in S_n} q_{ij}^n(a, \psi) h^{(1)}(j) \Big\} \leq \overline{g},
\end{aligned}$$

(interchange of limit and max-min in (3.4.8) follows because the action sets $A_n(i)$ and $B_n(i)$ are finite; see [12, Theorem 1]) from which (cf. proof of Lemma 3.4.5)

$$\inf_{\pi^2 \in \Pi_n^2} \sup_{\pi^1 \in \Pi_n^1} J_n(i, \pi^1, \pi^2) = g_n^* \leq \overline{g}$$

follows. This completes the proof of the result. □

It should be clear that if $\pi^1_{k+1} = \pi^1_k$ for some $k \geq 0$ then (g_k, h_k) is a solution of the average optimality equations for $\mathcal{G}_n$. Similarly, it can be derived from [11, Theorem 1] that if $g_{k+1} = g_k$ for some $k \geq 0$ then (g_k, h_k) is as well a solution of the average optimality equations for $\mathcal{G}_n$. Therefore, in either case g_k equals the value g_n^* of the game $\mathcal{G}_n$ and π^1_k is an optimal strategy for player 1. Finite convergence of the algorithm might not occur in general, however, because even if $A_n(i)$ is finite, $\overline{A}_n(i)$ is not. On the contrary, for finite state and action Markov decision processes, the policy iteration algorithm converges in a finite number of steps. The policy iteration algorithm has been shown to converge quadratically for certain classes of Markov decision processes [37], but no such convergence rates have been obtained for the policy iteration algorithm for Markov games.

3.5 An application

In this section we show an application of our numerical procedures to a game model based on a population system.

3.5.1 A dynamic population system

A population system is managed by players 1 and 2. The natural birth and death rates per individual are $\lambda > 0$ and $\mu > 0$, respectively. Player 1 is interested in the system having a large population and, to this end, player 1 can decrease the mortality rate (for instance, by using a suitable medical policy). On the other hand, the goal of player 2 is to have a small number of individuals; player 2 can choose policies that decrease the birth rate of the system (e.g., discouraging immigration).

We consider the following game model.

- The state space, standing for the number of individuals in the population, is $S = \{0, 1, 2, \ldots\}$.
- The action sets of the players are $A = B = [-1, 1]$, while $A(i) = B(i) = [-1, 1]$ for all $i \in S$.
- The system's transition rates $q_{ij}(a, b)$ satisfy $q_{ij}(a, b) = 0$ when $|i - j| > 1$. When $|i - j| \leq 1$ we let
$$q_{01}(a, b) = -q_{00}(a, b) = \lambda - C_b|b|,$$
and, for $i \geq 1$,
$$q_{i,i-1}(a, b) = \mu i - C_a|a|\sqrt{i}, \quad q_{i,i+1}(a, b) = \lambda i - C_b|b|i,$$
with $q_{ii}(a, b) = -(q_{i,i-1}(a, b) + q_{i,i+1}(a, b))$, for some constants $0 < C_a < \mu$ and $0 < C_b < \lambda$, and all $(a, b) \in A \times B$.
- The payoff rate (interpreted as a reward for player 1 and a cost for player 2) is given by
$$r(i, a, b) = p\, i + C_r ab\sqrt{i} \quad \text{for } i \in S \text{ and } -1 \leq a, b \leq 1,$$
for some constants $p > 0$ and $C_r > 0$.

In the above definitions, the term $\sqrt{i}$ models the fact that the payoff has a concave behavior with respect to the population size, while the term ab in the payoff rate captures the interplay between the actions of the players. Note that when the players take the actions $a = 0$ and $b = 0$ then they do not act on the dynamic system. In this case, the corresponding Markov process (referred to as the natural population system) is recurrent when $\lambda \leq \mu$ and transient when $\lambda > \mu$.

3.5.2 The discounted game

We consider the Lyapunov function w given by $w(i) = (\lambda + \mu + 1) \cdot (i + 1)$ for $i \in S$. Now we focus on the discounted payoff optimality criterion. Suppose that $\alpha > 0$ is the discount rate for the payoffs of the players.

Proposition 3.5.1 *Consider the population model $\mathcal{G}$ defined above. If the discount rate $\alpha > 0$ satisfies $\lambda - \mu < \alpha$ then Assumptions 3.1.2, 3.1.4, and 3.1.7 hold. If, in addition, we have $2(\lambda - \mu) < \alpha$ then Assumption 3.3.3 is satisfied.*

Proof: Fix an integer $k \geq 1$ and consider the Lyapunov function $i \mapsto w^k(i)$. Given a state $i \geq 1$ we have

$$\sum_{j \in S} q_{ij}(a,b) w^k(i) = (w^k(i-1) - w^k(i))(\mu i - C_a|a|\sqrt{i}) + (w^k(i+1) - w^k(i))(\lambda i - C_b|b|i).$$

Noting that

$$(i+2)^k - (i+1)^k = k(i+1)^{k-1} + \mathrm{O}(i^{k-2}) \quad \text{and} \quad i^k - (i+1)^k = -k(i+1)^{k-1} + \mathrm{O}(i^{k-2}),$$

some elementary calculations give

$$\sum_{j \in S} q_{ij}(a,b) w^k(i) = k(\lambda - \mu - C_b|b|) w^k(i) + \mathrm{O}(i^{k-\frac{1}{2}}) \leq k(\lambda - \mu) w^k(i) + \mathrm{O}(i^{k-\frac{1}{2}}).$$

Therefore, given an integer $k \geq 1$ and a constant $c_k < k(\mu - \lambda)$, there exists $d_k \geq 0$ such that

$$\sum_{j \in S} q_{ij}(a,b) w^k(i) \leq -c_k w^k(i) + d_k \quad \text{for all } (i,a,b) \in \mathbb{K}.$$

Note also that $-q_{ii}(a,b) \leq w(i)$ for all $(i,a,b) \in \mathbb{K}$, and so Assumptions 3.1.2 and 3.1.7(iii) hold.

If $\lambda - \mu < \alpha$ then choose $-\alpha < c_1 < \mu - \lambda$, and so Assumption 3.1.4(i) holds. Regarding the other assumptions, note that Assumption 3.1.4(ii) holds by letting $M = p + C_r$, while Assumptions 3.1.7(ii)–(iii) are straightforward.

It should be clear that Assumption 3.3.3(ii) is satisfied. If $2(\lambda - \mu) < \alpha$, then choose $\delta > 2$ and c_δ such that

$$-\alpha < c_\delta < \delta(\mu - \lambda),$$

and so Assumption 3.3.3(iii) holds. □

For each $n \geq 1$, consider now the finite state and actions game model $\mathcal{G}_n$ as described in Section 3.3.2. As a consequence of Theorems 3.3.1, 3.3.2, and 3.3.6 we obtain the following results.

(i) *Case $\lambda \leq \mu$ (the natural population system is recurrent).* Given arbitrary discount rate $\alpha > 0$ we have

$$\lim_{n \to \infty} V_n^\alpha(i) = V^\alpha(i) \quad \text{for all } i \in S.$$

Given arbitrary $k > 0$, by suitably choosing the action sets $A_n(i)$ and $B_n(i)$ we have

$$|V_n^\alpha(i) - V^\alpha(i)| = \mathrm{O}(n^{-k}) \quad \text{for each } i \in S.$$

(ii) *Case $\lambda > \mu$ (the natural population system is transient).* Given a discount rate $\lambda - \mu < \alpha$ we have

$$\lim_{n\to\infty} V_n^\alpha(i) = V^\alpha(i) \quad \text{for all } i \in S.$$

If the discount rate is such that $2(\lambda - \mu) < \alpha$ then for each $0 < k < \frac{\alpha}{\lambda-\mu} - 2$ we can choose the finite sets $A_n(i)$ and $B_n(i)$ such that

$$|V_n^\alpha(i) - V^\alpha(i)| = \mathrm{O}(n^{-k}) \quad \text{for each } i \in S.$$

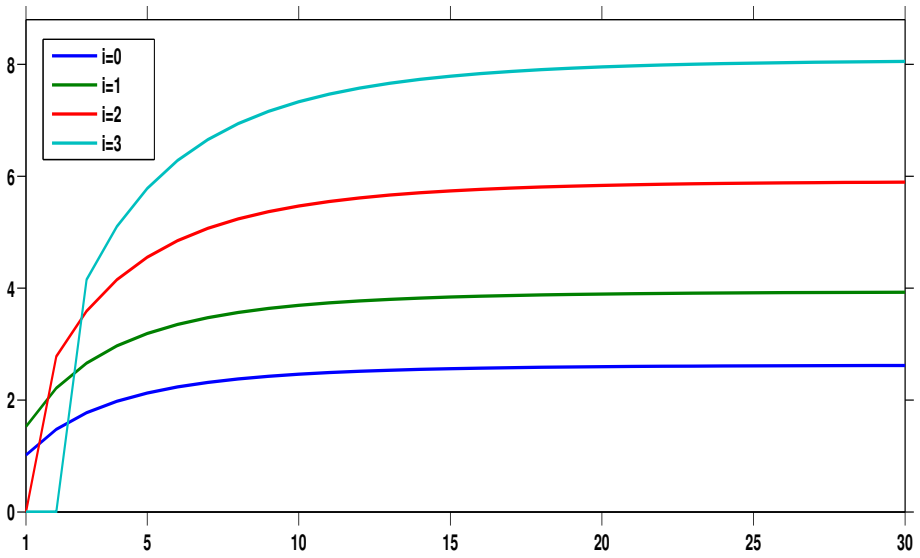


Figure 3.1: Value of the games $V_n^\alpha(i)$ for $n = 1, \ldots, 30$.

For the numerical experimentation we choose the following values of the parameters:

$$\lambda = 2.6, \quad \mu = 2.5, \quad \alpha = 1.2, \quad C_a = C_b = C_r = 0.2, \quad \text{and} \quad p = 3.$$

For each $n \geq 1$ we consider the truncated game model $\mathcal{G}_n$ with state space $\{0, 1, \ldots, n\}$. The action sets $A_n(i) \equiv A_n$ and $B_n(i) \equiv B_n$ consist of the $n+1$ points $\frac{2k}{n} - 1$ for $k = 0, 1, \ldots, n$.

For $n = 1, \ldots, 30$, we solve the finite game model $\mathcal{G}_n$ by using the value iteration procedure described in Section 3.3.3: we start from the initial value $u_0 = \mathbf{0}$, while $u_{k+1} = \widetilde{T}_n u_k$ for $k \geq 0$, and we let

$$\mathbf{q}_n = \max_{(i,a,b)\in\mathbb{K}_n} \{-q_{ii}^n(a,b)\} + 0.1;$$

recall (3.3.10). As a stopping criterion for the value iteration algorithm, we let $\varepsilon = 5 \times 10^{-5}$ and we stop at the iterate k when

$$||u_k - u_{k-1}|| \leq \varepsilon\alpha/\mathbf{q}_n,$$

		Actions in A_{30} and B_{30}.						
		-1.000	-0.933	-0.867	$\ldots$	0.867	0.933	1.000
Player 1	$i=0$	0.033	0.033	0.033	$\ldots$	0.033	0.033	0.000
	$i=1$	0.499	$< 10^{-4}$	$< 10^{-4}$	$\ldots$	$< 10^{-4}$	0.499	0.000
Player 2	$i=0$	0.499	$< 10^{-4}$	$< 10^{-4}$	$\ldots$	$< 10^{-4}$	0.499	0.000
	$i=1$	0.499	$< 10^{-4}$	$< 10^{-4}$	$\ldots$	$< 10^{-4}$	0.499	0.000

Table 3.1: Optimal strategies in $\overline{A}_{30}$ and $\overline{B}_{30}$ for $\mathcal{G}_{30}$.

which ensures that $\|u_k - V_n^\alpha\| \leq \varepsilon$ (this refers to the supremum norm in $\mathbb{R}^{n+1}$).

In Figure 3.1 we display the values $V_n^\alpha(i)$ for $i = 0, 1, 2, 3$ and $1 \leq n \leq 30$. We observe that the values of the games $\mathcal{G}_n$ become stable for relatively small values of the truncation size n, say for $n \geq 20$. We obtain the approximations

$$V^\alpha(0) \simeq 2.6179, \quad V^\alpha(1) \simeq 3.9269, \quad V^\alpha(2) \simeq 5.8948, \quad V^\alpha(3) \simeq 8.0524.$$

By Lemma 3.3.8, the approximation error (with respect to the value V_{30}^α of the game model $\mathcal{G}_{30}$) is less than 5×10^{-5}. Empirically, we observe that convergence seems to occur faster than at the convergence rate given in Theorem 3.3.6. This is because the bounds used to derive the convergence rate are very conservative.

Concerning the approximation of optimal strategies, for $n = 30$ we show in Table 3.1 the randomized strategies $\pi_*^1(\cdot|i)$ and $\pi_*^2(\cdot|i)$ for $i = 0$ and $i = 1$ as described in Lemma 3.3.8. Table 3.1 displays the corresponding probability distributions on the discretized sets of actions $A_{30} = B_{30}$. These are 10^{-4}-optimal strategies. Empirically, this suggests that the optimal strategy for player 1 in the game model $\mathcal{G}$ will be to choose his actions uniformly on $[-1, 1]$ in state $i = 0$, and to randomize between actions -1 and 1, with probabilities $1/2$, in state $i = 1$. For player 2, the estimation of an optimal strategy is to randomize between actions -1 and 1, with probabilities $1/2$, in both states $i = 0$ and $i = 1$.

3.5.3 The average game

Now we consider the game model $\mathcal{G}$ under the average payoff optimality criterion. The proof of our next result is omitted because it is similar to that of Proposition 3.5.1.

Proposition 3.5.2 *If $\mu > \lambda$ then Assumptions 3.1.11, 3.1.12, and 3.4.6 hold, with w the Lyapunov function $w(i) = (\lambda + \mu + 1) \cdot (i + 1)$ for $i \in S$.*

For $n \geq 1$, let $\mathcal{G}_n$ be the game model described previously, namely, its state space is $S_n = \{0, 1, \ldots, n\}$ and its action sets are $A_n(i) = B_n(i) = \{\frac{2k}{n} - 1 : k = 0, 1, \ldots, n\}$ for $i \in S_n$.

The game $\mathcal{G}_n$ has a value g_n^* (indeed, stationary policies are irreducible; recall Assumption 3.2.7(v) and Theorem 3.2.8) and, as a consequence of Theorem 3.4.7, we have that g_n^*

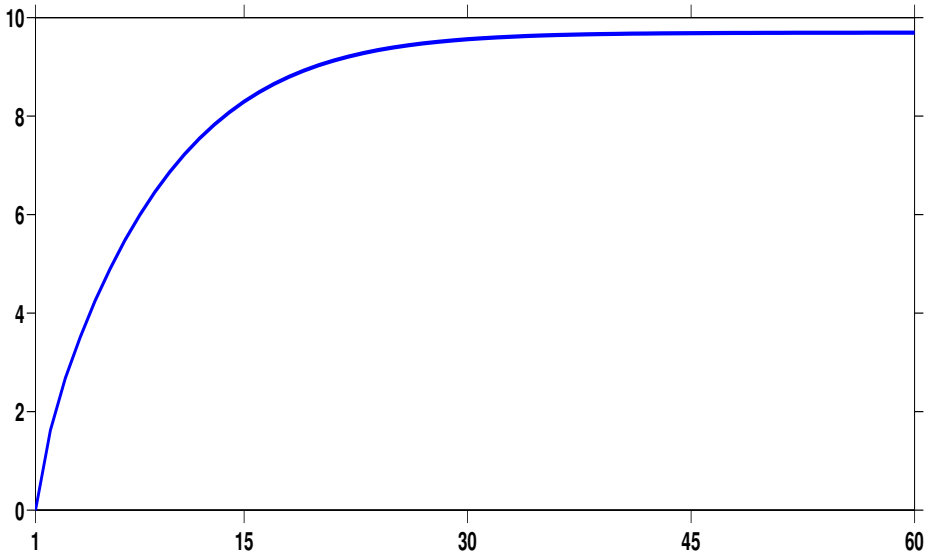


Figure 3.2: Value g_n^* of the game $\mathcal{G}_n$ for $n = 1, \ldots, 60$.

converges to the value g^* of $\mathcal{G}$ at a rate n^{-k} for any $k \geq 1$, by suitably choosing the finite actions sets, provided that $\mu > \lambda$.

For the numerical experimentation we choose the following values of the parameters: $\lambda = 2.2$, $\mu = 2.5$, $C_a = 0.8$, $C_b = C_r = 0.2$, and $p = 3$. For $n = 1, \ldots, 60$, we solve the finite game model $\mathcal{G}_n$ by using the policy iteration procedure described in Theorem 3.4.8. Figure 3.2 displays the value g_n^* as a function of $n = 1, \ldots, 60$. Empirically, we indeed observe a quick convergence and we obtain the approximate value $g^* \simeq 9.694$. This is in accordance with our theoretical results.

Chapter 4

Conclusions

In this thesis we have introduced the notion of convergence of control and game models; namely, we have studied the convergence of a sequence of approximating models to an original control model. This definition of convergence of $\mathcal{M}_n$ to $\mathcal{M}$, or $\mathcal{G}_n$ to $\mathcal{G}$, mainly relies on:

- Convergence of the state space;
- Convergence of the action sets in the Hausdorff metric;
- Uniform convergence of the transition and reward rates.

If the approximating models and the original model satisfy similar hypotheses, then the above referred convergence implies convergence of the optimal value function and the optimal policies. These theoretical results find direct applications by using the finite state and action truncations. Moreover, under some additional conditions, explicit rates of convergence of the optimal values can be obtained. Our numerical results show that the techniques developed herein are computationally efficient and can be used in practice to solve approximately control and game problems.

From a technical point of view, and as it has been seen, the analysis of game models is more complicated than studying control problems. Indeed, when solving a control problem, the corresponding dynamic programming equation is concerned with a "maximization" operator, whereas for game models it is a "maxmin" or "minmax" operator. This makes that our proofs for game models here are more subtle than those for control models. Another important difference between control and game models is that, for the discounted and average optimality criteria, deterministic stationary policies are a sufficient class for control problems, while in game models we need to consider randomized stationary strategies. This makes that the finite state and action truncated models are of a finite nature for control models (the family of deterministic stationary policies is finite), but the game model is still of a continuous nature (the class of stationary strategies is uncountable). A consequence of this fact is, as an illustration, that the policy iteration algorithm for control models converges in a finite number of steps, but, on the other hand, the policy iteration algorithm does not necessarily finitely converge for a game model.

A recurrent issue throughout this thesis is the use of Lyapunov conditions on the control and game models, related to a so-called Lyapunov function w. Such conditions can be seen as the core of the assumptions imposed on the control and game models. Indeed, the function w, which somehow bounds the reward and transition rates, and the Lyapunov conditions are used to:

(a) Prove the existence of the dynamic system itself (existence of the Markov process),

(b) Ensure the use of Dynkin's formula,

(c) Obtain convergence rates of the optimal value functions.

In this sense, one typically needs a Lyapunov condition on w to obtain (a), a Lyapunov condition on w^2 for (b), and a Lyapunov condition on w^δ for some $\delta > 2$ to obtain (c). Therefore, the technique of the Lyapunov conditions on the powers of w turns out to be a powerful tool to obtain interesting results on the control and game models studied here. In particular, the convergence rate of the optimal value functions closely depends on the maximal exponent $\delta > 2$ for which a suitable Lyapunov condition holds. Moreover, from a practical point of view, and as can be seen in the applications sections, verifying or discarding such Lyapunov conditions is usually a quite easy task, and sufficient conditions for these can be expressed, most of the time, by simple conditions on the parameters of the dynamic system.

Finally, let us mention some open issues. It would be interesting to study the approximation techniques developed in this thesis for finite horizon control and game models. In this case, optimal strategies are not, in general, stationary, and the corresponding optimality equation incorporates a differentiation with respect to the time component (such term does not appear in discounted and average models due to their stationary nature). It would be also interesting to know whether the approximation techniques can be used to study refined optimality criteria such as, e.g., bias optimality. In this case, optimal policies are derived by solving two nested optimality equations. Adapting our techniques to approximate such optimal policies is indeed a challenging open issue because the error in the first optimality equation would be somehow transferred to the second optimality equation, with, in addition, its own approximation error. Hence, it is not yet clear at all how to tackle such optimality criteria with our approximation techniques.

Bibliografía

[1] Akian, M., Cochet-Terrasson, J., Detournay, S., Gaubert, S. (2012). Policy iteration algorithm for zero-sum multichain stochastic games with mean payoff and perfect information. *arXiv:1208.0446*.

[2] Altman, E. (1994). Denumerable constrained Markov decision processes and finite approximations. *Math. Oper. Res.* **19**, pp. 169–191.

[3] Álvarez-Mena, J., Hernández-Lerma, O. (2002). Convergence of the optimal values of constrained Markov control processes. *Math. Methods Oper. Res.* **55**, pp. 461–484.

[4] Anderson, W.J. (1991). *Continuous-Time Markov Chains.* Springer, New York.

[5] Bertsekas, D.P. (2001). *Dynamic Programming and Optimal Control, Vol. II.* Athena Scientific, Belmont, MA.

[6] Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[7] Bogachev, V.I. (2007). *Measure Theory. Volume II.* Springer, New York.

[8] Bolley, F. (2008). Separability and completeness for the Wasserstein distance. In *Séminaire de probabilités XLI, pp. 371–377. Lecture Notes in Math.* **1934**, Springer, Berlin.

[9] Chang, H.S., Hu, J.Q., Fu, M.C., Marcus, S.I. (2010). Adaptive adversarial multi-armed bandit approach to two-person zero-sum Markov games. *IEEE Trans. Automat. Control* **55**, pp. 463–468.

[10] Filar, J.A., Schultz, T.A., Thuijsman, F., Vrieze, O.J. (1991). Nonlinear programming and stationary equilibria in stochastic games. *Math. Program.* **50**, pp. 227–237.

[11] Fisher, L. (1968). On recurrent denumerable decision processes. *Ann. Math. Statist.* **39**, pp. 424–432.

[12] Frenk, J.B.G., Kassay, G., Kolumbán, J. (2004). On equivalent results in minimax theory. *Euro. J. Oper. Res.* **157**, pp. 46–58.

[13] Guo, X.P., Hernández-Lerma, O. (2003). Continuous-time controlled Markov chains with discounted rewards. *Acta Appl. Math.* **79**, pp. 195–216.

[14] Guo, X.P., Hernández-Lerma, O. (2003). Drift and monotonicity conditions for continuous-time controlled Markov chains with an average criterion. *IEEE Trans. Automat. Control* **48**, pp. 236–244.

[15] Guo, X.P., Hernández-Lerma, O. (2003). Zero-sum games for continuous-time Markov chains with unbounded transition and average payoff rates. *J. Appl. Probab.* **40**, pp. 327–345.

[16] Guo, X.P., Hernández-Lerma, O. (2005). Zero-sum continuous-time Markov games with unbounded transition and discounted payoff rates. *Bernoulli* **11**, pp. 1009–1029.

[17] Guo, X.P., Hernández-Lerma, O. (2009). *Continuous-Time Markov Decision Processes: Theory and Applications*. Springer, New York.

[18] Guo, X.P., Hernández-Lerma, O., Prieto-Rumeau, T. (2006). A survey of recent results on continuous-time Markov decision processes. *Top* **14**, pp. 177–261.

[19] Guo, X.P., Zhang, W.Z. (2014). Convergence of controlled models and finite-state approximation for discounted continuous-time Markov decision processes with constraints. *Euro. J. Oper. Res.* **238**, pp. 486–496.

[20] Hernández-Lerma, O. (1989). *Adaptive Markov Control Processes*. Springer, New York.

[21] Jaśkiewicz, A., Nowak, A.S. (2006). Approximation of noncooperative semi-Markov games. *J. Optim. Theory Appl.* **131**, pp. 115–134.

[22] Kemeny, J.G., Snell, J.L., Thompson, G.L. (1974). *Introduction to Finite Mathematics. Third edition.* Prentice-Hall, Englewood Cliffs, NJ.

[23] Kushner, H.J., Dupuis, P. (2001). *Numerical Methods for Stochastic Control Problems in Continuous Time. Second Edition.* Springer, New York.

[24] Langen, H.J. (1981). Convergence of dynamic programming models. *Math. Oper. Res.* **6**, pp. 493–512.

[25] Leizarowitz, A., Shwartz, A. (2008). Exact finite approximations of average-cost countable Markov decision processes. *Automatica J. IFAC* **44**, pp. 1480–1487.

[26] Lorenzo, J.M., Hernández-Noriega, I., Prieto-Rumeau, T. (2015). Approximation of two-person zero-sum continuous-time Markov games with average payoff criterion. *Oper. Res. Lett.* **43**, pp. 110–116.

[27] Lund, R.B., Meyn, S.P., Tweedie, R.L. (1996). Computable exponential convergence rates for stochastically ordered Markov processes. *Ann. Appl. Prob.* **6**, pp. 218–237.

[28] Nowak, A.S., Altman, E. (2002). ϵ-equilibria for stochastic games with uncountable state space and unbounded costs. *SIAM J. Control Optim.* **40**, pp. 1821–1839.

[29] Prieto-Rumeau, T., Hernández-Lerma, O. (2010). Policy iteration and finite approximations to discounted continuous-time controlled Markov chains. In *Modern Trends in Controlled Stochastic Processes: Theory and Applications*, ed. by A. Piunovskiy, Luniver Press, pp. 84–101.

[30] Prieto-Rumeau, T., Hernández-Lerma, O. (2012). Discounted continuous-time controlled Markov chains: convergence of control models. *J. Appl. Probab.* **49**, pp. 1072–1090.

[31] Prieto-Rumeau, T., Hernández-Lerma, O. (2012). *Selected Topics on Continuous-Time Controlled Markov Chains and Markov Games*. Imperial College Press, London.

[32] Prieto-Rumeau, T., Hernández-Lerma, O. (2012). Uniform ergodicity of continuous-time controlled Markov chains: A survey and new results. *Ann. Oper. Res.* DOI:10.1007/s10479-012-1184-4.

[33] Prieto-Rumeau, T, Lorenzo, J.M. (2010). Approximating ergodic average reward continuous-time controlled Markov chains. *IEEE Trans. Automat. Control* **55**, pp. 201–207.

[34] Prieto-Rumeau, T., Lorenzo, J.M. (2015). Approximation of zero-sum continuous-time Markov games under the discounted payoff criterion. *Top* **23**, pp. 799–836.

[35] Puterman, M.L. (1994). *Markov Decision Processes. Discrete Stochastic Dynamic Programming*. Wiley, New York.

[36] Sennott, L.I. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems*. Wiley, New York.

[37] Santos, M.S., Rust, J. (2004). Convergence properties of policy iteration. *SIAM J. Control Optim.* **42**, pp. 2094–2115.

[38] Song, Q.S. (2008). Convergence of Markov chain approximation on generalized HJB equation and its applications. *Automatica J. IFAC* **44**, p. 761–766.

[39] Tidball, M.M., Lombardi, A., Pourtallier, O., Altman, E. (2000). Continuity of optimal values and solutions for control of Markov chains with constraints. *SIAM J. Control Optim.* **38** pp. 1204–1222.

[40] Whitt, W. (1978). Approximation of dynamic programs. *Math. Oper. Res.* **3**, pp. 231–243.