Staudigl, Mathias

Working Paper

A limit theorem for Markov decision processes

Working Papers, No. 475

Provided in Cooperation with:
Center for Mathematical Economics (IMW), Bielefeld University

Suggested Citation: Staudigl, Mathias (2013) : A limit theorem for Markov decision processes, Working Papers, No. 475, Bielefeld University, Institute of Mathematical Economics (IMW), Bielefeld, https://nbn-resolving.de/urn:nbn:de:0070-pub-26740390

This Version is available at:
https://hdl.handle.net/10419/81097

Working Papers

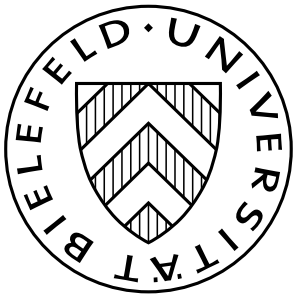
Institute of Mathematical Economics

475

February 2013

A limit theorem for Markov decision processes

Matthias Staudigl



IMW · Bielefeld University
Postfach 100131
33501 Bielefeld · Germany

email: imw@wiwi.uni-bielefeld.de
http://www.imw.uni-bielefeld.de/research/wp475.php
ISSN: 0931-6558

A limit theorem for Markov decision processes*

Mathias Staudigl†

February 20, 2013

Abstract

In this paper we prove a deterministic approximation theorem for a sequence of Markov decision processes with finitely many actions and general state spaces as they appear frequently in economics, game theory and operations research. Using viscosity solution methods no a-priori differentiabililty assumptions are imposed on the value function. Applications for this result can be found in large deviation theory, and some simple economic problems.

Keywords: Markov decision processes, Optimal Control, Viscosity solutions, Stochastic Approximation
JEL Classification Numbers: C02, C44, C61

1. Introduction

In this paper we study the following standard sequential decision problem. Consider a controlled Markov chain $\{X_n^\varepsilon\}_{n\in\mathbb{N}_0}$ defined on some probability space $(\Omega, \mathcal{F}, P)$, and taking values in $\mathbb{R}^d$. The evolution of this process is controlled by an action process $\{A_n^\varepsilon\}_{n\in\mathbb{N}_0}$, which is assumed to take values in a finite set of available actions $\mathcal{A}$. The controlled evolution of the state is assumed to follow the system equation

$$\begin{cases} X_{n+1}^\varepsilon = X_n^\varepsilon + \varepsilon f_{n+1}^\varepsilon(X_n^\varepsilon, A_n^\varepsilon) & \forall n \in \mathbb{N}_0 \\ X_0^\varepsilon = x \in \mathcal{X} \subset \mathbb{R}^d. \end{cases} \tag{1}$$

*This paper was written while I was visiting Nuffield College at the University of Oxford. I thank my sponsor, Peyton Young, for his support. I also thank Frank Riedel, Jan-Henrik Steg, Immanuel Bomze and Bill Sandholm for useful comments. Financial support from U.S. Air Force OSR Grant FA9550-09-0538 are gratefully acknowledged.

†Center of Mathematical Economics (IMW), Bielefeld University, Germany. e-mail: mathias.staudigl@uni-bielefeld.de; website: mwpweb.eu/MathiasStaudigl.

Assume that real time is a continuous variable, taking values in the set of non-negative real numbers $t \in \mathbb{R}_+$. Fitting the discrete process $\{(X_n^\varepsilon, \hat{A}_n^\varepsilon)\}_{n\in\mathbb{N}}$ into continuous time by defining processes

$$\hat{X}^\varepsilon(t) = X_n^\varepsilon, \text{and } \hat{A}^\varepsilon(t) = A_n^\varepsilon$$

for $t \in [n\varepsilon, (n+1)\varepsilon), n \in \mathbb{N}_0$, we obtain a jump process, with deterministic periods between consecutive jumps of length ε. Consider a decision maker, whose objective is to maximize his total sum of stage payoffs over an infinite time horizon and discount factor $\lambda_\varepsilon := e^{-r\varepsilon}$. Assume that the decision maker is an expected utility maximizer so that his preferences are given by

$$U(x,\sigma) = E_x^\sigma\left[\sum_{n=0}^{\infty}(1-\lambda_\varepsilon)\lambda_\varepsilon^n u(X_n^\varepsilon, A_n^\varepsilon)\right],$$

or in continuous-time formulation

$$E_x^\sigma\left[\int_0^\infty re^{-rt}u(\hat{X}^\varepsilon(t), \hat{A}^\varepsilon(t))\mathrm{d}t\right].$$

The mapping σ is a (behavior) *strategy* for the decision maker, essentially describing a probability distribution over actions at each decision node. Precise definitions are given in Section 2.1.
As a comparison problem consider the deterministic optimal control problem

$$\begin{aligned}&\sup_{\alpha\in\mathcal{S}}\int_0^\infty re^{-rt}u(y_x(t,\alpha),\alpha(t))\mathrm{d}t\\ &\text{s.t. } \dot{y}_x(t,\alpha) = b(y_x(t,\alpha),\alpha(t)),\ y_x(0,\alpha) = x\end{aligned}$$

where $\mathcal{S}$ is the set of measurable open-loop controls $\alpha : \mathbb{R}_+ \to \Delta(\mathcal{A})$, and b is a suitably defined Lipschitz continuous and bounded vector field. In this paper we address the question under which conditions solutions (i.e. value function and the strategies) of the stochastic sequential decision model, with decisions made on the discrete time grid $\{0, \varepsilon, 2\varepsilon, \ldots\}$, converge to solutions of the deterministic optimal control problem described above. The motivation for studying this question are two-fold. The first motivation is guided by practical considerations. There are some arguments in favor of using deterministic continuous optimal control problems over the stochastic discrete decision processes. Solving the stochastic decision problem numerically is often a computationally

very intensive task, due to the "curse of dimensionality" of dynamic programming.[1] The deterministic optimal control problem is often amenable to efficient numerical methods which seem to perform better than algorithms based on dynamic programming. Second, in some situations, the continuous deterministic formulation allows for an analytic treatment of the decision problem, using either Viscosity solution methods, or the more traditional Pontryagin Maximum principle. Hence, if one has the theoretical justification to replace the stochastic decision problem by a deterministic one, there are some good reasons to do that. My second motivation for investigating this question is in establishing convergence results for dynamic games in discrete time to dynamic games in continuous time. The present paper is therefore the basis for a model in which the limit dynamic game is characterized by a deterministic ordinary differential equation (i.e. a differential game). A task for future research is to extend this to allow stochastic limiting dynamics, in particular jump-diffusion processes.

Related convergence and approximation questions are at the core of optimal control theory. Indeed the present study is heavily influenced by the so-called Markov chain method developed by Kushner and Dupuis (2001). This is a powerful numerical approximation tool to obtain feedback controls in stochastic and deterministic optimal control problems. Similar approaches can be found in Capuzzo-Dolcetta and Ishii (1984); Gonzales and Rofman (1985); Falcone (1987) and Bardi and Capuzzo-Dolcetta (1997). Our proof method uses weak convergence arguments, as these are more naturally adapted to our probabilistic setting. The difference between these papers and the present one is the nature of the question I am addressing. While the above mentioned literature is interested to construct a numerical approximation scheme in order to approximate a given optimal control problem, I instead ask the question, given a discrete controlled Markov chain model, what is the limit as the discretization becomes arbitrarily fine? Therefore this paper is closer in spirit to stochastic approximation theory (Benaïm, 1998). While writing this paper I have learned from the paper by Gast et al. (2012). They establish a limit result for a finite-horizon Markov decision process converging to a deterministic optimal control problem. This paper differs from Gast et al. (2012) in the problem formulation as well as in the proof techniques. First I study infinite horizon problems with discounting. Second, my proof techniques are based on dynamic programming and viscosity solution techniques, whereas Gast et al. (2012) rely on ideas from stochastic approximation theory. Before developing the general analysis of the problem, let me introduce some concrete

[1]Note that for numerical implementation of the decision problem one needs to discretize the state space somehow. Usually at this stage the curse of dimensionality kicks in.

examples to which the limit results apply.

1.1 Examples

1.2 Optimal pricing policy of a Monopoly

Consider an infinitely lived monopoly, who sets prices $a \in \mathcal{A} = \{1, 2, \ldots, m\}$. The monopolist can announce prices at the periods $\{0, \varepsilon, 2\varepsilon, \ldots\}$. It faces a stochastic market demand, following a Markovian dynamics $\{X_n^\varepsilon\}_{n\in\mathbb{N}_0}$ with sample paths given by (1). The vector field $f_n^\varepsilon(x, a)$ capture the random changes in market demand, given the current demand is x and the quoted price is $a \in \mathcal{A}$. The probability measures $\mu_a^\varepsilon(\cdot|x)$ define the law of the random changes in demand, given the current demand is x and the monopolist announces a price a. The monopolist has a flow profit function $u(x, a)$. A strategy for the monopolist is to design an optimal pricing strategy $\{\sigma_n\}_{n=0}^\infty$, where σ_n is a function of the demand history to probability distributions over prices. Hence, the monopolists' problem is to maximize

$$U(x, \sigma) = E_x^\sigma \left[\sum_{n=0}^{\infty} (1 - \lambda_\varepsilon)\lambda_\varepsilon^n u(X_n^\varepsilon, A_n^\varepsilon) \right]$$

where $x \in \mathbb{R}$ is the initially given demand, assumed to be known to the monopolist. As $\varepsilon \to 0$ the monopolist is able to post prices in arbitrary short time spans, and thus can react arbitrarily fast to the random market demand. If the market is sufficiently stable where random fluctuations over very small time spans are negligible, a deterministic approximation to this models seems to the sensible.

1.2.1 Optimal stopping

A firm hast to decide when to exit an industry. The state of the market is modeled by a discrete-time Markov chain $\{X_n^\varepsilon\}_{n\in\mathbb{N}_0}$ which lives on $\mathbb{R}_+$. For concreteness think of X_n^ε as the market price in period n. Real time t takes values in the set of non-negative reals $\mathbb{R}_+$ and the firm receives information on the prevailing market price only at discrete points in time contained in the grid $\{0, \varepsilon, 2\varepsilon, \ldots\}$. The firm is small, and therefore cannot influence the evolution of the price dynamics. However, it has a model for the time series of prices, which is the AR(1) process given by eq. (1).

In each period the firm can decide whether to stay or exit the market. This is modeled by a binary action set $\mathcal{A} = \{0, 1\}$, where action 0 means to exit the market and 1 means to stay in the market. In each period in which the firm stays in the market it has to pay a

random fee $-r(X_n^\varepsilon) < 0$, and the state evolves according to an uncontrolled Markov chain with transition function q^ε on a set of possible prices $\mathcal{X} \subseteq \mathbb{R}_+$. If the firm decides to exit the market in period $N \in \mathbb{N}$ it gets a terminal reward $g(X_N^\varepsilon)$ and the evolution of prices stops (or the firm does simply not monitor the price evolution anymore). The function $g(\cdot)$ is non-negative (otherwise the firm would want to exit immediately) and bounded. This problem is contained in our model setup by specifying the following data. The transition dynamics are $\mu_0^\varepsilon(\cdot|x) = \delta_0$ and $\mu_1^\varepsilon(\cdot|x) = q^\varepsilon(\cdot|x) \in \mathsf{M}_1^+(\mathbb{R})$, where $q^\varepsilon(\cdot|x)$ is a given probability law modeling the uncontrolled evolution of the price time series. The utility rate function is given by

$$u(x,a) = \begin{cases} -r(x) & \text{if } a = 1, \\ g(x) & \text{if } a = 0. \end{cases}$$

The objective function of the decision maker is

$$U^\varepsilon(x,\sigma) = E_x^\sigma\left[\sum_{n=0}^{\infty}(1-\lambda_\varepsilon)\lambda_\varepsilon^n u(X_n^\varepsilon, A_n^\varepsilon)\right]$$

where σ is a measurable function mapping histories of the state process into probability distributions over actions (i.e. a *strategy*). Now suppose that the information about current prices appears in periods of length ε. In real time, the price time series evolves therefore according to the step process $\hat{X}^\varepsilon$, and the decision whether to exit the market or stay in the market can be made at all time points which are multiples of the step size ε. In the limit as ε approaches 0 the firm monitors the price evolution with more and more accuracy, and can also react to the price dynamics at virtually any point in real time. The results reported in this paper investigate such a scenario where in the limit as $\varepsilon \to 0$ the limit price dynamics can be modeled by a deterministic differential equation.

2. Problem formulation

2.1 The discrete problem

Let $\{(X_n^\varepsilon, A_n^\varepsilon)\}_{n\in\mathbb{N}_0}$ be a stochastic process taking values in the set $\mathbb{R}^d \times \mathcal{A}$, whose sample paths satisfy the dynamical systems equation (1). Each A_n^ε is an $\mathcal{A}$-valued random variable, adapted to the filtration $\mathcal{F}_n^\varepsilon = \sigma\left(X_0^\varepsilon, \ldots, X_n^\varepsilon\right)$, and controlling the evolution of the state process. The law of the random variables A_n^ε for $n = 0, 1, 2, \ldots$ are determined by a (behavior) *strategy*. A *strategy* is a collection of functions $\sigma = \{\sigma_n\}_{n\in\mathbb{N}_0}$, where each $\sigma_n(\cdot)$

is a probability distribution over the finite set of actions $\mathcal{A} := \{1, 2, \ldots, m\}$, adapted to the sigma-algebra $\mathcal{F}_n^\varepsilon$.[2] A strategy is *Markov* if for every n we can express the behavior strategy σ_n in terms of a single function $\underline{\alpha} : \mathbb{R}^d \to \Delta(\mathcal{A})$, so that

$$\sigma_n(a|x_0, \ldots, x_n) = \underline{\alpha}(a|x_n) \quad \forall n \geq 0, a \in \mathcal{A}. \tag{2}$$

Markov strategies are of fundamental importance in Markov decision processes, as we will see in due course. $\{f_n^\varepsilon(x,a)\}_{n\in\mathbb{N}}$ is a sequence of i.i.d random variables with common law $\mu_a^\varepsilon(\cdot|x)$ on $\mathbb{R}^d$. The collection of probability distributions $\mu_1^\varepsilon(\cdot|x), \ldots, \mu_m^\varepsilon(\cdot|x)$ defined on the Borel sets of $\mathbb{R}^d$ are the control measures of the Markov decision process.
Let $\Omega = (\mathbb{R}^d \times \mathcal{A})^{\mathbb{N}_0}$ denote the sample path space of the controlled Markov chain, and let $\mathcal{F}$ denote the σ-algebra generated by the finite cylinder sets. By the Ionescu-Tulcea Theorem (see e.g. Bertsekas and Shreve, 1978), each strategy σ defines a unique probability measure P_x^σ on $(\Omega, \mathcal{F})$ with the following characteristics

$$\begin{aligned}
&P_x^\sigma(X_0^\varepsilon \in \Gamma) = \delta_x(\Gamma),\\
&P_x^\sigma(X_{n+1}^\varepsilon \in \Gamma | X_n^\varepsilon = x, A_n^\varepsilon = a) = Q^\varepsilon(\Gamma|x,a),\\
&P_x^\sigma(A_n^\varepsilon = a | X_0^\varepsilon, \ldots, X_n^\varepsilon) = \sigma(a|X_0^\varepsilon, \ldots, X_n^\varepsilon),
\end{aligned}$$

where Γ is Borel measurable subset of $\mathbb{R}^d$. The probability measure $Q^\varepsilon(\cdot|x,a)$ models the evolution of the state process, and is defined by

$$Q^\varepsilon(\Gamma|x,a) = \mu_a^\varepsilon\left(\frac{1}{\varepsilon}(\Gamma - x)|x\right) \quad \forall (x,a) \in \mathbb{R}^d \times \mathcal{A}.$$

Under this (canonical) construction of the controlled Markov chain we think of the random variables X_n^ε and A_n^ε as the coordinate processes $X_n^\varepsilon(\omega) = x_n$ and $A_n^\varepsilon(\omega) = a_n$, for every $\omega = (x_0, a_0, \ldots, x_n, a_n, \ldots) \in \Omega$.
Given a strategy σ let E_x^σ denote expectations with respect to the probability measure P_x^σ. The objective of the decision maker is to maximize his normalized expected infinite horizon discounted utility

$$U^\varepsilon(x,\sigma) = E_x^\sigma\left[\sum_{n=0}^{\infty}(1-\lambda_\varepsilon)\lambda_\varepsilon^n u(X_n^\varepsilon, A_n^\varepsilon)\right]. \tag{3}$$

[2]Technically speaking, each σ_n is a stochastic kernel on $\mathcal{A}$ given $(\mathbb{R}^d)^{n+1}$. See Bertsekas and Shreve (1978) for the precise measure-theoretic definition of stochastic kernels.

The discount factor per unit time λ_ε is defined as $\lambda_\varepsilon = e^{-r\varepsilon}$. $r > 0$ is the discount rate, or the interest rate per unit time. The factor $(1 - \lambda_\varepsilon)$ provides the correct normalization of the stream of utilities. The maximized utility of the decision maker, or the *value function*, is defined as

$$V^\varepsilon(x) = \sup_\sigma U^\varepsilon(x, \sigma). \tag{4}$$

Here the supremum is taken over all strategies available to the decision maker. A standard result in Markov decision processes is that the decision maker does not gain much by using more complicated strategies than Markov strategies. Indeed, for every fixed $\varepsilon > 0$ it is well known (see e.g. Puterman, 1994) that the decision maker can choose a Markov strategy $\underline{\alpha}^\varepsilon : \mathbb{R}^d \to \Delta(\mathcal{A})$ which solves the decision problem, i.e.

$$V^\varepsilon(x) = U^\varepsilon(x, \underline{\alpha}^\varepsilon) \quad \forall x \in \mathbb{R}^d.$$

2.1.1 Standing hypothesis

This section provides a collection of all the technical assumptions we impose on the problem data. The first assumption is a continuity assumption on the *drift* of the state process $\{X_n^\varepsilon\}_{n\in\mathbb{N}_0}$, defined as the conditional mean increment of the process. We denote the drif $b^\varepsilon : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$ by

$$b^\varepsilon(x, a) := E_x^\sigma\left[\frac{1}{\varepsilon}\left(X_{n+1}^\varepsilon - X_n^\varepsilon\right) \mid X_n^\varepsilon = x, A_n^\varepsilon = a\right] = \int_{\mathbb{R}^d} z\mu_a^\varepsilon(\mathrm{d}z|x). \tag{5}$$

Assumption 2.1. *The function $b^\varepsilon : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$ is Lipschitz continuous and converges to a Lipschitz continuous function $b : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$ locally uniformly on compact sets.*

In the convergence proof we will make the following fairly standard uniform integrability assumption on the control measures.

Assumption 2.2. *The control measures $\mu_1^\varepsilon(\cdot|x), \ldots, \mu_m^\varepsilon(\cdot|x)$ are supported on a common compact subset $\mathcal{K} \subset \mathbb{R}^d$ for each $x \in \mathbb{R}^d$.*

This assumption implies that the vector fields $b(x, a)$ are contained in the closed convex hull of the compact set $\mathcal{K}$. Hence, the averaged vector field of the dynamics is uniformly bounded by some constant $M_b > 0$, so that

$$\sup_{x\in\mathbb{R}^d} \|b(x, a)\| \le M_b \quad \forall (x, a) \in \mathbb{R}^d \times \mathcal{A}. \tag{6}$$

Now we impose some restriction on the utility flow function of the decision maker.

Assumption 2.3. *The utility flow function $u : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ is uniformly bounded and Hölder continuous for each action $a \in \mathcal{A}$:*

$$\sup_{x \in \mathbb{R}^d} |u(x,a)| \leq M_u \quad \forall a \in \mathcal{A}, \text{ and} \tag{7}$$

$$|u(x,a) - u(y,a)| \leq M_u \|x - y\|^\gamma \quad \forall x, y \in \mathbb{R}^d, a \in \mathcal{A} \tag{8}$$

for some constants $M_u > 0$ and $\gamma \in [0,1]$.

The final assumption we make concerns the scaling relationship between the variance of the increments of the state process and the step size ε. This assumption is essential in making the deterministic approximation result work, as it says that in the limit of small step sizes, sample paths of the state process look like solutions of an ordinary differential equation with drift b. This will be made precise in Section 6, where the technical details are provided.

Assumption 2.4. *The covariance matrix of the increments of the state process $\{X_n^\varepsilon\}_{n \in \mathbb{N}_0}$ satisfies the scaling relationship*

$$\mathrm{Var}_x^\sigma \left[X_{n+1}^\varepsilon - X_n^\varepsilon | X_n^\varepsilon = x, A_n^\varepsilon = a \right] \leq \varepsilon^2 M_v \tag{9}$$

for every $(x,a) \in \mathbb{R}^d \times \mathcal{A}$, for some uniform constant $M_v \geq 0$.

2.2 The Limit Problem

The limit problem is a deterministic optimal control problem where the decision maker wants to maximize his total discounted utility over an infinite time horizon. Extend the utility rate function to the domain $\mathbb{R}^d \times \Delta(\mathcal{A})$ linearly, so that $u(x,\alpha) := \sum_{a \in \mathcal{A}} u(x,a)\alpha(a)$. Similarly, extend the drift b to $\mathbb{R}^d \times \Delta(\mathcal{A})$ by $b(x,\alpha) := \sum_{a \in \mathcal{A}} b(x,a)\alpha(a)$. The value function of the optimal control problem is defined as

$$v(x) := \sup_{\alpha \in \mathcal{S}} U(x,\alpha), \tag{OC}$$

where

$$U(x,\alpha) := \int_0^\infty r e^{-rt} u(y_x(t,\alpha), \alpha(t)) \mathrm{d}t \tag{10}$$

is the utility function of the decision maker under the deterministic strategy $\alpha \in \mathcal{S}$ which induces the state dynamics

$$\dot{y}_x(t,\alpha) = b(y_x(t,\alpha),\alpha(t)),\ y_x(0,\alpha) = x. \tag{11}$$

Existence and uniqueness to solutions of the differential equation (11) is guaranteed by Assumption 2.1. The set of strategies the decision maker can choose is the set of measurable functions $\alpha : \mathbb{R}_+ \to \Delta(\mathcal{A})$,

$$\mathcal{S} := \{\alpha : \mathbb{R}_+ \to \Delta(\mathcal{A}) | \alpha(\cdot) \text{ measurable}\}.$$

Note that these functions are defined without any reference to the current state and hence are open-loop controls.
The following technical lemma establishes that the value funcion of the deterministic optimal control problem (OC) is an element of the space of continuous bounded functions $v \in \mathcal{C}_b(\mathbb{R}^d : \mathbb{R})$.

Lemma 2.5. *Under Assumptions 2.1, 2.2 and 2.3 the value function $v : \mathbb{R}^d \to \mathbb{R}$ satisfies*

$$|v(x)| \leq M_u \qquad \forall x \in \mathbb{R}^d, \tag{12}$$

and it is Hölder continuous with coefficient $\gamma \in (0, \min\{\frac{r}{M_b}, 1\})$.

Proof. The proof of this Lemma is based upon standard arguments, which can be found in Bardi and Capuzzo-Dolcetta (1997). The uniform boundedness of the value function is a trivial consequence of the uniform boundedness of the utility flow function u, stated in Assumption 2.3. Indeed, for any strategy $\alpha \in \mathcal{S}$, we have

$$U(x,\alpha) = \int_0^\infty re^{-rt}u(y_x(t,\alpha),\alpha(t))\mathrm{d}t \leq M_u r \int_0^\infty e^{-rt}\mathrm{d}t = M_u.$$

For the second statement, pick two points $x_1, x_2 \in \mathbb{R}^d$ and fix a strategy $\alpha \in \mathcal{S}$ such that

$$v(x_1) - \delta \leq U(x_1,\alpha).$$

Such a strategy exists by definition of the supremum. Now $v(x_2) \geq U(x_2,\alpha)$, and w.l.o.g we assume that $v(x_1) > v(x_2)$. Then

$$|v(x_1) - v(x_2)| \leq |U(x_1,\alpha) + \delta - U(x_2,\alpha)|$$

$$= |\int_0^\infty re^{-rt}[u(y_{x_1}(t,\alpha),\alpha(t)) - u(y_{x_2}(t,\alpha),\alpha(t))]\mathrm{d}t + \delta|.$$

By eq. (8) and standard estimates on solutions to ordinary differential equations, we see that

$$\begin{aligned}|u(y_{x_1}(t,\alpha),\alpha(t)) - u(y_{x_2}(t,\alpha),\alpha(t))| &\leq M_u\|y_{x_1}(t,\alpha(t)) - y_{x_2}(t,\alpha(t))\|^\gamma \\ &\leq M_u\|x_1 - x_2\|^\gamma e^{-M_b\gamma t}.\end{aligned}$$

Using this estimate in the previous display shows that

$$|v(x_1) - v(x_2)| \leq M_u\|x_1 - x_2\|^\gamma|\int_0^\infty e^{(-r+\gamma M_b)t}\mathrm{d}t| + 2\delta.$$

To ensure that the integral on the right-hand side of this estimate converges, we consider three cases. If $r > M_b$ then the condition $\gamma < r/M_b$ is sufficient for convergence. In particular $\gamma = 1$ can be chosen, which shows that the value function is Lipschitz in this case. If $r = M_b$ any choice $\gamma \in (0,1)$ can be made. Finally if $r < M_b$ then we need to pick $0 \leq \gamma < r/M_b$. This completes the proof the Lemma. □

The dynamic programming approach to deterministic optimal control theory allows us to characterize the value function as a solution to a partial differential equation of the first-order, known as the Hamilton-Jacobi-Bellman equation. The Hamiltonian associated to the optimal control problem (OC) is given by

$$H(x,p) = \max_{a\in\mathcal{A}}\{\langle p, b(x,a)\rangle + ru(x,a)\}.$$

Note that here we have already used the fact that the maximum value of the Hamiltonian expression will be attained at a pure action. It is well-known that, under the technical assumptions made in this paper, the value function v is the unique viscosity solution of the Hamilton-Jacobi-Bellman equation

(HJB) $\qquad rv(x) - H(x, Dv(x)) = 0 \quad \forall x \in \mathbb{R}^d.$

See Bardi and Capuzzo-Dolcetta (1997), chapters II and III. Since the Hamiltonian maximization condition can be formulated to optimize over elements in the finite action set $\mathcal{A}$, it follows that

$$v(x) = \sup_{\alpha\in\mathcal{S}^\#}\int_0^\infty re^{-rt}u(y_x(t,\alpha),\alpha(t))\mathrm{d}t$$

where $\mathcal{S}^{\#} \subset \mathcal{S}$ is the space of measurable $\mathcal{A}$-valued open-loop strategies. Measurable functions may display very irregular behavior so that strategies in the set $\mathcal{S}^{\#}$ will not in general provide good candidates for discrete approximations. Rather we would like to exhibit controls which may only be δ-optimal, but be at least piecewise constant.[3] Adapting results reported in Capuzzo-Dolcetta (1983), we will show that such suboptimal controls generally exist for the problem at hand. The proof is constructive, and is strongly related to the Markov decision process introduced in the previous section. To construct piecewise constant suboptimal strategies we replace the optimal control problem by a deterministic dynamic programming problem, which can be interpreted as the mean-field model of the Markov decision process. For each $\varepsilon > 0$ let

$$\mathcal{S}_{\varepsilon}^{\#} := \{\alpha \in \mathcal{S} | \alpha(\cdot) \text{ is piecewise constant on } [n\varepsilon, (n+1)\varepsilon), n \in \mathbb{N}_0\}. \tag{13}$$

For each strategy $\alpha \in \mathcal{S}_{\varepsilon}^{\#}$ define a controlled trajectory recursively on the time grid $\{0, \varepsilon, 2\varepsilon, \ldots\}$ by

$$\begin{aligned} y_x^{\varepsilon}(n\varepsilon, \alpha) &= x + \varepsilon \sum_{k=0}^{n-1} b(y_x(k\varepsilon), \alpha), \alpha(k\varepsilon)), \\ y_x^{\varepsilon}(0, \alpha) &= x. \end{aligned}$$

Interpolate the state trajectory by setting $y^{\varepsilon}(t, \alpha) = y^{\varepsilon}(n\varepsilon, \alpha)$ for each $t \in [n\varepsilon, (n+1)\varepsilon), n \in \mathbb{N}_0$. In terms of this continuous time interpolation it is easily seen, recalling the identity $\lambda_{\varepsilon} = e^{-r\varepsilon}$, that

$$\begin{aligned} U(x, \alpha) &= \sum_{n=0}^{\infty} (1 - \lambda_{\varepsilon}) \lambda_{\varepsilon}^{n} u(y_x^{\varepsilon}(n\varepsilon, \alpha), \alpha(n\varepsilon)) \\ &= r \int_0^{\infty} e^{-rt} u(y_x^{\varepsilon}(t, \alpha), \alpha(t)) \mathrm{d}t \leq v(x) \end{aligned}$$

where the last inequality follows from the maximality of the value function. This holds for every piecewise constant strategy $\alpha \in \mathcal{S}_{\varepsilon}^{\#}$. The meaning of this is obvious. The decision maker cannot obtain a higher utility by constraining himself to the smaller set of strategies $\mathcal{S}_{\varepsilon}^{\#}$. Let

$$v^{\varepsilon}(x) := \sup_{\alpha \in \mathcal{S}_{\varepsilon}^{\#}} U(x, \alpha), \tag{OC$_{\varepsilon}$}$$

[3] Note that if $\alpha \in \mathcal{S}^{\#}$ is piecewise continuous it must be piecewise constant on the intervals of continuity.

and let us put the record that, for each $\varepsilon > 0$, we have $v^\varepsilon \leq v$ pointwise. We now establish some simple, but useful, properties of the value function v^ε.

Lemma 2.6. *The dynamic programming problem (OC_ε) has a solution, and the value function v^ε is unique. Moreover, it is uniformly bounded by the constant M_u and Hölder continuous with exponent $\gamma \in \left(0, \min\{\frac{r}{M_b}, 1\}\right)$.*

Proof. The proof of this Lemma is fairly standard, and so we only provide a sketch of the proof. First we show existence and uniqueness of solutions to (OC_ε), using standard arguments. Define the operator T_ε, acting on bounded functions $v : \mathbb{R}^d \to \mathbb{R}$, by

$$T_\varepsilon v(x) = \max_{a \in \mathcal{A}} \{(1 - \lambda_\varepsilon)u(x, a) + \lambda_\varepsilon v(x + \varepsilon b(x, a))\}$$

Since $\lambda_\varepsilon \in (0, 1)$ for each $\varepsilon > 0$, it is easy to see that T_ε defines a contraction mapping on the space of bounded functions on $\mathbb{R}^d$. With the supremum norm this is a Banach space, and the Banach fixed point theorem states that there exists a unique function v^ε such that $T_\varepsilon v^\varepsilon = v^\varepsilon$ pointwise. Standard arguments then show that v^ε is the value function of the restricted problem (OC_ε). The uniform boundedness and Hölder-continuity of the function v^ε follow directly from the proof of Lemma 2.5. □

Next, we construct a deterministic Markov strategy $\underline{a}^\varepsilon : \mathbb{R}^d \to \mathcal{A}$ which solves the problem (OC_ε). For each $x \in \mathbb{R}^d$ let

$$\underline{a}^\varepsilon(x) := \max \{a \in \mathcal{A} | v^\varepsilon(x) = (1 - \lambda_\varepsilon)u(x, a) + \lambda_\varepsilon v^\varepsilon(x + \varepsilon b(x, a))\}. \tag{14}$$

Based on this Markov strategy, we define a piecewise constant strategy in continuous time by setting

$$y_x^\varepsilon(t) = y_x^\varepsilon(n\varepsilon) = x + \varepsilon \sum_{k=0}^{n-1} b(y_x^\varepsilon(k\varepsilon), \underline{a}^\varepsilon) \quad \forall t \in [n\varepsilon, (n+1)\varepsilon), n \geq 0,$$

and, for fixed initial condition $x \in \mathbb{R}^d$,

$$\alpha^\varepsilon(t) := \underline{a}^\varepsilon(y_x^\varepsilon(t)) \quad \forall t \geq 0. \tag{15}$$

It follows that

$$v^\varepsilon(x) = \int_0^\infty re^{-rt} u(y_x^\varepsilon(t), \alpha^\varepsilon(t)) \mathrm{d}t = U(x, \alpha^\varepsilon).$$

This can be shown using the well-known one shot-deviation principle of discrete dynamic programming. It remains to check the consistency of the approximation procedure as $\varepsilon \to 0^+$.

Lemma 2.7. *$v^\varepsilon \to v$ as $\varepsilon \to 0^+$, where v is the unique viscosity solution to (HJB).*

Proof. For each $\varepsilon > 0$ the value function v^ε is uniformly bounded and Hölder continuous. By the Arzelà-Ascoli theorem we can assume that there exists a subsequence $\{v^{\varepsilon_j}\}_{j\in\mathbb{N}}$ such that $\varepsilon_j \to 0^+$ as $j \to \infty$, and along which $v^{\varepsilon_j} \to v$ locally uniformly on $\mathbb{R}^d$. To complete the proof, we will show that v is a viscosity solution of (HJB). This is done by showing that v is simultaneously a viscosity sub and supersolution of (HJB). Let $\phi \in \mathcal{C}^1(\mathbb{R}^d : \mathbb{R})$ be a given map. The bounded and continuous function $v \in \mathcal{C}_b(\mathbb{R}^d : \mathbb{R})$ is a viscosity subsolution of (HJB) if, whenever the function $v - \phi$ has a local maximum at a point x, then

$$rv(x) - H(x, \nabla\phi(x)) \leq 0. \tag{16}$$

$v \in \mathcal{C}_b(\mathbb{R}^d : \mathbb{R})$ is a viscosity supersolution of (HJB) if, whenever the function $v - \phi$ has a local minimum at a point x, then

$$rv(x) - H(x, \nabla\phi(x)) \geq 0. \tag{17}$$

Note that v in this characterization need not be differentiable in any sense. We now come to the verification. Take $\phi \in \mathcal{C}^1(\mathbb{R}^d : \mathbb{R})$ and $x_0 \in \mathbb{R}^d$ a local maximum point for $v - \phi$. Then there exists a closed ball $\mathbb{B}$ centered at x_0 such that

$$(v - \phi)(x_0) \geq (v - \phi)(x) \quad \forall x \in \mathbb{B}. \tag{18}$$

For each $j \in \mathbb{N}$ pick $x_0^j \in \arg\max_{x\in\mathbb{B}}(v^{\varepsilon_j} - \phi)(x)$. By the continuity of the value function v^{ε_j} and the local uniform convergence to v it follows that $x_0^j \to x_0$. Then, for j sufficiently large, the boundedness of the drift eq. (6) implies that $x_0^j + \varepsilon^j b(x_0^j, a) \in \mathbb{B}$ for all $a \in \mathcal{A}$. Therefore, eq. (18) implies that

$$v^{\varepsilon_j}(x_0^j + \varepsilon^j b(x_0^j, a)) - v^{\varepsilon_j}(x_0^j) \leq \phi(x_0^j + \varepsilon^j b(x_0^j, a)) - \phi(x_0^j) \qquad \forall a \in \mathcal{A}. \tag{19}$$

The discrete dynamic programming equation corresponding to problem (OC_ε) states that

$$0 = \max_{a\in\mathcal{A}} \left\{(1 - \lambda_{\varepsilon^j})u(x_0^j, a) + \lambda_{\varepsilon^j} v^{\varepsilon_j}(x_0^j + \varepsilon^j b(x_0^j, a)) - v^{\varepsilon_j}(x_0^j)\right\}$$

for every $j \in \mathbb{N}$. This, together with eq. (19), implies that

$$\begin{aligned} 0 &= \max_{a\in\mathcal{A}} \left\{(1-\lambda_{\varepsilon^j})[u(x_0^j,a) - v^{\varepsilon_j}(x_0^j + \varepsilon^j b(x_0^j,a))] + v^{\varepsilon_j}(x_0^j + \varepsilon^j b(x_0^j,a)) - v^{\varepsilon_j}(x_0^j)\right\} \\ &\leq \left\{(1-\lambda_{\varepsilon^j})[u(x_0^j,a) - v^{\varepsilon_j}(x_0^j + \varepsilon^j b(x_0^j,a))] + \phi^{\varepsilon_j}(x_0^j + \varepsilon^j b(x_0^j,a)) - \phi^{\varepsilon_j}(x_0^j)\right\}. \end{aligned}$$

Since $\phi \in \mathcal{C}^1(\mathbb{R}^d : \mathbb{R})$, the mean-value theorem implies that

$$\phi^{\varepsilon_j}(x_0^j + \varepsilon^j b(x_0^j,a)) - \phi^{\varepsilon_j}(x_0^j) = \varepsilon^j \langle \nabla\phi(x_0^j + \theta^j \varepsilon^j b(x_0^j,a)), b(x_0^j,a)\rangle$$

for every $j \in \mathbb{N}$ and some $\theta^j \in [0,1]$. Hence,

$$0 \leq \max_{a\in\mathcal{A}} \left\{(1-\lambda_{\varepsilon^j})[u(x_0^j,a) - v^{\varepsilon_j}(x_0^j + \varepsilon^j b(x_0^j,a))] + \varepsilon^j \langle \nabla\phi(x_0^j + \theta^j \varepsilon^j b(x_0^j,a)), b(x_0^j,a)\rangle\right\}$$

Dividing by ε^j and observing that

$$\frac{1}{\varepsilon^j}(1-\lambda_{\varepsilon^j}) = \frac{1}{\varepsilon^j}(1-e^{-r\varepsilon^j}) \to r$$

as $j \to \infty$, we conclude that

$$0 \leq -rv(x_0) + H(x, \nabla\phi(x_0)) \Leftrightarrow rv(x_0) - H(x_0, \nabla\phi(x_0)) \leq 0.$$

This shows that v satisfies the viscosity subsolution condition (16). The proof that v it also satisfies the viscosity supersolution condition (17) is done, mutatis mutandis, in the same way and omitted. □

Proposition 2.8. *The sequence of strategies $\{\alpha^\varepsilon\}_{\varepsilon\in(0,1)}$ is a maximizing sequence:*

$$U(x,\alpha^\varepsilon) \to \sup_{\alpha\in\mathcal{S}^\#} U(x,\alpha) = v(x).$$

as $\varepsilon \to 0^+$.

Proof. For each $\varepsilon > 0$ we know that $v^\varepsilon(x) = U(x,\alpha^\varepsilon)$. By the arguments of the previous Lemma, the value function v^ε converges locally uniformly to the viscosity solution v. By uniqueness of solutions in the present case, it follows that v is the value function of the optimal control problem (OC). □

By means of this proposition the strategies α^ε guarantee the decision maker a suboptimal payoff which approximates the maximal payoff when ε is only chosen sufficiently

small. In particular, for every $\delta > 0$ there exists a $\varepsilon_\delta > 0$ such that

$$U(x, \alpha^\varepsilon) \geq v(x) - \delta \quad \forall \varepsilon \in (0, \varepsilon_\delta).$$

3. The Main result

Having described in detail the Markov decision process, and its limit problem, we come now to the main result of this paper.

Theorem 3.1. *Under assumptions 2.1-2.4 we have $V^\varepsilon \to v$ as $\varepsilon \to 0$.*

The proof of this theorem is based on weak convergence arguments as used in Kushner and Dupuis (2001) and Dupuis and Ellis (1997). The main steps of the proof are as follows. First we define continuous-time interpolations of the controlled Markov chain and the action process which will provide the approximation of the controlled pairs for the limit problem. Consider the step-functions

$$\hat{X}^\varepsilon(t) = X_n, \ \hat{A}^\varepsilon(t) = A_n^\varepsilon \quad \forall t \in [n\varepsilon, (n+1)\varepsilon), n \in \mathbb{N}_0. \tag{20}$$

$\hat{X}^\varepsilon$ is a random element of the space of right-continuous functions with left limits, denoted by $\mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d)$, and $\hat{A}^\varepsilon$ is a random element of $\mathcal{D}(\mathbb{R}_+ : \mathcal{A})$. Both these spaces are complete separable metric spaces, when endowed with the Skorohod metric (see e.g. Billingsley, 1999). In terms of these step functions the utility to the decision maker under the strategy σ is given by

$$\begin{aligned} U^\varepsilon(x, \sigma) &= E_x^\sigma \left[\int_0^\infty r e^{-rt} u(\hat{X}^\varepsilon(t), \hat{A}^\varepsilon(t)) \mathrm{d}t \right] \\ &= E_x^\sigma \left[\sum_{n=0}^\infty (1 - \lambda_\varepsilon) \lambda_\varepsilon^n u(\hat{X}^\varepsilon(n\varepsilon), \hat{A}^\varepsilon(n\varepsilon)) \right]. \end{aligned}$$

Hence, by just transporting the controlled Markov chain and the action process to their respective function spaces does not change the value the decision maker can achieve. In Section 6.1 we show that the sequence of interpolated processes $\{(\hat{X}^\varepsilon(t), \hat{A}^\varepsilon(t)), t \geq 0\}$ are tight in their respective function spaces and suitable topologies. By the Prohorov theorem this guarantees that every sequence has a convergent subsequence. Using a suitable representation of the action process in terms of mixed actions (this will be made precise in section 6), this sequential compactness result allows us to prove that there exists a well defined limit process $(\bar{X}, \nu)$, where $\bar{X}$ is a stochastic process taking values in the space of

continuous functions $\mathcal{C}(\mathbb{R}^d : \mathbb{R})$ and ν is a stochastic process taking values in $\mathcal{S}$.[4] The two are coupled by the stochastic integral equation

$$\bar{X}(t) = x + \int_0^t b(\bar{X}(s), \nu(s))\mathrm{d}s. \tag{21}$$

For every element of the probability space variable, the pair $(\bar{X}(\omega), \nu(\omega))$ defines an admissible control pair for the deterministic optimal control problem (OC). Consequently, the strategy $\nu(\omega)$ is an element of the set $\mathcal{S}$, and therefore cannot give the decision maker a larger utility as he could obtain by solving the deterministic problem directly. This forms the basis for the proof that $\limsup_{\varepsilon\to 0} V^\varepsilon(x) \leq v(x)$. To show equality of the value functions, we need to show that also $\liminf_{\varepsilon\to 0} V^\varepsilon(x) \geq v(x)$. This will be shown by adapting the deterministic piecewise constant strategy α^ε constructed in eq. (15), and using this strategy as a strategy for the Markov decision process. The details of all these arguments are provided in Section 6.

4. Conclusion

We have focused in this paper on a very standard stochastic optimal control problem, and studied its convergence to a deterministic continuous-time problem. The key assumption which allowed us to prove this deterministic limit result was the "asymptotically vanishing" variance of the increments of the state process. Without this assumption a diffusion, or even a jump-diffusion approximation would be more appropriate. Second, we have focused in this paper on the theoretically important case in which the decision maker has only finitely many actions among which he can choose. There are no problems in allowing the set of actions to be a convex compact subset of $\mathbb{R}^m$. The arguments of this paper go through without any change at all, and in fact turn out to be simpler, as under this assumption, paired with the continuity hypothesis on the utility flow function u, it is well-known that the decision maker can use a deterministic Markov strategy which as optimal strategy in the Markov decision process (see e.g. Puterman, 1994). A more challenging question, and one which actually motivated me to look at this problem, is to extend the current result to stochastic games with imperfect public monitoring. In this extended setting the state process $\{X_n^\varepsilon\}_{n\in\mathbb{N}_0}$ is interpreted as the public signal process the players can observe. Strategies as defined in this paper, which are contingency plans conditioning only on the realizations of the signal process, are called public strategies.

[4] In the control-theoretic literature this relaxation procedure is standard since the classical works of Warga (1972). See section 6.1 for the precise definition of the relaxed representation of the action process.

The deterministic limit case is then only one of many scenarios one could study, and in fact might not be the most interesting one. A challenging problem is to prove a limit theorem where the limit signal process evolves according to a jump diffusion process as in the recent paper by Sannikov and Skrypacz (2010). We leave this problem for future research.

5. Proofs

Let $\{(X_n^\varepsilon, A_n^\varepsilon)\}_{n\in\mathbb{N}_0}$ be the Markov decision process. For each $\varepsilon > 0$ the law of the action process is described by a feedback strategy σ^ε, being a stochastic kernel on $\mathcal{A}$ given $\mathbb{R}^d$. In the following we assume that the initial condition of the state process is the fixed point $x \in \mathcal{X} \subset \mathbb{R}^d$. There are no problems in making the alternative assumption that the initial condition is drawn from a distribution ρ supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$. The pair process $\{(X_n^\varepsilon, A_n^\varepsilon)\}_{n\in\mathbb{N}_0}$ induces the law $P_x^{\sigma^\varepsilon} \equiv P^\varepsilon$ defined on the measure space $(\Omega, \mathcal{F})$, where $\Omega := (\mathcal{X} \times \mathcal{A})^{\mathbb{N}_0}$ and $\mathcal{F}$ the sigma-field generated by the finite cylinder sets. Weak convergence arguments are used in the sequel, investigating the limit behavior of the sequence of laws $\{P^\varepsilon\}_{\varepsilon\in(0,1)}$. Recall that a sequence of probability measures $\{P^\varepsilon\}$ converges weakly to a limit measure P if

$$\lim_{\varepsilon\to 0} \int_\Omega f(\omega) \mathrm{d}P^\varepsilon(\omega) = \int_\Omega f(\omega) \mathrm{d}P(\omega)$$

for all bounded continuous random variables $f : \Omega \to \mathbb{R}$. We will use this notion of convergence to speak about limits of the continuous-time process $\{(\hat{X}^\varepsilon(t), \hat{A}^\varepsilon(t)\}_{t\geq 0}$, defined in (20). Once we have settled the convergence issue, we will be able to determine the limit of the value function $\{V^\varepsilon\}_{\varepsilon\in(0,1)}$.

5.1 Convergence of interpolated processes

By definition, we have

$$X_n^\varepsilon(\omega) = X_0^\varepsilon(\omega) + \varepsilon \sum_{k=0}^{n-1} f_{k+1}^\varepsilon(X_k^\varepsilon(\omega), A_k^\varepsilon(\omega)).$$

Call $Z_{n+1}^\varepsilon = f_{n+1}^\varepsilon(X_n^\varepsilon, A_n^\varepsilon)$ the random (normalized) increment of the state process in stage n of the algorithm, and $\hat{Z}^\varepsilon(t) = Z_{n+1}^\varepsilon$ for $t \in [n\varepsilon, (n+1)\varepsilon)$ its corresponding step process. Using the step processes $\hat{X}^\varepsilon$ and $\hat{A}^\varepsilon$ introduced in eq. (20), we can write the above recursive

relation as an integral equation

$$\hat{X}^\varepsilon(t,\omega) = \hat{X}^\varepsilon(0,\omega) + \int_0^{n\varepsilon} \hat{Z}^\varepsilon(s,\omega)\mathrm{d}s.$$

Introducing the random variable

$$M_n^\varepsilon = \varepsilon\left(Z_{n+1}^\varepsilon - b^\varepsilon(X_n^\varepsilon, A_n^\varepsilon)\right),$$

we obtain the representation

$$\hat{X}^\varepsilon(t,\omega) = \hat{X}^\varepsilon(0,\omega) + \int_0^{n\varepsilon} b^\varepsilon(\hat{X}^\varepsilon(s,\omega), \hat{A}^\varepsilon(s,\omega))\mathrm{d}s + \sum_{k=0}^{n} M_k^\varepsilon(\omega).$$

Given the definition of the function b^ε, the following Lemma is very simple.

Lemma 5.1. *The process* $\{\sum_{k=0}^n M_k^\varepsilon\}_{n\in\mathbb{N}_0}$ *is a martingale with respect to the filtration* $\mathcal{G}_n^\varepsilon = \sigma\left(X_0^\varepsilon, A_0^\varepsilon, \ldots, X_n^\varepsilon, A_n^\varepsilon\right)$.

It follows that $\{\|\sum_{k=0}^n M_k^\varepsilon\|^2\}_{n\in\mathbb{N}_0}$ is a submartingale with respect to $\mathcal{G}_n^\varepsilon$. This translates in a straightforward way to the continuous-time submartingale $t \mapsto \|\hat{M}^\varepsilon(t)\|^2$, where

$$\hat{M}^\varepsilon(t) = \int_0^{n\varepsilon} (\hat{Z}^\varepsilon(s) - b^\varepsilon(\hat{X}^\varepsilon(s), \hat{A}^\varepsilon(s))\mathrm{d}s \qquad \forall t \in [n\varepsilon, (n+1)\varepsilon).$$

An application of the submartingale inequality (Theorem 3.8 in Karatzas and Schreve, 2000) gives the bound

$$P_x^\sigma\left[\sup_{0\le t\le T} \|\hat{M}^\varepsilon(t)\|^2 \ge \lambda\right] \le \frac{1}{\lambda} E_x^\sigma \|\hat{M}^\varepsilon(T)\|^2$$

for every $\lambda > 0$ and $T < \infty$.

Using assumptions 2.4 and 2.2, the expectation on the right-hand side of this inequality is on the order of $o(\varepsilon)$. Therefore

$$\lim_{\varepsilon\to 0} P_x^\sigma\left[\sup_{0\le t\le T} \|\hat{M}^\varepsilon(t)\|^2 \ge \lambda\right] = 0 \tag{22}$$

for every strategy σ and initial state $x \in \mathcal{X}$.

The pair $(\hat{X}^\varepsilon, \hat{A}^\varepsilon)$ can be thought of as mappings from Ω to the space of cadlag functions with image in $\mathbb{R}^d \times \mathcal{A}$, denoted by $\mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d \times \mathcal{A})$. The problem with this space is that

it does not have the useful compactness properties to speak about convergence of the action process. The action process $\{\hat{A}^\varepsilon(t,\omega), t \geq 0\}$ is for each $\omega \in \Omega$ a deterministic right-continuous step function taking values in the discrete set $\mathcal{A}$. Given its discrete nature we cannot talk about function convergence in an ordinary sense. To speak about convergence of this process we interpret the pure action $\hat{A}^\varepsilon(t)$ as a behavior strategy taking values in the simplex $\Delta(\mathcal{A})$. To achieve this, we define the mixed action process by

$$\nu_a^\varepsilon(t,\omega) = \delta_{\hat{A}^\varepsilon(t,\omega)}(a) := \begin{cases} 1 & \text{if } \hat{A}^\varepsilon(t,\omega) = a, \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

Clearly the random variable $\nu^\varepsilon(t,\omega)$ is an element of the mixed action simplex $\Delta(\mathcal{A})$ and the map $t \mapsto \nu^\varepsilon(t,\omega)$ is an element of the space of open-loop controls for the limit problem $\mathcal{S} = \{\alpha : \mathbb{R}_+ \to \Delta(\mathcal{A}) | \alpha(\cdot) \text{ measurable}\}$ for each fixed $\omega \in \Omega$.[5] Denote by $\mathcal{S}|_{[0,T]}$ the subspace of open loop controls restricted to the domain $[0,T]$. The usefulness of introducing this abstract concept comes from the following technical fact. Say that a sequence $\{\nu^j\}_{j\in\mathbb{N}} \subset \mathcal{S}|_{[0,T]}$ converges weak* to a limit $\nu \in \mathcal{S}|_{[0,T]}$ if for every integrable function $f : \mathcal{A} \times [0,T] \to \mathbb{R}$ we have

$$\lim_{j\to\infty} \int_0^T \sum_{a\in\mathcal{A}} f(a,t)\nu_a^j(t)\mathrm{d}t = \lim_{j\to\infty} \int_0^T \sum_{a\in\mathcal{A}} f(a,t)\nu_a(t)\mathrm{d}t$$

The following result follows from general functional analytic facts (essentially Alaoglu's theorem).

Lemma 5.2. *The set $\mathcal{S}|_{[0,T]}$ is sequentially compact in the weak* topology, i.e. every sequence $\{\alpha^j\} \subset \mathcal{S}|_{[0,T]}$ has a weak* convergent subsequence with limit in $\mathcal{S}|_{[0,T]}$.*

Proof. See e.g. Lemma 5.1. in Capuzzo-Dolcetta and Ishii (1984). □

Defining a topology on $\mathcal{S}$ by saying that a sequence of open-loop controls $\{\alpha^j\}$ converges weak* to a limit α if and only if each restriction $\alpha^j|_{[0,T]}$ converges to the restriction $\alpha|_{[0,T]}$ shows that $\mathcal{S}$ is a weak* compact subset of $L^\infty(\mathbb{R}_+, \Delta(\mathcal{A}))$. To summarize, for every $\omega \in \Omega$ and every sequence of relaxed controls $\{\nu^{\varepsilon_j}(\omega)\}_{j\in\mathbb{N}}$ with $\varepsilon_j \to 0$ as $j \to \infty$, there exists a weak* converging subsequence with limit $\nu(\omega) \in \mathcal{S}$. Therefore, we can state the following technical fact.

Lemma 5.3. *The family of open-loop strategies $\{\nu^\varepsilon\}_{\varepsilon\in(0,1)}$ is sequentially compact in $\mathcal{S}$ with respect to the weak* convergence. Therefore, for every $\omega \in \Omega$ there exists a sequence $\varepsilon_j(\omega) \to 0$ as $j \to$ for*

[5] $t \mapsto \nu^\varepsilon(t,\omega)$ is a step function, thus trivially measurable.

which $\nu^{\varepsilon_j(\omega)}(\omega) \to \nu(\omega)$ in the weak sense. The so obtained limit process is denoted by ν and is a random element of the space of open loop strategies $\mathcal{S}$.*

We now finalize the proof of the convergence of the interpolated sample paths by showing that the family of $\mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d)$ valued random variables $\{\hat{X}^\varepsilon(t); t \geq 0\}_{\varepsilon\in(0,1)}$ is relatively compact with limit in $\mathcal{C}(\mathbb{R}_+ : \mathbb{R}^d)$.

Lemma 5.4. *The family of $\mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d)$ valued processes $\{\hat{X}^\varepsilon\}_{\varepsilon\in(0,1)}$ is relatively compact, i.e. for every subsequence there exists a subsubsequence such that $\hat{X}^\varepsilon_j \Rightarrow \bar{X}$ in distribution.*

Proof. For $\mathbf{x} \in \mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d)$ define the modulus of continuity by

$$w(\mathbf{x}, \delta, T) = \inf_{\{t_i\}} \max_{1\leq i\leq n} \sup_{s,t\in[t_{i-1},t_i)} \|\mathbf{x}(s) - \mathbf{x}(t)\|_\infty$$

where the sequence $\{t_i\}$ ranges of all partitions of the form $0 = t_0 < t_1 < \ldots < t_{n-1} < T \leq t_n$ with $\min_{1\leq i\leq n}(t_i - t_{i-1}) > \delta$. For every fixed $\varepsilon > 0$ pick $\delta = \frac{\varepsilon}{2}$. Then the sequence $t_i = i\varepsilon, i = 0, 1, \ldots, \lceil T/\varepsilon \rceil$ is admissible, and we see that

$$\max_i \max_{s,t\in[(i-1)\varepsilon,i\varepsilon)} \|\hat{X}^\varepsilon(t) - \hat{X}^\varepsilon(s)\|_\infty = \max_{1\leq i\leq n} \|\varepsilon f^\varepsilon_i(X^\varepsilon_{i-1}, A^\varepsilon_{i-1})\|_\infty.$$

By Assumption 2.2, the random vector fields $\{f^\varepsilon_n\}$ take values in the compact set $\mathcal{K}$ and can therefore be uniformly embedded in a compact cube $\Gamma \subset \mathbb{R}^d$. It follows that for every $\omega \in \Omega, \varepsilon > 0$ and $T > 0$ we have

$$\lim_{\delta\to 0} w(\hat{X}^\varepsilon(\omega), \delta, T) = 0.$$

Using assumption 2.2 once again, we see that for every $T > 0$ the sample paths of the step process $\hat{X}^\varepsilon$ are contained in a compact cube $\Gamma_T \subset \mathbb{R}^d$ with probability 1. Theorem 7.2 in Ethier and Kurtz (1986) states that under these two conditions the family of processes $\{\hat{X}^\varepsilon\}_{\varepsilon\in(0,1)}$ is relatively compact in $\mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d)$. □

By now we have shown that the random pair $(\hat{X}^\varepsilon, \nu^\varepsilon) : \Omega \to \mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d \times \mathcal{A})$ converges in law to a pair process $(\hat{X}, \nu)$. In terms of the induced probability measures this has the following meaning. Let $\hat{P}^\varepsilon = P^\varepsilon_x \circ (\hat{X}^\varepsilon, \nu^\varepsilon)^{-1}$ be the induced law of the process $(\hat{X}^\varepsilon, \nu^\varepsilon)$ on $\mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d \times \mathcal{A})$, with marginals $\hat{P}^\varepsilon_1$ on $\mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d)$ and $\hat{P}^\varepsilon_2$ on $\mathcal{D}(\mathbb{R}_+ : \Delta(\mathcal{A})) \subset \mathcal{S}$, respectively.[6] We now proceed to show that the limit state process $\bar{X}$ has almost surely

[6]Note that the space $\mathcal{D}(\mathbb{R}_+ : \Delta(A))$ is very similar to the space $\mathcal{S}^\#_\varepsilon$, defined in eq. (13). However, now we allow the step function to take values at any point on the simplex $\Delta(\mathcal{A})$.

continuous sample paths.
For every $\mathbf{x} \in \mathcal{D}(\mathbb{R}_+ : \mathbb{R}^d)$ define

$$J(\mathbf{x}) := \int_0^\infty e^{-s} \min\{J(\mathbf{x}, s), 1\} \mathrm{d}s$$

with

$$J(\mathbf{x}, s) := \sup_{0 \le t \le s} \|\mathbf{x}(t) - \mathbf{x}(t-)\|_\infty$$

and $\mathbf{x}(t-) \equiv \lim_{\tau \to t^-} \mathbf{x}(\tau)$. Then, for every $s > 0$, it follows that

$$J(\hat{X}^\varepsilon, s) \le \varepsilon \sup_{k \in \mathcal{K}} \|k\|_\infty$$

and therefore

$$J(\hat{X}^\varepsilon(\omega)) \le \varepsilon \sup_{k \in \mathcal{K}} \|k\|_\infty \to 0 \text{ for } \varepsilon \to 0$$

for every $\omega \in \Omega$. Passing to a subsequence, we can assume that $\hat{X}^\varepsilon \Rightarrow \bar{X}$ in distribution. Theorem 10.2 in Ethier and Kurtz (1986) implies that $\bar{X}$ is almost surely a random process in $\mathcal{C}(\mathbb{R}_+ : \mathbb{R}^d)$. To characterize this process, define the process

$$\begin{aligned} \bar{X}^\varepsilon(t, \omega) &= \hat{X}^\varepsilon(0, \omega) + \int_0^t b^\varepsilon(\hat{X}^\varepsilon(s, \omega), \hat{A}^\varepsilon(s, \omega)) \mathrm{d}s \\ &= \hat{X}^\varepsilon(0, \omega) + \int_0^t b^\varepsilon(\hat{X}^\varepsilon(s, \omega), \nu^\varepsilon(s, \omega)) \mathrm{d}s. \end{aligned}$$

Here we have extended the domain of the drift b to $\mathbb{R}^d \times \Delta(\mathcal{A})$ in the obvious way. Passing to subsequences if necessary, we can assume that $(\hat{X}^\varepsilon, \nu^\varepsilon) \Rightarrow (\bar{X}, \nu)$ in distribution. By assumption 2.1 the drift converges locally uniformly to a Lipschitz continuous function b. Together with the continuous mapping theorem (Ethier and Kurtz, 1986, p.103), this implies that

$$\lim_{\varepsilon \to 0} \int_0^t b^\varepsilon(\hat{X}^\varepsilon(s), \nu^\varepsilon(s)) \mathrm{d}s = \int_0^t b(\bar{X}(s), \nu(s)) \mathrm{d}s$$

for every $t > 0$ and in distribution. Since $\hat{X}^\varepsilon(0, \omega) = x$ with probability 1, we obtain $\bar{X}^\varepsilon \Rightarrow \bar{X}$, with

$$\bar{X}(t) = x + \int_0^t b(\bar{X}(s), \nu(s)) \mathrm{d}s. \tag{24}$$

This completes the proof of the convergence of sample paths of the controlled Markov chain.

$$\hat{P}_1^\varepsilon \to \hat{P}_1,\ \hat{P}_2^\varepsilon \to \hat{P}_2 \in \mathsf{M}_1^+(\mathcal{S})$$

5.2 Convergence of Values

To complete the proof of the main result we show that $V^\varepsilon \to V$ for a compact set of initial conditions $\mathcal{X} \subset \mathbb{R}^d$. Let $\{\sigma^\varepsilon\}_{\varepsilon\in(0,1)}$ be the sequence of optimal Markov strategies for the decision maker, so that for each $\varepsilon \in (0,1)$ we have

$$V^\varepsilon(x) = E_x^\varepsilon\left[\int_0^\infty re^{-rt}u(\hat{X}^\varepsilon(t), \hat{A}^\varepsilon(t))\mathrm{d}t\right]$$

Passing to a subsequence we may assume that $(\hat{X}^\varepsilon, \nu^\varepsilon) \Rightarrow (\bar{X}, \nu)$ in distribution. By the Skorohod representation theorem (Billingsley, 1999) there exists a probability space $(\bar{\Omega}, \bar{\mathcal{G}}, P)$ on which we can define random variables $(Y^\varepsilon, \rho^\varepsilon)$, with joint law $\hat{P}^\varepsilon$, and which converge almost surely to the processes $(\bar{X}, \nu)$. Using this abstract results, we will not distinguish between the random elements $(\hat{X}^\varepsilon, \nu^\varepsilon)$ and $(Y^\varepsilon, \rho^\varepsilon)$, as they describe the same processes in distribution. Then we actually have convergence of the random processes except on a set of $P-$measure 0, denote by N. This construction allows us to use the continuous mapping theorem as follows. Define the continuous function $g : \mathcal{D}(\mathbb{R}_+, \mathbb{R}^d) \times \mathcal{S} \to \mathbb{R}$ by

$$g(\phi, \alpha) := \int_0^\infty re^{-rt}u(\phi(t), \alpha(t))\mathrm{d}t.$$

Then, for each $\omega \in \bar{\Omega}$ the number $g(\hat{X}^\varepsilon(\omega), \nu^\varepsilon(\omega))$ is the payoff of the decision maker under the control pair $(\hat{X}^\varepsilon, \nu^\varepsilon)$. Since u is continuous, it follows from the continuous mapping theorem (Billingsley, 1999) that, for each $\bar{\omega} \in \bar{\Omega} \setminus N$

$$\begin{aligned}
\lim_{\varepsilon\to 0} g(\hat{X}^\varepsilon(\omega), \nu^\varepsilon(\omega)) &= g(\bar{X}(\omega), \nu(\omega)) \\
&= \int_0^\infty re^{-rt}u(\bar{X}(t,\omega), \nu(t,\omega))\mathrm{d}t \\
&= U(x, \nu(\omega)) \\
&\leq v(x).
\end{aligned}$$

Applying the dominated convergence theorem, it follows that

$$\lim_{\varepsilon\to 0}\int_{\bar\Omega} g(\hat X^\varepsilon(\omega),\nu^\varepsilon(\omega))\mathrm{d}P(\omega)=\int_{\bar\Omega} g(\bar X(\omega),\nu(\omega))\mathrm{d}P(\omega). \tag{25}$$

Let E denote expectation on the probability space $(\bar\Omega,\bar{\mathcal{F}},P)$ provided by the Skorohod representation. Then, eq. (25) implies that

$$\begin{aligned}\limsup_{\varepsilon\to 0} V^\varepsilon(x) &= \limsup_{\varepsilon\to 0} E_x^\varepsilon\left[\int_0^\infty re^{-rt}u(\hat X^\varepsilon(t),\nu^\varepsilon(t))\mathrm{d}t\right]\\ &= \mathsf{E}_x\left[\int_0^\infty re^{-rt}u(\bar X(t),\nu(t))\mathrm{d}t\right]\\ &\le v(x).\end{aligned}$$

To finish the proof of Theorem 3.1, we need to show that also

$$\liminf_{\varepsilon\to 0^+} V^\varepsilon(x)\ge v(x).$$

The proof of this assertion is rather straightforward, thanks to the explicit approximation procedure described in section 2.2. For each $\varepsilon>0$ let α^ε denote the piecewise constant control, taking values in the pure action set $\mathcal{A}$, constructed in eq. (15). From Proposition 2.8 we know that for every $\delta>0$ there exists a $\varepsilon_\delta>0$ sufficiently small so that

$$U(x,\alpha^\varepsilon)\ge v(x)-\delta\quad\forall\varepsilon\in(0,\varepsilon_\delta).$$

We adapt this strategy for the controlled Markov chain as follows. For each $n\in\mathbb{N}_0$ we define a deterministic action process $A_n^\varepsilon:=\alpha^\varepsilon(n\varepsilon)$, without any explicit reference to the current value of the state process. Hence, independent of the probability space variable ω, we always implement the same action process $\{A_n^\varepsilon\}_{n\in\mathbb{N}}$. This defines an admissible strategy for the decision maker which gives him a payoff of[7]

$$E_x\left[\sum_{n=0}^\infty(1-\lambda_\varepsilon)\lambda_\varepsilon^n u(X_n^\varepsilon,A_n^\varepsilon)\right]\le V^\varepsilon(x^\varepsilon).$$

With a slight abuse of notation, denote the left-hand side of this equation by $U^\varepsilon(x,\alpha^\varepsilon)$. Set $\hat X^\varepsilon(t)=X_n^\varepsilon$ and $\nu^\varepsilon(t)=\delta_{A_n^\varepsilon}$ for each $t\in[n\varepsilon,(n+1)\varepsilon)$. Then, it follows from the sequential compactness of relaxed controls (Lemma 6.3) that, passing if necessary to a subsequence,

[7] We omit to index the probability measure and its expectation by the strategy, as it is in one-to-one correspondence with the deterministic action process in this case.

the deterministic limit

$$\lim_{\varepsilon\to 0} \nu^\varepsilon = \lim_{\varepsilon\to 0} \delta_{\alpha^\varepsilon(\cdot)} = \nu$$

exists and defines an open-loop control in $\mathcal{S}$ (see also Capuzzo-Dolcetta and Ishii, 1984, for a related argument). Along the same subsequence, it follows from arguments used in section 6.1 that $\hat{X}^\varepsilon \Rightarrow \bar{X}$ in distribution, where

$$\bar{X}(t) = x + \int_0^t b(\bar{X}(s), \nu(s))\mathrm{d}s.$$

Since the strategy used in this integral equation is deterministic and the initial condition is fixed, this equation has a unique deterministic solution. Therefore $\bar{X}$ is a deterministic process which, by uniqueness, corresponds to the limit process of the controlled pair $(y_x^\varepsilon, \alpha^\varepsilon)$. Therefore, Proposition 2.8 implies that

$$\liminf_{\varepsilon\to 0} U^\varepsilon(x, \alpha^\varepsilon) = \int_0^\infty re^{-rt} u(\bar{X}(t), \nu(t))\mathrm{d}t = v(x).$$

As $V^\varepsilon(x) \geq U^\varepsilon(x, \alpha^\varepsilon)$ for every ε, we conclude that

$$\liminf_{\varepsilon\to 0} V^\varepsilon(x) \geq \liminf_{\varepsilon\to 0} U^\varepsilon(x, \alpha^\varepsilon) = v(x).$$

This completes the proof of Theorem 3.1.

References

Bardi, M. and Capuzzo-Dolcetta, I. (1997). *Optimal control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Birkhäuser - Systems & Control: Foundations & Applications.

Benaïm, M. (1998). Recursive algorithms, urn processes, and the chaining number of chain recurrent sets. *Ergodic Theory and Dynamical Systems*, 18:53–87.

Bertsekas, D. and Shreve, S. E. (1978). *Stochastic optimal control - The discrete time case*. Academic Press.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley Series in Probability and Statistics, second edition.

Capuzzo-Dolcetta, I. (1983). On a discrete approximation of the hamilton-jacobi-bellman equation of dynamic programming. *Applied Mathematics & Optimization*, 10:367–377.

Capuzzo-Dolcetta, I. and Ishii, H. (1984). Approximate solutions of the bellman equation of deterministic control theory. *Applied Mathematics & Optimization*, 11:161–181.

Dupuis, P. and Ellis, R. S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley Series in Probability and Statistics.

Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.

Falcone, M. (1987). A numerical approach to the infinite horizon problem of deterministic control theory. *Applied Mathematics & Optimization*, 15:1–13.

Gast, N., Gaujal, B., and Le Boudec, J.-Y. (2012). Mean field for markov decision processes: From discrete to continuous optimization. *IEEE Transactions on Automatic Control*, 57(9):2266 –2280.

Gonzales, R. and Rofman, E. (1985). On deterministic control problems: An approximation procedure for the optimal cost i. the stationary problem. *SIAM Journal on Control and Optimization*, 23(2):242–266.

Hörner, J., Sugaya, T., Takahashi, S., and Vieille, N. (2011). Recursive methods in discounted stochastic games: An algorithm for $\delta \to 1$ and a folk theorem. *Econometrica*, 79:1277–1318.

Karatzas, I. and Schreve, S. E. (2000). *Brownian Motion and Stochastic Calculus*. Springer-Verlag, 2nd edition.

Kushner, H. J. and Dupuis, P. (2001). *Numerical Methods for Stochastic control problems in continuous time*. Springer, New York, second edition.

Puterman, M. L. (1994). *Markov Decision Processes - Discrete Stochastic Programming*. Wiley-Interscience.

Sannikov, Y. and Skrypacz, A. (2010). The role of information in repeated games with frequent actions. *Econometrica*, Vo. 78, No. 3:847–882.

Warga, J. (1972). *Optimal Control of Differential and Functional Equations*. Academic Press.