ON THE POLICY IMPROVEMENT ALGORITHM IN CONTINUOUS TIME

SAUL D. JACKA AND ALEKSANDAR MIJATOVIĆ

Abstract. We develop a general approach to the Policy Improvement Algorithm (PIA) for stochastic control problems for continuous-time processes. The main results assume only that the controls lie in a compact metric space and give general sufficient conditions for the PIA to be well-defined and converge in continuous time (i.e. without time discretisation). It emerges that the natural context for the PIA in continuous time is weak stochastic control. We give examples of control problems demonstrating the need for the weak formulation as well as diffusion-based classes of problems where the PIA in continuous time is applicable.

1. Introduction

The **Policy Improvement Algorithm (PIA)** has played a central role in control and optimisation for over half a century, see e.g. Howard's monograph [4]. The PIA yields an intuitive constructive approach to optimal control by generating a sequence of policies $(\pi_n)_{n\in\mathbb{N}}$ whose payoffs $(V^{\pi_n})_{n\in\mathbb{N}}$ are improved at every step. Put differently, the payoffs $(V^{\pi_n})_{n\in\mathbb{N}}$ form a sequence of functions on the state space converging monotonically to the value function of the problem (see e.g. Section 3 below for the precise definition). In the stochastic setting, the PIA is perhaps most widely applied in the theory of Markov decision processes, see e.g. [3, 7, 8] and the references therein. Most of the literature on the PIA makes assumptions either on the process (e.g. finite/countable state space or discrete time) or on the set of available controls (e.g. a finite set [1]). In contrast, the present paper presents an abstract approach to the PIA in continuous time, allowing for an uncountable set of controls.

The main aim of this work is two-fold: (1) define a general weak formulation for optimal control problems in continuous time, without restricting the set of available controls, and (2) develop an abstract framework for in this setting for the PIA to work. The latter task involves stating a general set of assumptions (see **(As1)**–**(As8)** in Section 3 below), under which the sequence of policies $(\pi_n)_{n\in\mathbb{N}}$ can be constructed, prove that the PIA yields an increasing sequence of payoffs $(V^{\pi_n})_{n\in\mathbb{N}}$ (see Theorem 1 below), which converges to the value function of the stochastic control problem (see Theorem 2 below), and prove that a subsequence of policies $(\pi_n)_{n\in\mathbb{N}}$ converges uniformly on compacts to an optimal policy π^* with the payoff V^{π^*} equal to the value function (see Theorem 3 below). In particular, our results imply that under general assumptions **(As1)**–**(As8)**, an optimal policy π^* exists.

1991 *Mathematics Subject Classification.* 93E20.

Key words and phrases. Stochastic control in continuous time, policy improvement algorithm, general state space, general controls.

The present paper presents a unified language for stating and solving general stochastic control problems in continuous time, which can in particular be used to describe simultaneously our recent results on the PIA for diffusions over the infinite [5] and finite [6] time horizons. The key distinction between this work and [5, 6] lies in the fact that here we assume that the payoff V^{π_n} is sufficiently regular for every policy π_n produced by the PIA, which appears to be necessary for the algorithm to converge. In contrast, in [5] (resp. [6]) we prove that this assumption is satisfied in the context of control problems for continuous-time diffusion processes over an infinite (resp. finite) time horizon.

The remainder of the paper is organised as follows: Section 2 gives the general weak formulation of the control problem and presents examples demonstrating the necessity of the weak formulation. Section 3 describes the PIA and states our main results. Section 4 presents examples of the PIA in the context of diffusion processes, based on [5, 6]. The proofs of the results are given in Section 5.

2. The general problem: setting and examples

2.1. **Setting.** Consider the following weak formulation of a general *optimal control problem*. Given continuous functions $f : S \times A \to \mathbb{R}_+$ and $g : S \to \mathbb{R}_+$, find for each $x \in S$

$$V(x) := \sup_{\Pi \in \mathcal{A}_x} \mathbb{E}\left[\int_0^\tau f(X_t^\Pi, \Pi_t)dt + g(X_\tau^\Pi)1_{(\tau<\infty)}\right] \tag{2.1}$$

where

(1) the control process Π, defined on some filtered probability space $(\Omega, (\mathcal{F}_t)_{t\in\mathbb{R}_+}, \mathcal{F}, \mathbb{P})$, takes values in a compact metric space A and is $(\mathcal{F}_t)$-adapted. The topological space S is the state space of the controlled process and D is a domain (i.e. an open and connected subset) in S, such that $D = \cup_{n=1}^\infty K_n$, where $\{K_n\}$ are an increasing family of compact sets in S with K_n contained in the interior of K_{n+1} for all $n \in \mathbb{N}$;

(2) for each $a \in A$ we assume that X^a is a strong Markov process with state space S and a given (martingale) infinitesimal generator $\mathcal{L}^a$ and domain $\mathbf{D}^a$. Furthermore, we assume that there exists a nonempty subset $\mathbf{C}$ of $\cap_{a\in A}\mathbf{D}^a$ with the property that the map $(x, a) \mapsto \mathcal{L}^a\phi(x)$ is jointly continuous on $D \times A$ for each $\phi \in \mathbf{C}$;

(3) $\mathcal{A}_x$ consists of all control processes Π such that there exists an $(\mathcal{F}_t)$-adapted, right-continuous S-valued process X^Π satisfying

  (i) $X_0^\Pi = x$;

  (ii) the law of (X^Π, Π) is unique;

  (iii) for each $\phi \in \mathbf{C}$,

$$\phi(X_{t\wedge\tau}^\Pi) - \int_0^{t\wedge\tau} \mathcal{L}^{\Pi_s}\phi(X_s^\Pi)ds \quad \text{is a martingale,} \tag{2.2}$$

  where the stopping time τ is the first exit time of X^Π from D;

  (iv) defining J by

$$J(x, \Pi) := \int_0^\tau f(X_t^\Pi, \Pi_t)dt + g(X_\tau^\Pi)1_{(\tau<\infty)},$$

we have

$$\int_0^{t\wedge\tau} f(X_s^\Pi, \Pi_s)ds + g(X_\tau^\Pi)1_{(\tau<\infty)} \xrightarrow{L^1} J(x,\Pi) \qquad \text{as } t\to\infty.$$

We refer to the elements of $\mathcal{A}_x$ as controls.

Remark 1. The stochastic basis, i.e. the filtered probability space $(\Omega, (\mathcal{F}_t)_{t\in\mathbb{R}_+}, \mathcal{F}, \mathbb{P})$, in the definition of the control process Π in (1) above may depend on Π. In particular, the expectation in (2.1) corresponds to the probability measure $\mathbb{P}$ under which the control Π is defined. In the weak formulation, we are not required to fix a filtered probability space in advance but instead allow the control, together with its corresponding controlled process, to be defined on distinct stochastic bases for different controls.

We now recall the definition of a key class of controls, namely Markov policies. A *Markov policy* π is a function $\pi : S \to A$ such that for each $x \in D$ there exists an adapted process X on a stochastic basis satisfying

(i) $X_0 = x$;
(ii) $\Pi = \pi(X)$, defined by $\Pi_t := \pi(X_t)$ for $t \geq 0$, is in $\mathcal{A}_x$;
(iii) the processes (X^Π, Π) and $(X, \pi(X))$ have the same law.

Hereafter we denote such an X by X^π. Note that (ii) in the definition of a Markov policy implies the existence of the process X^Π and the uniqueness of the law of (X^Π, Π). Part (iii) stipulates that the law of $(X, \pi(X))$ coincides with it.

Remark 2. As mentioned in Remark 1 above, our formulation of the control problem in (2.1) does not make a reference to a particular filtered probability space. This allows us to consider the Markov control $\pi = \text{sgn}$, see e.g. examples (I) and (II) in Section 2.2.3 below. It is well known that the SDE in (2.5) (with $a = \text{sgn}(X)$), arising in these examples, does not possess a strong solution, and hence a strong formulation of the stochastic control problem would have to exclude such natural Markov controls. Furthermore, these examples show that such controls arise as the optimal controls in certain problems.

Given $x \in S$ and a policy $\Pi \in \mathcal{A}_x$ (resp. a Markov policy π), we define the *payoff* to be

$$V^\Pi(x) := \mathbb{E}[J(x,\Pi)] \qquad (\text{resp. } V^\pi(x) := \mathbb{E}[J(x,\pi(X^\pi)]). \tag{2.3}$$

Hence the value function V, defined in (2.1), can be expressed in terms of the payoffs V^Π as

$$V(x) := \sup_{\Pi\in\mathcal{A}_x} V^\Pi(x) \qquad \text{for any } x \in S. \tag{2.4}$$

2.2. **Examples.** There are numerous specific stochastic control problems that lie within the setting described in Section 2.1. We mention two classes of examples.

2.2.1. *Discounted infinite horizon problem.* Let X^a be a killed Markov process with $S = D \cup \{\partial\}$ with ∂ an isolated cemetery state. Killing to ∂ occurs at a (possibly state and control-dependent) rate α and τ is the death time of the process. A special case of the controlled (killed) Itô diffusion process will be described in Section 4.1. The detailed proofs that the Policy Improvement Algorithm from Section 3 below can be applied in this case are given in [5].

Remark 3. The general setting allows us to consider more general problems where τ is the earlier of the killing time and exit from a domain. As is usual, we may also assume that the killing time is unobserved so that, conditioning on the sample path and control we revise problem (2.1) to the standard killed version, where $V^{\Pi}(x)$ and $V(x)$ are given in (2.3) and (2.4), respectively, with

$$J(x,\Pi) := \int_0^\tau \exp\left(-\int_0^t \alpha(X_s^\Pi, \Pi_s)ds\right) f(X_t^\Pi, \Pi_t)dt + \exp\left(-\int_0^\tau \alpha(X_s^\Pi, \Pi_s)ds\right) g(X_\tau^\Pi)1_{(\tau<\infty)}.$$

2.2.2. *The finite horizon problem.* Let Y^a be a Markov process on a topological space S' with infinitesimal generator $\mathcal{G}^a$ and τ the time to the horizon T. Define $S := S' \times \mathbb{R}$ and $D := S' \times \mathbb{R}_+$, so if $x = (y, T)$ then $X_t^a = (Y_t^a, T - t)$, $\tau = T$ and $\mathcal{L}^a = \mathcal{G}^a - \frac{\partial}{\partial t}$. The detiled proofs that the PIA in Section 3 below works in this setting are given in [6].

2.2.3. *The weak formulation of the control problem is essential.* In this example we demonstrate that it is necessary to formulate the stochastic control setting in Section 2.1 using the weak formulation in order not to exclude natural examples of the control problems.

(I) In our formulation it is possible for two controls Π and Σ to have the same law but the pairs (X^Π, Π) and (X^Σ, Σ) not to. Consider $S := \mathbb{R}$, $A := \{-1, 1\}$ and, for $a \in A$, the strong Markov process X^a is given by

$$dX_t^a = a\, dV_t, \tag{2.5}$$

where V is any Brownian motion. Let W be a fixed Brownian motion on a stochastic basis. Define $\Pi := \operatorname{sgn}(W)$ (with $\operatorname{sgn}(0) := 1$) and in (2.5) let V be defined by the stochastic integral $V_t := \int_0^t \operatorname{sgn}(W_s)dW_s$. Then $X^\Pi = W$ and hence $(X^\Pi, \Pi) = (W, \operatorname{sgn}(W))$. Take $\Sigma := \operatorname{sgn}(W)$ and in (2.5) let $V := W$. Then, by the Tanaka formula, we have

$$X_t^\Sigma = \int_0^t \operatorname{sgn}(W_s)\, dW_s = |W_t| - L_t^0(W),$$

where $L^0(W)$ is the local time of W at zero. It is clear that X^Σ is a Brownian motion and hence $X^\Pi \stackrel{d}{=} X^\Sigma$ and $\Pi \stackrel{d}{=} \Sigma$. However the random vectors (X^Π, Π) and (X^Σ, Σ) have distinct joint laws, e.g. $\mathbb{P}(X_t^\Pi > 0, \Pi_t = -1) = 0 < \mathbb{P}(X_t^\Sigma > 0, \Sigma_t = -1)$ for any $t > 0$.

In order to show that such strategies can arise as optimal strategies, consider (in the context of Section 2.2.1) the controlled process in (2.5) with $D := (-1, 1)$ and

$$J(x,\Pi) := \exp\left(-\int_0^\tau \alpha(X_t^\Pi, \Pi_t)dt\right) \cdot g(X_\tau^\Pi), \quad \text{where} \quad \alpha(x,a) := \begin{cases} 2 + a, & x \in D, \\ \infty, & x \notin D, \end{cases}$$

τ is the first exit of X^Π from the interval $(-1,1)$ and $g:\{-1,1\}\to\mathbb{R}$ is given by

$$g(1) := -\sinh(\sqrt{6}), \qquad g(-1) := \sqrt{3}\sinh(\sqrt{2}).$$

Define the function $\widehat{V}: S\to\mathbb{R}$ by

$$\widehat{V}(x) := \begin{cases} -\sinh(\sqrt{6}x), & x\geq 0,\\ -\sqrt{3}\sinh(\sqrt{2}x), & x<0,\end{cases}$$

and note that $\widehat{V}$ is C^1, piecewise C^2 and the following equalities hold for all $x\in D\setminus\{0\}$:

$$\operatorname{sgn}(\widehat{V}(x)) = -\operatorname{sgn}(x) \quad \text{and} \quad \widehat{V}''(x) = (2+4\cdot 1_{\{x>0\}})\,\widehat{V}(x).$$

Hence the following HJB equation holds (recall $a\in A=\{-1,1\}$ and thus $a^2=1$):

$$\sup_{a\in A}\left[\frac{a^2}{2}\widehat{V}'' - (a+2)\widehat{V}\right] = 0, \qquad \text{with boundary condition } \widehat{V}|_{\partial D} = g,$$

and the supremum is attained at $a=\operatorname{sgn}(x)$. Now, a standard application of martingale theory and stochastic calculus implies that the Markov policy $\pi(x):=\operatorname{sgn}(x)$ is optimal for problem (2.1) (with the controlled process given in (2.5)) and its payoff V^π equals the value function V in (2.1): $V(x)=V^\pi(x)=\widehat{V}(x)$ for all $x\in D$.

(II) It may appear at first glance that the weak formulation of the solution only played a role in Example (I) due to the fact that the space of controls in (I) was restricted to $A=\{-1,1\}$. Indeed, it holds that if in Example (I) we allow controls in the interval $[-1,1]$, then the Markov control $\pi(x)=\operatorname{sgn}(x)$ is no longer optimal (as the HJB equation is no longer satisfied). However, the weak formulation of the control problem is essential even if we allow the controller to choose from an uncountable set of actions at each moment in time. We now illustrate this point by describe an example where the Markov control $\pi(x)=\operatorname{sgn}(x)$ is optimal, while the controls take values in the closed interval.

Consider the controlled process X^a in (2.5) with S and D as in Example (I). Let $A:=[-1,1]$ and define

$$J(x,\Pi) := \exp\left(-\int_0^\tau \alpha(X_t^\Pi,\Pi_t)dt\right)\cdot g(X_\tau^\Pi) - \int_0^\tau \exp\left(-\int_0^t \alpha(X_s^\Pi,\Pi_s)ds\right)\cdot f(X_t^\Pi)dt, \tag{2.6}$$

where τ is the first exit of X^Π from the interval $D=(-1,1)$,

$$\alpha(x,a) := \begin{cases} 4a+9/2, & x\in D,\ a\in A,\\ \infty, & x\notin D,\ a\in A,\end{cases}, \qquad f(x) := \frac{13}{2}\sinh(2\max\{x,0\}), \quad \text{for } x\in\mathbb{R},$$

and $g:\{-1,1\}\to\mathbb{R}$ is given by

$$g(1) := -\sinh(2), \qquad g(-1) := 2\sinh(1).$$

Define the function $\widehat{V}: S\to\mathbb{R}$ by

$$\widehat{V}(x) := \begin{cases} -\sinh(2x), & x\geq 0,\\ -2\sinh(x), & x<0.\end{cases}$$

Note that $\widehat{V}$ is C^1, piecewise C^2 and, for all $x \in D \setminus \{0\}$, it holds

$$\operatorname{sgn}(\widehat{V}(x)) = -\operatorname{sgn}(x) \quad \text{and} \quad \widehat{V}''(x) = (1 + 3 \cdot 1_{\{x>0\}})\, \widehat{V}(x).$$

We now show that the HJB equation

$$\sup_{a\in[-1,1]} \left[\frac{a^2}{2} \widehat{V}'' - (4a + 9/2)\widehat{V} \right] - f = 0 \tag{2.7}$$

holds with boundary condition $\widehat{V}|_{\partial D} = g$ and the supremum attained at $a = \operatorname{sgn}(x)$. We first establish (2.7) for $x > 0$. In this case (2.7) reads

$$\sup_{a\in[-1,1]} \left[\sinh(2x) \left((4a + 9/2) - 2a^2 \right) \right] - \frac{13}{2} \sinh(2x) = 0.$$

Now, the function $a \mapsto 4a - 2a^2$ is increasing on $[-1, 1]$. Hence the supremum is attained at $a = 1$ and the equality follows. In the case $x < 0$, the HJB equation in (2.7) takes the form

$$\sup_{a\in[-1,1]} \left[-\sinh(x) \left(a^2/2 - (4a + 9/2) \right) \right] = 0.$$

The function $a \mapsto a^2/2 - 4a - 9/2$ is decreasing on the interval $[-1, 1]$ and has a zero at $a = -1$. Hence the HJB equation in (2.7) holds with the stated boundary condition. The classical martingale argument implies that the Markov policy $\pi(x) := \operatorname{sgn}(x)$ is optimal for problem (2.1) (with $J(x, \Pi)$ given in (2.6)) and its payoff V^π equals the value function V in (2.1): $V(x) = V^\pi(x) = \widehat{V}(x)$ for all $x \in D$.

3. The policy improvement algorithm (PIA)

In order to develop the policy improvement algorithm, we first have to define the notion of an improvable Markov policy.

Definition 1. A Markov policy π is *improvable* if $V^\pi \in \mathbf{C}$. The collection of improvable Markov policies is denoted by I. A Markov policy π' is an *improvement* of $\pi \in I$ if,

(I) for each $x \in D$

$$\pi'(x) \in \arg\max_{a\in A} [\mathcal{L}^a V^\pi(x) + f(x, a)],$$

or equivalently put

$$\mathcal{L}^{\pi'(x)} V^\pi(x) + f(x, \pi'(x)) = \sup_{a\in A} [\mathcal{L}^a V^\pi(x) + f(x, a)],$$

and

(II) π' is also a Markov policy.

3.1. **Improvement works.** The PIA works by defining a sequence of improvements and their associated payoffs. More specifically, π_{n+1} is the improvement of the improvable Markov policy π_n (in the sense of Definition 1). With this in mind, we make the following assumptions:

(As1): there exists a non-empty subset I^* of I such that $\pi_0 \in I^*$ implies that, for each $n \in \mathbb{N}$, the Markov policy π_n is a continuous function in I^*;

(As2): for any Markov policy $\pi_0 \in I^*$, let the difference of consecutive payoff processes converge in L^1 to a non-negative random variable:

$$\lim_{t\uparrow\infty} \left(V^{\pi_{n+1}}(X^{\pi_{n+1}}_{t\wedge\tau}) - V^{\pi_n}(X^{\pi_{n+1}}_{t\wedge\tau})\right) \overset{L^1}{=} Z_x \geq 0 \text{ a.s. for each } x \in D.$$

Remark 4. The key assertions in Assumption **(As1)** are that, for every $n \in \mathbb{N}$, the payoff V^{π_n} is in $\mathbf{C}$ (see (2) in Section 2.1 for the definition of $\mathbf{C}$) and that the sup in (I) of Definition 1 is attained.

The following theorem asserts that the algorithm, under the assumptions above, actually improves improvable policies. We prove it in Section 5.1 below.

Theorem 1. *Under Assumptions **(As1)** and **(As2)**, the inequality*

$$V^{\pi_{n+1}}(x) \geq V^{\pi_n}(x) \qquad \textit{holds for each } n \in \mathbb{N} \textit{ and all } x \in S.$$

3.2. **Convergence of payoffs.** Assume from now on that Assumptions **(As1)** and **(As2)** hold and that we have fixed an improvable Markov policy π_0 in I^*. Denote by $(\pi_n)_{n\in\mathbb{N}}$ the sequence of Markov policies in I^* defined by the PIA started at π_0 (see the beginning of Section 3.1).

(As3): The value function V, defined in (2.1), is finite on the domain D.

(As4): There is a subsequence $(n_k)_{k\in\mathbb{N}}$ such that

$$\lim_{k\nearrow\infty} \mathcal{L}^{\pi_{n_k+1}} V^{\pi_{n_k}}(x) + f(x, \pi_{n_k+1}(x)) = 0 \quad \text{uniformly in } x \in D.$$

(As5): For each $x \in S$, each $\Pi \in \mathcal{A}_x$ and each $n \in \mathbb{N}$ the following limit holds:

$$V^{\pi_n}(X^{\Pi}_{t\wedge\tau}) \xrightarrow{L^1} g(X^{\Pi}_{\tau})1_{(\tau<\infty)} \qquad \text{as } t \to \infty.$$

The next result states that the PIA works. Its proof is in Section 5.2 below.

Theorem 2. *Under Assumptions **(As1)**–**(As5)**, the following limit holds:*

$$V^{\pi_n}(x) \uparrow V(x) \qquad \textit{for all } x \in S.$$

3.3. **Convergence of policies.** Assume from now on that Assumptions **(As1)**–**(As5)** hold and that, as before, we have fixed a π_0 in I^* together with the sequence of improved Markov policies $(\pi_n)_{n\in\mathbb{N}}$.

(As6): For any $\pi_0 \in I^*$, the sequence $(\pi_n)_{n\in\mathbb{N}}$ is sequentially precompact in the topology of uniform convergence on compacts on the space of continuous functions $C(S, A)$.

(As7): For any sequence $(\rho_n)_{n\in\mathbb{N}}$ in I^*, such that

(i) $\exists$ Markov policy ρ, such that $\rho_n \overset{n\to\infty}{\longrightarrow} \rho$ uniformly on compacts in S,
(ii) $\phi_n \in \mathbf{C}$ for all $n \in \mathbb{N}$ and $\phi_n \overset{n\to\infty}{\longrightarrow} \phi$ pointwise,
(iii) $\mathcal{L}^{\rho_n}\phi_n \overset{n\to\infty}{\longrightarrow} Q$ uniformly on compacts in S,
then

$$\phi \in \mathbf{C}, \quad \mathcal{L}^{\rho}\phi = Q \text{ and } \mathcal{L}^{\rho}\phi_n - \mathcal{L}^{\rho_n}\phi_n \overset{n\to\infty}{\longrightarrow} 0$$

uniformly on compacts in S.

(As8): For each $x \in D$ and each $\Pi \in \mathcal{A}_x$,

$$V(X^{\Pi}_{t\wedge\tau}) \overset{L^1}{\longrightarrow} g(X^{\Pi}_{\tau})1_{(\tau<\infty)} \qquad \text{as } t \to \infty,$$

holds.

The next theorem states that the sequence of policies produced by the PIA contains a uniformly convergent subsequence. We give a proof of this fact in Section 5.3 below.

Theorem 3. *Under Assumptions **(As1)**–**(As8)**, for any π_0 in I^* and the corresponding sequence of improved Markov policies $(\pi_n)_{n\in\mathbb{N}}$, there exists a subsequence $(\pi_{n_k})_{k\in\mathbb{N}}$ such that $\pi_{n_k} \overset{n\to\infty}{\longrightarrow} \pi^*$ in the topology of uniform convergence on compacts and $V^{\pi^*} = V$.*

4. Examples of the PIA

4.1. **Discounted infinite horizon controlled diffusion.** This section gives an overview of the results in [5]. Define $D := \mathbb{R}^d$ and $S := \mathbb{R}^d \cup \{\partial\}$ and let $\mathbf{C} = C^2_b(\mathbb{R}^d, \mathbb{R})$ be the space of bounded, C^2, real-valued functions on $\mathbb{R}^d$. Suppose that X is a controlled (killed) Itô diffusion in $\mathbb{R}^d$, so that

$$\mathcal{L}^a\phi(\cdot) = \frac{1}{2}\sigma(\cdot,a)^T H\phi\sigma(\cdot,a) + \mu(\cdot,a)^T\nabla\phi - \alpha(\cdot,a)\phi, \tag{4.1}$$

where $H\phi$ (resp. $\nabla\phi$) denotes the Hessian (resp. gradient) with entries $\frac{\partial^2\phi}{\partial x_i\partial x_j}$, $1 \le i,j \le d$ (resp. $\frac{\partial\phi}{\partial x_i}$, $1 \le i \le d$). Furthermore, we make the following assumptions on the deterministic characteristics of the model:

(ℵ1): $\sigma(x,a)$, $\mu(x,a)$, $\alpha(x,a)$ and $f(x,a)$ are uniformly (in a) Lipschitz on compacts in $\mathbb{R}^d$ and are continuous in a; α is bounded below by a positive constant $\lambda > 0$, σ is uniformly elliptic and f is uniformly bounded by a (large) constant M.

(ℵ2): The control set A is a compact interval $[a,b]$.

For every $h \in \mathbf{C}$ and $x \in \mathbb{R}^d$, let $I_h(x)$ denote an element of $\arg\max_{a\in A}[\mathcal{L}^a h(x,a) + f(x,a)]$.

(ℵ3): If the sequence of functions $(h_n)_{n\in\mathbb{N}}$ is in C^2 and the sequence $(Hh_n)_{n\in\mathbb{N}}$ is uniformly bounded on compacts, then we may choose the sequence of functions $(I_{h_n})_{n\in\mathbb{N}}$ to be uniformly Lipschitz on compacts.

Remark 5. (1) The assumption in **(ℵ3)** is very strong. Neverthless, if σ is independent of a and bounded, $\mu(x,a) = \mu_1(x) - ma$, $\alpha(x,a) = \alpha_1(x) + ca$ and $f(x,a) = f_1(x) - f_2(a)$ with

$f_2 \in C^1$ and with strictly positive derivative on A, and Assumptions **(ℵ1)** and **(ℵ1)** hold, then **(ℵ3)** holds.

(2) We stress that the assumptions in **(ℵ1)**–**(ℵ3)** do not depend on the stochastic behaviour of the model but are given explicitly in terms of its deterministic characteristics. This makes the PIA provably convergent for a broad class of diffusion control problems.

Proposition 4. *Under Assumptions **(ℵ1)**–**(ℵ3)**, Assumptions **(As1)**–**(As8)** hold for the (possibly killed) controlled diffusion process with generator* (4.1) *and the PIA converges when started at any locally Lipschitz Markov policy π_0.*

Proof. Note that $\mathcal{L}^a\phi$ is jointly continuous if ϕ is in $\mathbf{C}$ and (with the usual trick to deal with killing) (2.2) holds for any control Π such that there is a solution to the killed equation

$$X_t^\Pi = (x + \int_0^t \sigma(X_s^\Pi, \Pi_s) dB_s + \int_0^t \mu(X_s^\Pi, \Pi_s) ds) 1_{(t<\tau)} + \partial 1_{(t\geq\tau)}.$$

Furthermore, any locally Lipschitz π is a Markov policy by strong uniqueness of the solution to the SDE. We now establish Assumptions **(As1)**–**(As8)**.

(As1) If π_0 is Lipschitz on compacts then by Assumption **(ℵ3)**, **(As1)** holds.

(As3) Boundedness of V in **(As3)** follows from the boundedness of f and the fact that α is bounded away from 0.

(As6) Assumption **(ℵ3)** implies that (π_n) are uniformly Lipschitz and hence sequentially pre-compact in the sup-norm topology (A6) by the Arzela-Ascoli Theorem.

(As5) $g = 0$ and since α is bounded away from 0, for any Π, $X_t^\Pi \to \partial$. Now $V^{\pi_n}(\partial) = 0$ and so, by bounded convergence, **(As5)** holds:

$$V^{\pi_n}(X^\Pi_{t\wedge\tau}) \xrightarrow{L^1} g(X^\Pi_\tau) 1_{(\tau<\infty)} \qquad \text{as } t \to \infty.$$

(As2) Similarly, **(As2)** holds:

$$V^{\pi_{n+1}}(X^{n+1}_{t\wedge\tau}) - V^{\pi_n}(X^{n+1}_{t\wedge\tau}) \xrightarrow{L^1} 0 \qquad \text{as } t \to \infty.$$

(As4) The statement in **(As4)** is trickier to establish. Note that we have **(As1)** and **(As2)**, by Theorem 1, we know that $V^{\pi_n}(x)$ is a non-decreasing sequence. Moreover, since **(As3)** holds, $V^{\pi_n} \uparrow V^{lim}$. Now take a subsequence $(n_k)_{k\in\mathbb{N}}$ such that $(\pi_{n_k}, \pi_{n_k+1}) \to (\pi^*, \tilde{\pi})$ uniformly on compacts. Then the corresponding σ etc. must also converge. Denote the limits by σ^*, $\tilde{\sigma}$ etc. Then, $V^{lim} \in \mathbf{C}_b^2$ (see the argument in [5], based on coupling and the classical PDE theory from Friedman [2]) and

$$\lim_{k\to\infty} \nabla V^{\pi_{n_k}} = \lim_{k\to\infty} \nabla V^{\pi_{n_k+1}} = \nabla V^{lim} \quad \text{and} \quad \lim_{k\to\infty} HV^{\pi_{n_k}} = \lim_{k\to\infty} HV^{\pi_{n_k+1}} = HV^{lim}$$

uniformly on compacts and $\mathcal{L}^{\tilde{\pi}} V^{lim} + f(\cdot, \tilde{\pi}(\cdot)) = 0$. Now, from the convergence of the derivatives of $V^{\pi_{n_k}}$, we obtain

$$\mathcal{L}^{\pi_{n_k+1}} V^{\pi_{n_k}} + f(\cdot, \pi_{n_k+1}(\cdot)) \to \mathcal{L}^{\tilde{\pi}} V^{lim} + f(\cdot, \tilde{\pi}(\cdot)) = 0$$

uniformly on compacts.

(As7) and **(As8)** follow from Friedman [2]. See [5] for details. ♢

4.2. **Finite horizon controlled diffusion.** This is very similar to the previous example if we add the requirement that g is Lipschitz and bounded. The details can be found in [6].

Remark 6. In both examples we need to prove that V^{π_n} is continuous before we can apply the usual PDE arguments. This crucial step is carried out in [5] and [6] respectively.

5. Proofs

Lemma 5. *Under Assumptions **(As1)** and **(As2)**, it holds that*

$$\mathcal{L}^{\pi_n}V^{\pi_n}(x) + f(x, \pi_n(x)) = 0 \qquad \text{for all } x \in D \text{ and } n \in \mathbb{N}.$$

Proof. We know that

$$V^{\pi_n}(X^{\pi_n}_{t\wedge\tau}) - \int_0^{t\wedge\tau} \mathcal{L}^{\pi_n}V^{\pi_n}(X^{\pi_n}_s)ds$$

is a martingale and the usual Markovian argument shows that therefore

$$\int_0^{t\wedge\tau} (\mathcal{L}^{\pi_n}V^{\pi_n} + f(\cdot, \pi_n(\cdot))(X^{\pi_n}_s)ds = 0.$$

The result then follows from continuity of $\mathcal{L}^{\pi_n}V^{\pi_n} + f(\cdot, \pi_n(\cdot))$ (see (2) in Section 2.1) and the right continuity of X^{π_n}. ♢

5.1. **Proof of Theorem 1.** Take $\pi_0 \in I^*$ and $x \in D$ and let $(\pi_n)_{n\in\mathbb{N}}$ be the sequence of policies produced by the PIA. For any $n \in \mathbb{N}$ define

$$S_t := (V^{\pi_{n+1}} - V^{\pi_n})(X^{\pi_{n+1}}_{t\wedge\tau}), \qquad t \geq 0.$$

By assumption, both payoffs $V^{\pi_{n+1}}$ and V^{π_n} are in $\mathbf{C}$. Hence the process

$$V^{\pi_k}(X^{\pi_{n+1}}_{t\wedge\tau}) - \int_0^{t\wedge\tau} \mathcal{L}^{\pi_{n+1}}V^{\pi_k}(X^{\pi_{n+1}}_s)ds, \qquad \text{for } k = n,\, n+1,$$

is a martingale. So,

$$S_t = (V^{\pi_{n+1}} - V^{\pi_n})(x) + M_{t\wedge\tau} + \int_0^{t\wedge\tau} (\mathcal{L}^{\pi_{n+1}}V^{\pi_{n+1}} - \mathcal{L}^{\pi_{n+1}}V^{\pi_n})(X^{\pi_{n+1}}_s)ds,$$

where M is a martingale. Thus

$$S_t = (V^{\pi_{n+1}} - V^{\pi_n})(x) + M_{t\wedge\tau} - \int_0^{t\wedge\tau} \sup_{a\in A}[\mathcal{L}^a V^{\pi_n} + f(\cdot, a)](X^{\pi_{n+1}}_s)ds,$$

by Lemma 5 and the definition of π_{n+1}. Appealing to Lemma 5 again, the integrand is non-negative and hence S is a supermartingale. Taking expectations and letting $t \to \infty$ we obtain the result using **(As2)**. □

5.2. **Proof of Theorem 2.** From Theorem 1 and **(As3)**, $V^{\pi_n}(x) \uparrow V^{lim}(x)$ holds for any $x \in S$ as $n \to \infty$, where the function V^{lim} is finite and bounded above by V. Fix $x \in D$ and $\Pi \in \mathcal{A}_x$ and take the subsequence $(\pi_{n_k})_{k\in\mathbb{N}}$ in **(As4)**. Set

$$S_t^k = V^{\pi_{n_k}}(X^{\Pi}_{t\wedge\tau}) + \int_0^{t\wedge\tau} f(X_s^{\Pi}, \Pi_s)ds.$$

It follows that there is a martingale M^k such that

$$\begin{aligned} S_t^k &= S_0^k + M^k_{t\wedge\tau} + \int_0^{t\wedge\tau} [\mathcal{L}^{\Pi_s} V^{\pi_{n_k}} + f(\cdot, \Pi_s)](X_s^{\Pi})ds \\ &\leq S_0^k + M^k_{t\wedge\tau} + \int_0^{t\wedge\tau} \sup_{a\in A}[\mathcal{L}^{a} V^{\pi_{n_k}} + f(\cdot, a)](X_s^{\Pi})ds \\ &= S_0^k + M^k_{t\wedge\tau} + \int_0^{t\wedge\tau} [\mathcal{L}^{\pi_{n_k+1}} V^{\pi_{n_k}} + f(\cdot, \pi_{n_k+1}(\cdot))](X_s^{\Pi})ds \end{aligned}$$

So

$$\mathbb{E}S_t^k \leq V^{\pi_{n_k}}(x) + \mathbb{E}\int_0^{t\wedge\tau} [\mathcal{L}^{\pi_{n_k+1}} V^{\pi_{n_k}} + f(\cdot, \pi_{n_k+1}(\cdot))](X_s^{\Pi})ds. \tag{5.1}$$

Letting $k \to \infty$ in (5.1) we obtain, by **(As4)**, together with dominated convergence, and monotone convergence, that

$$V^{lim}(x) \geq \mathbb{E}[V^{lim}(X^{\Pi}_{t\wedge\tau}) + \int_0^{t\wedge\tau} f(X_s^{\Pi}, \Pi_s)ds] \qquad \text{for all } t \geq 0. \tag{5.2}$$

Now **(As5)** and Fatou's lemma (recall $V^{lim} \geq V^{\pi_{n_k}} \geq 0$ for any index k) imply

$$\liminf_{t\to\infty} \mathbb{E}V^{lim}(X^{\Pi}_{t\wedge\tau}) \geq \liminf_{t\to\infty} \mathbb{E}V^{\pi_{n_k}}(X^{\Pi}_{t\wedge\tau}) \geq \mathbb{E}\liminf_{t\to\infty} V^{\pi_{n_k}}(X^{\Pi}_{t\wedge\tau}) = \mathbb{E}g(X^{\Pi}_{\tau})1_{(\tau<\infty)},$$

and so (5.2) yields $V^{lim}(x) \geq V^{\Pi}(x)$ for each $\Pi \in \mathcal{A}_x$. Hence $V^{lim} \geq V$ on the domain D (on the complement we clearly have $V^{lim} = V$). However, since by definition $V^{\pi_n} \leq V^{lim}$ on S for all $n \in \mathbb{N}$, $V^{lim} = \lim_{n\to\infty} V^{\pi_n} \leq V$ on S so in fact we have equality. □

5.3. **Proof of Theorem 3.** Let $(\pi_{n_j})_{j\in\mathbb{N}}$ be a subsequence, guaranteed by **(As6)**, of the sequence of Markov policies $(\pi_n)_{n\in\mathbb{N}}$ in I^* produced by the PIA. Put differently, the limit

$$\lim_{j\to\infty} \pi_{n_j} = \pi^*$$

holds uniformly on compacts in S for some Markov policy π^*. Hence, for any $x \in D$, there exists (by the definition of a Markov policy) a controlled process X^{π^*} defined on some filtered probability space.

Fix $x \in D$, define the process $S^j = (S_t^j)_{t\geq 0}$,

$$S_t^j := V^{\pi_{n_j}}(X^{\pi^*}_{t\wedge\tau}) + \int_0^{t\wedge\tau} f(X_s^{\pi^*}, \pi^*(X_s^{\pi^*}))ds, \tag{5.3}$$

and note that the following equality holds

$$S_t^j = V^{\pi_{n_j}}(x) + M_t + \int_0^{t\wedge\tau} \left[\mathcal{L}^{\pi^*} V^{\pi_{n_j}} + f(\cdot, \pi^*(\cdot))\right](X_s^{\pi^*})ds \qquad \text{for any } t \geq 0, \tag{5.4}$$

where the martingale $M = (M_t)_{t\geq 0}$ is given by

$$M_t := V^{\pi_{n_j}}(X^{\pi^*}_{t\wedge\tau}) - V^{\pi_{n_j}}(x) - \int_0^{t\wedge\tau} \mathcal{L}^{\pi^*} V^{\pi_{n_j}}(X^{\pi^*}_s))ds.$$

By Lemma 5 and the representation in (5.4) we obtain

$$S^j_t = V^{\pi_{n_j}}(x) + M_t + \int_0^{t\wedge\tau} [(\mathcal{L}^{\pi^*} - \mathcal{L}^{\pi_{n_j}})V^{\pi_{n_j}} + f(\cdot,\pi^*(\cdot)) - f(\cdot,\pi_{n_j}(\cdot))](X^{\pi^*}_s)ds.$$

Take expectations on both sides of this identity. By localising, applying **(As7)** and using Theorem 2 we obtain

$$V(x) = \lim_{j\to\infty} \mathbb{E}[S^j_t] \qquad \text{for any } t \geq 0. \tag{5.5}$$

Definition (5.3) and Theorem 2 imply a.s. monotone convergence

$$S^j_t \nearrow V(X^{\pi^*}_{t\wedge\tau}) + \int_0^{t\wedge\tau} f(X^{\pi^*}_s, \pi^*(X^{\pi^*}_s))ds \qquad \text{as } j \to \infty \text{ for any } t \geq 0.$$

Hence by the monotone convergence theorem and (5.5) we find

$$V(x) = \mathbb{E}[\lim_{j\to\infty} S^j_t] = \mathbb{E}\left[V(X^{\pi^*}_{t\wedge\tau}) + \int_0^{t\wedge\tau} f(X^{\pi^*}_s, \pi^*(X^{\pi^*}_s))ds\right] \qquad \text{for any } t \geq 0.$$

Letting $t \to \infty$, applying **(As8)** and recalling the definition of V^{π^*} yields the result that $V^{\pi^*} = V$. □

References

[1] B. T. Doshi, Continuous Time Control of Markov Processes on an Arbitrary State Space: Discounted Rewards. *The Annals of Statistics*, 4(6):1219–1235, 1976.

[2] A Friedman, *Partial Differential Equations of Parabolic Type*. Prentice-Hall, Englewood Cliffs, N.J., 1964.

[3] A. Hordijk and M. L. Puterman, On the convergence of policy iteration in finite state undiscounted Markov decision processes: the unichain case. *Mathematics of Operations Research*, 12(1):163–176, 1987.

[4] R. A. Howard, *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, 1960.

[5] S. D. Jacka, A. Mijatović and D. Širaj, Policy Improvement Algorithm for Controlled Multidimensional Diffusion Processes, *in preparation*.

[6] S. D. Jacka, A. Mijatović and D. Širaj, Coupling of Diffusions and the Policy Improvement Algorithm for a Finite Horizon Problem, *in preparation*.

[7] S. P. Meyn, The policy iteration algorithm for average reward Markov decision processes with general state space. *IEEE Transactions on Automatic Control*, 42(12):1663–1680, 1997.

[8] M. S. Santos and J. Rust, Convergence properties of policy iteration. *SIAM Journal on Control and Optimization*, 42(6):2094–2115, 2004.

Department of Statistics, University of Warwick, UK

E-mail address: s.d.jacka@warwick.ac.uk

Department of Mathematics, Imperial College London, UK

E-mail address: a.mijatovic@imperial.ac.uk