APPLICATIONES MATHEMATICAE
47,2 (2020), pp. 225–253

B. A. Escobedo-Trujillo (Coatzacoalcos)
O. Hernández-Lerma (Ciudad de México)
F. A. Alaffita-Hernández (Coatzacoalcos)

ADAPTIVE CONTROL OF DIFFUSION PROCESSES WITH A DISCOUNTED REWARD CRITERION

Abstract. The optimal control problem we are dealing with in this paper is to determine control policies that maximize a discounted reward criterion when the dynamic system evolves as a stochastic differential equation (SDE). Both the instantaneous reward function and the SDE's drift coefficient may depend on an unknown parameter. We give conditions ensuring the existence of an asymptotically optimal policy using the so-called Principle of Estimation and Control. We illustrate our results with several examples.

1. Introduction. An adaptive control problem is an optimal control problem (OCP) that is not completely specified in the sense that the system dynamics or the optimality criterion depend on an unknown parameter θ. In this paper the optimal control problem is to maximize a discounted reward criterion when the system evolves as a diffusion process. To analyze our adaptive control problem we follow the Principle of Estimation and Control (PEC), which can be traced back to Kurano (1972) and Mandl (1974). The idea is simply to estimate the parameter θ, and then solve the OCP when (the unknown) θ is replaced by its estimated values. Here we present conditions under which there exists a sequence of estimators θ_m that converge to θ, and therefore the following happens: (a) For each m, there is an estimator θ_m such that $\theta_m \to \theta$ as $t \to \infty$; (b) the optimal control policy π^*_m corresponding to the θ_m-OCP converges (in some sense) to the optimal policy π^*

2020 *Mathematics Subject Classification*: 93E10, 93E20, 93E24, 60J60.
Key words and phrases: adaptive policy, consistent estimator.
Received 28 August 2020.
Published online 9 December 2020.

DOI: 10.4064/am2421-10-2020

for the θ-OCP; and also (c) the corresponding optimal reward function $V^*_{\theta_m}$ converges to V^*_θ.

The PEC is another name for what Mandl called *the Method of Substituting the Estimates into Optimal Stationary Controls*, and which, except for small variations, is also found in the stochastic control theory literature under the name of *Certainty Equivalence Controller*.

In the remainder of this section we present summaries of parameter estimation techniques (Section 1.1), adaptive control approaches (Section 1.2), and our approach and main contributions in this paper (Section 1.3). We conclude this introduction with some remarks on terminology and notation we use (Section 1.4).

1.1. Parameter estimation techniques in diffusion processes. For computational reasons, in this paper we only consider parameter estimation techniques that estimate the unknown θ at a sequence of times $0 < t_1 < t_2 < \cdots$, with $t_m \to \infty$ (see [1, 2, 13, 14, 22, 31, 32, 34]). There are online and offline estimation methods. In the former, the estimate θ_m is obtained based on the information available up to time t_m only. In contrast, in an offline estimation method, first all the input/output data are collected and then the parameter estimates are obtained. In this paper, we use offline methods to generate consistent estimators.

If the transition density of the dynamic system is known, then it is common to choose the maximum likelihood estimator (MLE), which maximizes the log-likelihood function. The MLE is both intuitive and flexible, and as such the estimator has become a dominant means of statistical inference; see [1, 2, 22] and the references therein. But, most often, the transition density is unknown and so the log-likelihood function of the dynamic system is unknown as well. Hence, other methods of estimation have to be considered.

The paper [32] presents an overview focused on parameter estimation methods for discretely observed stochastic differential equations (SDEs) over the period 1981–1999. These methods are: Generalized Method of Moments (GMM), Efficient Method of Moments, and three methods based on the likelihood function: (a) discretization of the likelihood function that, under some technical conditions, follows from an assumption of continuous observations being available, (b) the likelihood function is derived for a discretized version of the SDE, where the discretization time step δ is equal to the sampling interval Δ; this method is called an approximate maximum likelihood method, and (c) the extension proposed in [34] of an approximate maximum likelihood method which assumes that $\delta \ll \Delta$.

The paper [33] presents a partial survey on the statistical estimation of diffusion processes. This survey was concentrated on contributions to the literature based on three different approaches in which the likelihood func-

tion of the observations is not directly computable: (1) the Euler–Maruyama discretization scheme, (2) martingale estimating functions (martingale theory), and (3) Generalized Method of Moments. The techniques (1)–(2) are based on replacements of the true likelihood function (which is not known) by some approximation, whereas the estimator mentioned in (3) is a vector that minimizes a distance function, properly defined, of the sample moments from zero. Under some assumptions, these estimators are consistent and asymptotically normal.

For linear dynamic systems, the main techniques for the identification of unknown parameters are the Least Squares Estimator (LSE), and the approximate maximum likelihood estimator (AMLE). As mentioned in [27], the LSE is suitable when the disturbance is white noise. If the noise is colored, more complex methods are needed to avoid bias and to identify the disturbance process. The AMLE is based on Gaussian approximation of the transition density and can be interpreted as based on maximization of a discretized continuous-time log-likelihood function as well. These estimators are consistent and asymptotically normally distributed [34, 35, 42].

1.2. Adaptive control approaches. The paper [8] studies the *self-tuning scheme* for the adaptive control of a diffusion process with long-run average cost criterion and maximum likelihood estimation of parameters. Asymptotic optimality under a suitable identifiability condition is established under two alternative sets of hypotheses: a Lyapunov-type stability criterion, and a condition on the running cost that penalizes instability. The self-tuning method of adaptive control of diffusions consists of estimating the unknown parameter online by some standard scheme (e.g. maximum likelihood); then, at each time, the current estimate is taken as the true parameter for the selection of the control [6]. The estimation for diffusion processes by discrete observation has been studied by several authors; see [22, 42, 37] and their references. Other papers on estimation for diffusion processes are [13, 14].

Adaptive control of continuous-time linear stochastic systems has been studied since 1990 in [9, 12, 11]. These papers concern adaptive control problems with average (or ergodic) quadratic cost criteria. The estimation methods used are maximum likelihood, least squares based on continuous observations and modified weighted least squares. The adaptive control of a linear diffusion process with regard to the discounted cost criterion is studied in [5]. The author shows that the certainty-equivalence type of control, analogous to the one considered by Duncan and Pasik-Duncan [12], is asymptotically discount optimal in the sense of Schäl [41]. Adaptive optimal control for continuous-time linear systems based on policy iteration is studied in [43]. On the other hand, [10] considers a Bayesian adaptive control problem for

ergodic diffusions. A nearly self-optimizing strategy is constructed on the basis of discrete-time observations of the state process.

The concept of asymptotic discount optimality was introduced by Schäl [41] in connection with adaptive control of discrete-time Markov control processes. The problem of constructing asymptotically optimal adaptive policies has been studied in several contexts; see [20, 21, 29, 30] and their references. For instance, [29] and [30] introduced the Principle of Estimation and Control (PEC) for Markov decision models with finite state space and bounded rewards. They establish the existence of an optimal policy based on a consistent estimator for the unknown parameter which is optimal uniformly in the parameter. The paper [21] considers discrete-time stochastic control systems.

1.3. Our approach and main contributions. Using the AMLE or the LSE to estimate θ, we construct via the PEC a policy which is asymptotically optimal [29, 30]. Our paper can be considered an extension of [8], which studies policy convergence ($\pi_m \to \pi_{\theta^*}$) in the framework of relaxed controls (see Section 4.1). Here, we also consider convergence in the sense of Schäl, which is more convenient in some cases. Moreover, we illustrate our results with applications that clearly show the importance of choosing the right parameter estimation scheme.

1.4. Terminology and notation. For vectors x and matrices A we use the usual Euclidean norm

$$|x|^2 := \sum_k x_k^2 \quad \text{and} \quad |A|^2 := \operatorname{Tr}(AA^T) = \sum_{k,p} A_{k,p}^2,$$

where A^T and $\operatorname{Tr}(\cdot)$ denote the transpose and the trace of a square matrix, respectively. Sometimes we use the notation $\partial_i := \frac{\partial}{\partial x_i}$ and $\partial_{ij}^2 := \frac{\partial^2}{\partial x_i \partial x_j}$. Given a Borel set B, we denote its Borel σ-algebra as $\mathcal{B}(B)$. Furthermore, $\mathcal{P}(B)$ stands for the family of probability measures on $\mathcal{B}(B)$. For any given set $\mathcal{O} \subset \mathbb{R}^n$, $W^{l,p}(\mathcal{O})$ is the Sobolev space of real-valued measurable functions on $\mathcal{O}$ whose generalized derivatives up to order $l \geq 0$ are in $L^p(\mathcal{O})$ for $p \geq 1$. Further, $\mathcal{C}^\kappa(\mathcal{O})$ represents the space of real-valued continuous functions on $\mathcal{O}$ with continuous lth partial derivative in x_i for $i = 1, \ldots, n$, and $l = 0, 1, \ldots, \kappa$. In particular, when $\kappa = 0$, $\mathcal{C}^0(\mathcal{O})$ stands for the space of real-valued continuous functions on $\mathcal{O}$. Moreover, $\mathcal{C}^{\kappa,\beta}(\mathcal{O})$ is the subspace of $\mathcal{C}^\kappa(\mathcal{O})$ consisting of all functions whose partial derivatives up to order κ are Hölder continuous with exponent $\beta \in (0, 1]$; $\mathfrak{B}(\mathcal{O})$ stands for the space of measurable bounded functions on $\mathcal{O}$; and $C_b(\mathcal{O})$ is the space of continuous bounded functions on $\mathcal{O}$.

2. Model and main assumptions. Consider a controlled n-dimensional diffusion process $x(\cdot)$ evolving according to the stochastic differential

equation

$$dx(t) = b(x(t), u(t), \theta)dt + \sigma(x(t))dW(t), \qquad x(0) = x_0,\ t \geq 0, \tag{2.1}$$

where $b : \mathbb{R}^n \times U \times \Theta \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times d}$ are given functions, and $W(\cdot)$ is a d-dimensional standard Brownian motion. The compact set $U \subset \mathbb{R}^{n_1}$ is called the *control* (or *action*) *set*. Moreover, $u(\cdot)$ in (2.1) is a U-valued stochastic process representing the controller's action at each time $t \geq 0$. Let $\Theta \subset \mathbb{R}^m$ be a compact set called the *parameter set*, and $\theta \in \Theta$ is the unknown parameter.

All random processes in (2.1) live in a complete probability space $(\Omega, \mathcal{F}, \mathbb{P}^u_\theta)$, where $\mathbb{P}^u_\theta$ denotes the law of the state process $x(\cdot)$ given the parameter $\theta \in \Theta$ and the control $u(\cdot)$. The σ-algebra $\mathcal{F}$ is $\mathbb{P}^u_\theta$-complete. We impose the following standard assumptions on the drift b and the diffusion matrix σ to guarantee existence and uniqueness of solutions to (2.1).

Assumption 2.1.

(a) *The drift function $b(\cdot,\cdot,\cdot)$ in (2.1) is continuous and Lipschitz in the first and third arguments uniformly in u, that is, there exist nonnegative constants K_θ and D such that, for all $u \in U$, $\theta_1, \theta_2 \in \Theta$ and $x, y \in \mathbb{R}^n$,*

$$|b(x,u,\theta) - b(y,u,\theta)| \leq K_\theta |x-y|,$$
$$|b(x,u,\theta_1) - b(x,u,\theta_2)| \leq D|\theta_1 - \theta_2|.$$

Moreover, $u \mapsto b(x,u,\theta)$ is continuous on U.

(b) *The diffusion coefficient $\sigma(\cdot)$ satisfies a Lipschitz condition: there exists a positive constant K_1 such that, for all $x, y \in \mathbb{R}^n$,*

$$|\sigma(x) - \sigma(y)| \leq K_1 |x-y|.$$

(c) (Uniform ellipticity) *The matrix $a(x) := \sigma(x)\sigma(x)^T$ has the property that, for some constant $K_2 > 0$,*

$$x^T a(y) x \geq K_2 |x|^2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Remark 2.1. The Lipschitz conditions on b and σ in Assumption 2.1(a, b), along with the compactness of U, imply that there exists a constant $\tilde{K} \geq K_1 + K_2 + K_\theta$ such that

$$\sup_{(u,\theta)\in U\times\Theta} |b(x,u,\theta)| + |\sigma(x)| \leq \tilde{K}(1+|x|) \quad \text{for all } x \in \mathbb{R}^n.$$

Control policies. Let $\mathcal{P}(U)$ be the space of probability measures on U endowed with the topology of weak convergence, and let $\mathcal{B}(U)$ be the Borel σ-algebra of U.

Let $\mathbb{M}$ be the family of measurable functions $f : [0,\infty) \times \mathbb{R}^n \to U$, and $\mathbb{F} \subset \mathbb{M}$ the subfamily of functions $f : \mathbb{R}^n \to U$. A strategy $u(t) := f(t, x(t))$, for some $f \in \mathbb{M}$, is called a *Markov policy*, whereas $u(t) := f(x(t))$ for some $f \in \mathbb{F}$ is said to be a *stationary Markov policy*.

DEFINITION 2.2. A *randomized policy* is a family $\pi := \{\pi_t, t \geq 0\}$ of stochastic kernels on $\mathcal{B}(U) \times \mathbb{R}^n$ satisfying:

(a) for each $t \geq 0$ and $x \in \mathbb{R}^n$, $\pi_t(\cdot|x) \in \mathcal{P}(U)$, and for each $t \geq 0$ and $D \in \mathcal{B}(U)$, $\pi_t(D|\cdot)$ is a Borel function on $\mathbb{R}^n$; and
(b) for each $D \in \mathcal{B}(U)$ and $x \in \mathbb{R}^n$, the function $\pi_t(D|x)$ is Borel measurable in $t \geq 0$.

A randomized policy is sometimes called a *relaxed control*.

DEFINITION 2.3. A randomized policy is said to be *stationary* if there is a stochastic kernel $\pi \in \mathcal{B}(U) \times \mathbb{R}^n$ such that $\pi_t(\cdot|x) = \pi(\cdot|x)$ for all $t \geq 0$ and $x \in \mathbb{R}^n$. The set of randomized stationary policies (also known as *relaxed controls*) is denoted by Π.

For $u \in U$, $\theta \in \Theta$ and $\nu_\theta \in W^{2,p}(\mathbb{R}^n)$, let

$$\mathcal{L}^{\theta,u}\nu_\theta(x) := \sum_{i=1}^{m} b^i(x,u,\theta)\partial_i\nu_\theta(x) + \frac{1}{2}\sum_{i,j=1}^{m} a^{ij}(x)\partial^2_{ij}\nu_\theta(x), \tag{2.2}$$

where b^i is the ith component of b, and a^{ij} is the (i,j)-component of the matrix $a(\cdot)$ defined in Assumption 2.1(c).

For each randomized stationary policy $\pi \in \Pi$ we write both the drift coefficient b and the operator $\mathcal{L}^{\theta,u}$ defined in (2.1) and (2.2) respectively as

$$b(x,\pi,\theta) := \int_U b(x,u,\theta)\,\pi(du|x), \qquad \mathcal{L}^{\theta,\pi}\nu_\theta(x) := \int_U \mathcal{L}^{\theta,u}\nu_\theta(x)\,\pi(du|x). \tag{2.3}$$

REMARK 2.4. Observe that every $f \in \mathbb{F}$ can be identified with a strategy in Π by means of the $\mathcal{P}(U)$-valued trajectory δ_f, where $\delta_{f(x)}$ represents the Dirac measure at $f(x)$, i.e., $\pi(\cdot|x) = \delta_{f(x)}(\cdot)$.

Under Assumption 2.1, for each policy $\pi \in \Pi$ and $\theta \in \Theta$ there exists an almost surely unique strong solution $x^{\theta,\pi}(\cdot)$ of (2.1) which is a Markov–Feller process. Furthermore, for each policy $\pi \in \Pi$ and $\theta \in \Theta$, the operator $\mathcal{L}^{\theta,\pi}\nu(x)$ in (2.3) becomes the infinitesimal generator of the process $x^{\theta,\pi}(\cdot)$. (For more details, see the arguments in [3, Theorem 2.2.7] and [17, Theorem 2.1] for instance.) Moreover, by the same reasoning as for [3, Theorem 4.3], we find that for each $\pi \in \Pi$ and $\theta \in \Theta$, the transition probability measure $\mathbb{P}^\pi_\theta(t,x,\cdot)$ of $x^{\theta,\pi}(\cdot)$ is absolutely continuous with respect to Lebesgue's measure for every $t \geq 0$ and $x \in \mathbb{R}^n$. Hence, there exists a transition density function $p^\pi_\theta(t,x,y) \geq 0$ such that

$$\mathbb{P}^\pi_\theta(t,x,B) = \int_B p^\pi_\theta(t,x,y)\,dy$$

for every Borel set $B \subset \mathbb{R}^n$.

Stability assumptions. The following assumption is a standard Lyapunov stability condition for continuous time (controlled and uncontrolled) Markov processes; see, for instance, [3, 25]. It gives, in particular, inequality (2.4) below.

ASSUMPTION 2.2. *There exists a function* $w \geq 1$ *in* $C^2(\mathbb{R}^n)$ *and constants* $d \geq c > 0$ *such that*

(a) $\lim_{|x|\to\infty} w(x) = \infty$.
(b) $\mathcal{L}^{\theta,\pi} w(x) \leq -cw(x) + d$ *for all* $\pi \in \Pi$, $\theta \in \Theta$ *and* $x \in \mathbb{R}^n$.

For every $\pi \in \Pi$, $\theta \in \Theta$, $x \in \mathbb{R}^n$ and $t \geq 0$, an application of Dynkin's formula to the function $v(t,x) := e^{ct}w(x)$ and Assumption 2.2(b) yield

$$\mathbb{E}_x^{\pi,\theta}[w(x(t))] \leq e^{-ct}w(x) + \frac{d}{c}(1 - e^{-ct}). \tag{2.4}$$

We now introduce the concept of the w-weighted norm, where w is the function in Assumption 2.2.

DEFINITION 2.5. Let $\mathcal{B}_w(\mathbb{R}^n)$ denote the Banach space of real-valued measurable functions v on $\mathbb{R}^n$ with

$$\|v\|_w := \sup_{x\in\mathbb{R}^n} \frac{|v(x)|}{w(x)} < \infty.$$

The reward rate. Let $r : \mathbb{R}^n \times U \times \Theta \to \mathbb{R}$ be a measurable function, which we call the *reward rate*, and which satisfies the following conditions:

ASSUMPTION 2.3.

(a) *The function* $r(x,u,\theta)$ *is continuous on* $\mathbb{R}^n \times U \times \Theta$ *and locally Lipschitz in* x *uniformly with respect to* $u \in U$ *and* $\theta \in \Theta$, *i.e., for each* $R > 0$, *there exists a constant* $K(R) > 0$ *such that*

$$\sup_{(u,\theta)\in U\times\Theta} |r(x,u,\theta) - r(y,u,\theta)| \leq K(R)|x-y| \quad \text{for } |x|, |y| \leq R.$$

(b) $r(\cdot,u,\theta)$ *is in* $\mathcal{B}_w(\mathbb{R}^n)$ *uniformly in* $u \in U$ *and* $\theta \in \Theta$, *that is, there exists* $M > 0$ *such that for all* $x \in \mathbb{R}^n$,

$$\sup_{(u,\theta)\in U\times\Theta} |r(x,u,\theta)| \leq Mw(x).$$

Similar to (2.3), for each $\pi \in \Pi$ we write the reward rate as

$$r(x,\pi,\theta) := \int_U r(x,u,\theta)\,\pi(du|x). \tag{2.5}$$

3. Discounted optimality criterion. In this section we give conditions for the existence of α-optimal policies and asymptotically optimal α-policies for the discounted optimality criterion defined as follows.

DEFINITION 3.1 (α-discount criterion). The expected payoff under the α-discount criterion when using the policy $\pi \in \Pi$, given the initial state $x(0) = x \in \mathbb{R}^n$ and the parameter value $\theta \in \Theta$, is

$$V(x,\pi,\theta) := \mathbb{E}_x^{\pi,\theta}\Big[\int_0^\infty e^{-\alpha t} r(x(t),\pi,\theta)\,dt\Big]. \tag{3.1}$$

As a consequence of Assumption 2.3(b) and (2.4), the discounted reward is finite-valued. In fact,

$$|V(x,\pi,\theta)| \le M\int_0^\infty e^{-\alpha t}\mathbb{E}_x^\pi[w(x)]\,dt \le M(\alpha)w(x) \tag{3.2}$$

with $M(\alpha) := M(\alpha+d)/(\alpha c)$. Here, c and d are as in Assumption 2.2, and M is the constant in Assumption 2.3(b). Hence, $V(\cdot,\pi,\theta)$ is in $\mathcal{B}_w(\mathbb{R}^n)$, and moreover the *optimal discounted reward*

$$V_\theta^*(x) := \sup_{\pi\in\Pi} V(x,\pi,\theta) \tag{3.3}$$

satisfies $|V_\theta^*(x)| \le M(\alpha)w(x)$. Thus $V_\theta^*(\cdot)$ is also in $\mathcal{B}_w(\mathbb{R}^n)$.

The discounted reward optimality equation. The following proposition gives a characterization of the discounted reward. For a proof see [26, Proposition 3.1.5]. The proof in [26] uses both Dynkin's formula and the inequality (2.4).

PROPOSITION 3.2. *Under Assumptions* 2.1–2.3 *the α-discount reward $V(\cdot,\pi,\theta)$ belongs to $\mathcal{W}^{2,p}(\mathbb{R}^n)\cap\mathcal{B}_w(\mathbb{R}^n)$ and, for every given $x\in\mathbb{R}^n$, $\pi\in\Pi$ and $\theta\in\Theta$,*

$$\alpha V(x,\pi,\theta) = r(x,\pi,\theta) + \mathcal{L}^{\theta,\pi}V(x,\pi,\theta). \tag{3.4}$$

Conversely, if a function $\varphi_\theta \in W^{2,p}(\mathbb{R}^n)\cap\mathcal{B}_w(\mathbb{R}^n)$ satisfies (3.4)*, then $\varphi_\theta(x) = V(x,\pi,\theta)$ for all $x\in\mathbb{R}^n$. If the equality in* (3.4) *is replaced by $\ge$ or $\le$, then $\varphi_\theta(x)\ge V(x,\pi,\theta)$ or $\varphi_\theta(x)\le V(x,\pi,\theta)$, respectively.*

DEFINITION 3.3. A policy $\pi^*\in\Pi$ is said to be *α-discount optimal*, given that $\theta\in\Theta$ is the true parameter value, if

$$V(x,\pi^*,\theta) = V_\theta^*(x) \quad \text{for all } x\in\mathbb{R}^n,$$

where $V_\theta^*(x) := \sup_{\pi\in\Pi} V(x,\pi,\theta)$ denotes the optimal α-discount reward.

The following proposition shows that the optimal discounted reward $V_\theta^*(\cdot)$ is a solution of a suitably defined optimality equation, and it also proves the existence of optimal stationary policies $f_\theta^*\in\mathbb{F}$.

PROPOSITION 3.4 ([8, 25, 24]). *Suppose that Assumptions* 2.1–2.3 *hold. Then:*

(i) *The optimal α-discount reward V_θ^* belongs to $\mathcal{W}^{2,p}(\mathbb{R}^n) \cap \mathcal{B}_w(\mathbb{R}^n)$ and it satisfies the discounted reward Hamilton–Jacobi–Bellman (HJB) equation: for all $x \in \mathbb{R}^n$ and $\theta \in \Theta$,*

$$\alpha V_\theta^*(x) = \sup_{u \in U}\{r(x,u,\theta) + \mathcal{L}^{\theta,u} V_\theta^*(x)\}. \tag{3.5}$$

Conversely, if a function $\varphi_\theta \in W^{2,p}(\mathbb{R}^m) \cap \mathcal{B}_w(\mathbb{R}^n)$ satisfies (3.5) then $\varphi_\theta(x) = V_\theta^*(x)$ for all $x \in \mathbb{R}^n$.

(ii) *There exists a stationary policy $f_\theta^* \in \mathbb{F}$ which, for each $x \in \mathbb{R}^n$, maximizes the right-hand side of (3.5), that is,*

$$\alpha V_\theta^*(x) = r(x, f_\theta^*, \theta) + \mathcal{L}^{\theta, f_\theta^*} V_\theta^*(x) \quad \text{for all } x \in \mathbb{R}^n, \tag{3.6}$$

and f_θ^* is α-discount optimal.

REMARK 3.5. (a) In the expression (3.5) we can write $\sup_{f\in\mathbb{F}}$ or $\sup_{\pi\in\Pi}$ instead of $\sup_{u\in U}$. For instance, (3.5) can be written as

$$\alpha V_\theta^*(x) = \sup_{\pi \in \Pi}\{r(x,\pi,\theta) + \mathcal{L}^{\theta,\pi} V_\theta^*(x)\} \tag{3.7}$$

(see [38]).

(b) Lemma 9.1 and Theorem 9.1 in [16] ensure that if $V_\theta^* \in \mathcal{W}^{2,p}(\mathbb{R}^n) \cap \mathcal{B}_w(\mathbb{R}^n)$ is a solution of the HJB equation (3.7), then $V_\theta^*(x) = \sup_{\Pi_D} V(x,\pi,\theta)$, where Π_D (or $\mathbb{F}_D$) is the class of stationary policies $\pi_\theta \in \Pi$ (or $f_\theta \in \mathbb{F}$) for which

$$\lim_{t\to\infty} \mathbb{E}_x^{\theta,\pi_\theta}[e^{-\alpha t} V_\theta^*(x(t))] = 0 \quad \forall x \in \mathbb{R}^n. \tag{3.8}$$

4. Adaptive control. To construct adaptive policies we will use the so-called *principle of estimation and control* [30] in which the unknown parameter θ in the control problem (2.1)–(3.1) is replaced by estimates θ_m, $m = 1, 2, \ldots$. Thus, for each θ_m, Proposition 3.4(ii) ensures the existence of an α-discount optimal stationary policy $f_{\theta_m} \in \mathbb{F}$. The policy $f_{\theta_m} \in \mathbb{F}$ is called an *adaptive policy*. Moreover, if for each θ_m, there exists a function $V_{\theta_m} \in \mathcal{W}^{2,p}(\mathbb{R}^n) \cap \mathcal{B}_w(\mathbb{R}^n)$ that satisfies the HJB equation (3.6) with f_{θ_m}, then $(f_{\theta_m}, V_{\theta_m}(\cdot))$ is an optimal pair for the *θ_m-control problem*. In this section we prove that the sequence $\{(f_{\theta_m}, V_{\theta_m}(\cdot))\}_{m\geq 1}$ of optimal pairs converges in some sense to an optimal pair $\{(f_\theta, V_\theta(\cdot))\}$ of the θ-control problem.

4.1. Convergence of stationary policies. As is well known [19, 21], if for each θ_m there is a *unique* α-optimal stationary policy f_{θ_m} as in Proposition 3.4(ii), then one usually gets pointwise convergence $f_{\theta_m}(x) \to f(x)$ for all $x \in \mathbb{R}^n$. This is the case, for example, in LQ problems, that is, linear systems with quadratic costs. In general, however, this uniqueness property does not hold, and so the pointwise convergence of control policies may not make sense. We consider convergence in the sense of Schäl [40] (see [19,

Lemma 6.5]), as well as convergence of stationary randomized controls (or relaxed controls) [3, 15, 44].

The topology on the space Π of relaxed controls, which is determined by the convergence criterion in the following definition, renders Π a compact metric space (see [44] for instance).

DEFINITION 4.1. We say that a sequence $\{\pi_m\}$ in Π *converges to* $\pi \in \Pi$, and we write $\pi_m \xrightarrow{w} \pi$, if

$$\int_{\mathbb{R}^n} g(x)h(x,\pi_m)\,dx \to \int_{\mathbb{R}^n} g(x)h(x,\pi)\,dx \tag{4.1}$$

for all $g \in L^1(\mathbb{R}^n)$ and $h \in C_b(\mathbb{R}^n \times U)$, where $h(x,\pi_m) := \int_U h(x,u)\,\pi_m(du|x)$ and $h(x,\pi) := \int_U h(x,u)\,\pi(du|x)$.

In addition to (4.1), we will use another notion of convergence of π_m to π.

DEFINITION 4.2. A sequence $\{\pi_m\} \subset \Pi$ converges in the sense of Schäl to $\pi \in \Pi$ if, for each $x \in \mathbb{R}^n$, there is a subsequence $m_k \equiv m_k(x)$ of $\{m\}$ such that $\pi_{m_k}(\cdot|x) \to \pi(\cdot|x)$ as $k \to \infty$ in the topology of weak convergence in $\mathcal{P}(U)$, that is, for each $h \in C_b(U)$, $\int_U h(u)\,\pi_{m_k}(du|x) \to \int_U h(u)\,\pi(du|x)$.

REMARK 4.3. (a) Suppose that Assumptions 2.1–2.3 hold. Then Proposition 3.4 in [24] establishes that for fixed $\alpha > 0$ and $\theta \in \Theta$, the mapping $\pi \mapsto V(x,\pi,\theta)$ is continuous with respect to the topology of relaxed controls. The proof in [24] uses a theorem similar to our Theorem 9.1 below without considering the θ parameter.

(b) Theorem 2.4.2 of [3] or Lemma 3.5 of [7] ensure that under the topology of relaxed controls the solutions to (2.1) depend continuously on the controls in the sense that if $\theta_m \to \theta$, then

$$\|p_{\theta_m}^{\pi_{\theta_m}}(t,x,\cdot) - p_\theta^{\pi_\theta}(t,x,\cdot)\|_{L^1(\mathbb{R}^n)} \to 0 \quad \text{as } m \to \infty,$$

where $p_{\theta_m}^{\pi_{\theta_m}}(t,x,y)$ and $p_\theta^{\pi_\theta}(t,x,y)$ denote the transition probability densities for $x^{\theta_m,\pi_{\theta_m}}(\cdot)$ and $x^{\theta,\pi_\theta}(\cdot)$, respectively. So, $x^{\theta_m,\pi_{\theta_m}}(\cdot) \to x^{\theta,\pi}(\cdot)$ in law as $\theta_m \to \theta$; in addition, $\pi_{\theta_m} \xrightarrow{w} \pi_\theta$. Moreover, $x^{\theta,\pi_\theta}(\cdot)$ is a diffusion satisfying (2.1) with true parameter θ for some Markov control π_θ.

DEFINITION 4.4. A sequence $\{\theta_m\}_{m\geq 1}$ of measurable functions $\theta_m : \Omega \to \Theta$ is said to be a sequence of *uniformly strongly consistent* (USC) *estimators* of $\theta \in \Theta$ if as $m \to \infty$,

$$\theta_m(\omega) \to \theta \quad \mathbb{P}_\theta^\pi\text{-a.s. for all } \pi \in \Pi.$$

For simplicity of notation, we write $\theta_m := \theta_m(\omega) \in \Theta$.

4.2. Optimal discounted rewards. We will show that, as $\theta_m \to \theta$, the optimal discounted rewards $V_{\theta_m}^*(\cdot)$ for the θ_m-control problem converge

almost surely to the optimal discounted reward $V_\theta^*(\cdot)$ for the θ-control problem, and $\pi_{\theta_m} \to \pi \in \Pi$ in the relaxed controls topology, and also in Schäl's sense.

THEOREM 4.5. *Let $\{\theta_m\} \subset \Theta$ be a sequence of USC estimators of θ. For each m, let π_{θ_m} be a maximizer of the right-hand side of* (3.7) *given that $\theta = \theta_m$. Then there exists a subsequence $\{m_k\}$ of $\{m\}$ and a policy π^* such that $\pi_{m_k} \xrightarrow{w} \pi^*$ in the relaxed controls topology. Moreover, under the assumptions of Proposition* 3.2*, as $m \to \infty$,*

$$V_{\theta_m}^*(x) \to V_\theta^*(x) \quad \mathbb{P}_\theta^\pi\text{-a.s. for each } x \in \mathbb{R}^n. \tag{4.2}$$

Proof. The proof is based on Theorem 9.1 (in the Appendix). Consider a sequence of USC estimators $\theta_m \in \Theta$ such that $\theta_m \to \theta$ as $m \to \infty$. Let $R > 0$, and take the open ball $B_R := \{x \in \mathbb{R}^n \mid |x| < R\}$. For $x \in B_R$ and each $m = 1, 2, \ldots$, let $\pi_{\theta_m} \in \Pi$ be a maximizer of the right-hand side of (3.7). Note that the equality $r(x, \pi_{\theta_m}, \theta_m) + \mathcal{L}^{\theta_m, \pi_{\theta_m}} V_{\theta_m}^*(x) - \alpha V_{\theta_m}^*(x) = 0$ can be expressed in terms of the operator (9.3) as

$$L^{\theta_m, \pi_{\theta_m}} V_{\theta_m}^*(x) = r(x, \pi_{\theta_m}, \theta_m) + \mathcal{L}^{\theta_m, \pi_{\theta_m}} V_{\theta_m}^*(x) - \alpha V_{\theta_m}^*(x) = 0. \tag{4.3}$$

We will now proceed to verify the hypotheses (a)–(e) of Theorem 9.1, with $\mathcal{O} = B_R$. By (4.3), both hypotheses (a) and (c) hold with $\xi_m \equiv \xi = 0$. Moreover, the assumptions of our Theorem 4.5 ensure that hypotheses (d) and (e) also hold. On the other hand, by [18, Theorem 9.11], there exists a constant C_0 depending on R such that, for a fixed $p > n$ (n being the dimension of (2.1)), we have

$$\begin{aligned} \|V_{\theta_m}^*(\cdot)\|_{W^{2,p}(B_R)} &\le C_0\big(\|V_{\theta_m}^*(\cdot)\|_{L^p(B_{2R})} + \|r(\cdot, \pi_{\theta_m}, \theta_m)\|_{L^p(B_{2R})}\big) \\ &\le C_0\big(\|V_{\theta_m}^*(\cdot)\|_{L^p(B_{2R})} + M\|w\|_{L^p(B_{2R})}\big) \\ &\le C_0\Big(\|V_{\theta_m}^*(\cdot)\|_{L^p(B_{2R})} + |\bar{B}_{2R}|^{1/p} \max_{x \in \bar{B}_{2R}} w(x)\Big) < \infty, \end{aligned} \tag{4.4}$$

where $|\bar{B}_{2R}|$ represents the volume of the closed ball with radius $2R$ and M is the constant in Assumption 2.3(b). This implies the hypothesis (b) in Theorem 9.1.

Now note that Theorem 9.1 ensures the existence of a function $V \in W^{2,p}(B_R)$ together with a subsequence $\{m_k\}$ such that $V_{\theta_{m_k}}^* \to V$ uniformly in B_R and pointwise on $\mathbb{R}^n$ as $k \to \infty$ and $\pi_{m_k} \xrightarrow{W} \pi^*$. Furthermore, V satisfies

$$\alpha V(x) = r(x, \pi^*, \theta) + \mathcal{L}^{\theta, \pi^*} V(x) \mathbb{P}_\theta^{\pi^*}\text{-a.s.} \tag{4.5}$$

Since R was arbitrary, the convergence (4.5) holds for all $x \in \mathbb{R}^n$. Thus, Proposition 3.2 asserts that $V(x)$ actually coincides with $V(x, \pi^*, \theta)$.

To complete the proof of Theorem 4.5 we will next show that the function $V(\cdot)$ in (4.5) coincides with $V_\theta^*(\cdot)$. To this end, observe from (3.7) that

$$\alpha V^*_{\theta_{m_k}}(x) \geq r(x,\pi,\theta) + \mathcal{L}^{\theta,\pi} V^*_{\theta_{m_k}}(x) \quad \forall \pi \in \Pi.$$

Taking the limit as $m \to \infty$ and using the same arguments as above, we obtain

$$\alpha V(x) \geq r(x,\pi,\theta) + \mathcal{L}^{\theta,\pi} V(x) \quad \forall \pi \in \Pi.$$

From this inequality and (4.5) it follows that

$$\alpha V(x) = \sup_{\pi\in\Pi}\{r(x,\pi,\theta) + \mathcal{L}^{\theta,\pi} V(x)\} \quad \mathbb{P}_\theta^{\pi^*}\text{-a.s.}$$

Therefore, by (3.7), $V(\cdot) = V_\theta^*(\cdot)$. In other words, as $k \to \infty$,

$$V^*_{\theta_{m_k}}(x) \to V_\theta^*(x) \quad \mathbb{P}_\theta^{\pi}\text{-a.s.} \ \blacksquare$$

Remark 4.6. Theorem 4.5 holds if convergence in the sense of Schäl is used instead of convergence in the topology of relaxed controls.

4.3. Stationary policies. We will next consider again a sequence $\{\theta_m\}$ of USC estimators of θ, and the corresponding sequence of stationary optimal policies $\{\pi_{\theta_m}\}$ for the θ_m-control problem. We will show that $\{\pi_{\theta_m}\}$ (or a subsequence thereof) converges in the topology of relaxed controls (Definition 4.1) and also in the sense of Schäl (Definition 4.8) to a stationary policy π_θ that is optimal for the θ-control problem.

Theorem 4.7. *Suppose that Assumptions* 2.1–2.3 *are satisfied, and let* $\{\theta_m\} \subset \Theta$ *be a sequence of USC estimators of* θ. *For each* m, *let* π_{θ_m} *be a maximizer of the right-hand side of* (3.7) *given that* $\theta = \theta_m$. *Then there exists a subsequence* $\{m_k\}$ *of* $\{m\}$ *and a policy* π^* *such that* $\pi_{m_k} \xrightarrow{w} \pi^*$, *and moreover* π^* *is optimal for the* θ*-control problem* $\mathbb{P}_\theta^{\pi^*}$*-a.s.*

Proof. As was already noted, when endowed with the topology of relaxed controls, Π becomes a compact metric space. Therefore, the sequence $\{\pi_{\theta_m}\}$ has a subsequence $\{\pi_{\theta_{m_k}}\}$ that converges to some $\pi^* \in \Pi$. Now, to prove the optimality of the policy π^* for the α-discount criterion observe that, for each m_k, the policy $\pi_{\theta_{m_k}}$ is a maximizer of (3.7), so

$$\alpha V^*_{\theta_{m_k}}(x) = r(x,\pi_{\theta_{m_k}},\theta_{m_k}) + \mathcal{L}^{\theta_{m_k},\pi_{\theta_{m_k}}} V^*_{\theta_{m_k}}(x). \tag{4.6}$$

Now observe that, for $x \in B_R$, we can express (4.6) in terms of the operator $L^{\theta,\pi}$ in (9.3) (in the Appendix) as

$$L^{\theta_{m_k},\pi_{\theta_{m_k}}} V^*_{\theta_{m_k}}(x) = 0, \tag{4.7}$$

with $h_{m_k} \equiv V^*_{\theta_{m_k}}$, $\pi_m \equiv \pi_{\theta_{m_k}}$ and $\xi_m \equiv 0$. Consequently, as in the proof of Theorem 4.5, we can invoke Theorem 9.1 to see that, as $k \to \infty$, (4.7)

converges to

$$\alpha V_\theta^*(x) = r(x,\pi^*,\theta) + \mathcal{L}^{\theta,\pi^*}V_\theta^*(x) \quad \text{for } x \in B_R,\ \mathbb{P}_\theta^{\pi^*}\text{-a.s.} \tag{4.8}$$

Furthermore, since R was arbitrary, we conclude that (4.6) converges to (4.8) for all $x \in \mathbb{R}^n$.

On the other hand, from (3.7), we obtain, for each $k \geq 1$ and $x \in \mathbb{R}^n$,

$$\alpha V_{\theta_{m_k}}^*(x) \geq r(x,\pi,\theta_{m_k}) + \mathcal{L}^{\theta_{m_k},\pi}V_{\theta_{m_k}}^*(x) \quad \text{for all } \pi \in \Pi. \tag{4.9}$$

Hence, taking $k \to \infty$ and using Theorem 9.1 again, we have

$$\alpha V_\theta^*(x) \geq r(x,\pi,\theta) + \mathcal{L}^{\theta,\pi}V_\theta^*(x) \quad \text{for all } \pi \in \Pi. \tag{4.10}$$

Thus, by (4.8) and (4.10), we have

$$\alpha V_\theta^*(x) = \sup_{\pi\in\Pi}\{r(x,\pi,\theta) + \mathcal{L}^{\theta,\pi}V_\theta^*(x)\}, \tag{4.11}$$

implying that π^* is optimal for the θ-control problem. ∎

PROPOSITION 4.8. *Let $\{\pi_{\theta_m}\} \subset \Pi$ be a sequence of maximizers of the right-hand side of* (3.7) *given that $\theta = \theta_m$. If Assumptions* 2.1–2.3 *hold, then there exists $\pi^* \in \Pi$ such that π^* is the limit in the sense of Schäl of $\{\pi_{\theta_m}\}$.*

Proof. Fix an arbitrary $x \in \mathbb{R}^n$. Since $\mathcal{P}(U)$ is compact, $\{\pi_{\theta_m}(\cdot|x)\}$ has a subsequence $\{\pi_{\theta_{m_k(x)}}(\cdot|x)\}$ that converges to some $\pi^*(\cdot|x) \in \mathcal{P}(U)$. Furthermore, for all $B \subset \mathcal{B}(U)$, by [39, Lemma 4], $\pi^*(B|x)$ is measurable on $x \in \mathbb{R}^n$. So, $\pi^*(\cdot|x)$ is in Π. This proves the result. ∎

REMARK 4.9. Recall that for $x \in \mathbb{R}^n$, a stationary randomized policy $\pi(\cdot|x)$ is a probability measure in $\mathcal{P}(U)$. Under the Assumptions 2.1(a) and 2.3(a, b) the functions $r(x,\cdot,\theta)$ and $b(x,\cdot,\theta)$ are continuous and attain its supremum in U for each $x \in \mathbb{R}^n$ and $\theta \in \Theta$. Hence, by the definition of weak convergence on $\mathcal{P}(U)$, the infinitesimal generator $\mathcal{L}^{\theta,\pi(\cdot|x)}$ and the reward rate $r(x,\pi(\cdot|x),\theta)$ are continuous on $\mathcal{P}(U)$.

PROPOSITION 4.10. *Let $\{\pi_{\theta_m}\} \subset \Pi$ be a sequence of maximizers of* (3.7) *such that π_{θ_m} converges to $\pi^* \in \Pi$ in Schäl's sense. If Assumptions* 2.1–2.3 *hold, then π^* is an α-discount optimal policy for the θ-control problem $\mathbb{P}_\theta^{\pi}$-a.s.*

Proof. This follows by the same arguments used in the proof of Theorem 4.6, but using convergence in the sense of Schäl rather than convergence in the topology of relaxed controls. ∎

5. Estimation methods. The solution $x^{\theta,u}(\cdot)$ of (2.1) induces for each $\theta \in \Theta$ a probability measure $\mathbb{P}_\theta^u$ on the space $C([0,\infty),\mathbb{R}^n)$ of continuous trajectories from $[0,\infty)$ into $\mathbb{R}^n$ endowed with its Borel σ-algebra $\mathcal{F}$. In practice, $x^{\theta,u}(t)$ can only be observed up to a finite time, say T. Therefore, choose T as large as practically possible with respect to computation time, computer power, measurement instruments, etc.

In this work, we consider maximum likelihood estimation for stochastic differential equations based on finitely many discrete observations $X_T := \{x_{t_i} \mid 0 \le i \le m\}$ of a trajectory $\{x^{\theta,u}(t) \mid t \in [0,T]\}$ at times $0 = t_0 < t_1 < t_m := T$ when $\mathbb{P}_\theta^u$ is not known. Namely, the parameters will be estimated using the *discrete approximate likelihood ratio function* $\mathrm{MLR}(X_T,\theta)$ [34], which is defined as

$$\begin{aligned}(5.1)\quad \mathrm{MLR}(X_T,\theta) := {} & \sum_{i=1}^{m} b(x_{t_{i-1}},u_{t_{i-1}},\theta)[\sigma(x_{t_{i-1}})\sigma(x_{t_{i-1}})^T]^{-1}(x_{t_i}-x_{t_{i-1}}) \\ & - \frac{1}{2}\sum_{i=1}^{m}\{b(x_{t_{i-1}},u_{t_{i-1}},\theta)^T[\sigma(x_{t_{i-1}})\sigma(x_{t_{i-1}})^T]^{-1} \\ & \qquad \cdot b(x_{t_{i-1}},u_{t_{i-1}},\theta)(t_i-t_{i-1})\},\end{aligned}$$

with b and σ as in (2.1). The MLR function generates the *discrete approximate likelihood ratio estimator*, θ_{LR},

$$(5.2)\qquad \theta_{LR} \equiv \theta_{LR}(X_T) := \mathrm{Argmax}_{\theta\in\Theta}\,\mathrm{MLR}(X_T,\theta).$$

Shoji [42] shows that when a one-dimensional stochastic differential equation with a constant diffusion coefficient is considered, optimization based on the MLR function is equivalent to optimization based on the *least square function* $\mathrm{LSE}(X_T,\theta)$,

$$(5.3)\qquad \mathrm{LSE}(X_T,\theta) := \sum_{i=1}^{m}(x_{t_i}-x_{t_{i-1}}-b(x_{t_{i-1}},u_{t_{i-1}},\theta)(t_i-t_{i-1}))^2.$$

In this case, the *least square estimator* θ_{LSE} is given by

$$(5.4)\qquad \theta_{\mathrm{LSE}} \equiv \theta_{\mathrm{LSE}}(X_T) := \mathrm{Argmin}_{\theta\in\Theta}\,\mathrm{LSE}(X_T,\theta).$$

Consistency and asymptotic normality of θ_{LSE} and θ_{LR} are studied in [34, 35, 36, 37, 42].

Remark 5.1. (a) The log-likelihood function for θ based on continuous observations of $x^{\theta,u}(t)$ in the time interval $[0,T]$ when the transition densities of $x^{\theta,u}(t)$ are unknown is

$$\begin{aligned}(5.5)\qquad \mathrm{MLR}(X_T,\theta) := {} & \int_0^T b(x(t),u(t),\theta)^T[\sigma(x(t))\sigma(x(t))^T]^{-1}\,dx(t) \\ & - \frac{1}{2}\int_0^T b(x(t),u(t),\theta)^T[\sigma(x(t))\sigma(x(t))^T]^{-1}b(x(t),u(t),\theta)\,dt\end{aligned}$$

(see [34]). The approximation of the integrals in (5.5) by finite Itô and Riemann sums, respectively, leads to the approximate likelihood function (5.1).

(b) In [34, 35, 36, 42] the differential $dx(t)$ in (5.5) is approximated by a backward difference formula $x_{t_i} - x_{t_{i-1}}$. However, in our work this differential

will be replaced by the central difference $dx(t) \approx \frac{x_{t_{i+1}} - x_{t_{i-1}}}{2(t_{i+1} - t_{i-1})}$ since it yields a more accurate approximation.

(c) The discrete approximate likelihood ratio estimator will be used in this work since we will assume that the diffusion matrix $\sigma(x(t))$ is either constant or of the form $\sigma(x(t)) = \theta\hat{\sigma}(x(t))$. In cases where the diffusion matrix depends on θ in a more general way, we recommend the reader to try several estimation methods and select the one that best suits his/her application.

(d) In the applications presented below we consider linear dynamic systems and quadratic costs. Therefore, Assumptions 2.1 and 2.3(a) are obviously satisfied. To prove Assumptions 2.2 and 2.3(b), use the function $w(x) := x^2 + 1$.

(e) The root mean square error (RMSE) will be used to measure the differences between variables predicted by the estimators and the real values.

6. Simulation study. Consider the one-dimensional diffusion process $x^{\theta,u}(t)$ defined as the solution of the stochastic differential equation with unknown parameter θ,

$$dx(t) = -(u(t) + \theta x(t))dt + \sigma dW(t), \tag{6.1}$$

with $\sigma > 0$ and $W(t)$ a standard Brownian motion. We assume that there are no control constraints, so $U = \mathbb{R}$, and $\theta \in [0, \infty) =: \Theta$. Let

$$V(x, u, \theta) = \mathbb{E}_x^{\theta,u}\Big[\int_0^\infty e^{-\alpha t}\tfrac{1}{2}[x^2(t) + \lambda(\theta)u^2(t)]\,dt\Big] \tag{6.2}$$

be the α-discount reward, with $\lambda : \Theta \to \mathbb{R}$, positive continuously differentiable function.

The objective of the θ-control problem is to design an admissible control process so that the α-discounted cost (6.2) is minimized. When all the parameters in (6.1)–(6.2) are known, Proposition 3.4(i) ensures that the optimal discounted cost $V_\theta^*(x) = \inf_{\pi \in \Pi} V(x, \pi, \theta)$ is a solution of the HJB equation (3.5), which in view of (6.1)–(6.2) becomes

$$\alpha V_\theta^*(x) = \min_{u \in U}\big\{\tfrac{1}{2}x^2 + \tfrac{1}{2}\lambda(\theta)u^2 + (-u - \theta x)\partial_x V_\theta^*(x) + \tfrac{1}{2}\sigma^2\partial_{xx}^2 V_\theta^*(x)\big\}. \tag{6.3}$$

To solve (6.3), we propose a solution $V_\theta \in C^2(\mathbb{R}) \cap \mathcal{B}_w(\mathbb{R})$ of the form

$$V_\theta(x) = \tfrac{1}{2}\lambda(\theta)h(\theta)x^2 + K, \tag{6.4}$$

where $h : \Theta \to \mathbb{R}$ is a nonnegative continuous function on Θ and K is a constant, both to be determined. Then the derivatives of V_θ are given by $\partial_x V_\theta(x) = \lambda(\theta)h(\theta)x$ and $\partial_{xx}^2 V_\theta(x) = \lambda(\theta)h(\theta)$. Hence, since $h \geq 0$, the function $x \mapsto V_\theta(x)$ is convex. So, substituting the derivatives of V_θ in (6.3),

we see that the minimizer f_θ^* of the HJB equation is

$$f_\theta^*(x(t)) = h(\theta)x(t). \tag{6.5}$$

Inserting the derivatives of V_θ and the minimizer f_θ^* in (6.3), we find that h is the nonnegative solution of the second order equation $-h^2(\theta) - (2\theta+\alpha)h(\theta) + \frac{1}{\lambda(\theta)} = 0$ and $K = \lambda(\theta)h(\theta)/\alpha$, that is,

$$h(\theta) = \frac{(2\theta+\alpha) - \sqrt{(2\theta+\alpha)^2 - 4(-1)/\lambda(\theta)}}{-2}.$$

REMARK 6.1. (a) Take $u(t) = f_\theta^*(x(t))$ in (6.1), to obtain $dx(t) = -p(\theta)x(t)dt + \sigma dW(t)$, with $p(\theta) = h(\theta) + \theta$, which is the so-called Langevin equation. As is well known (see, for instance, [4, Section 8.3]), the solution of the Langevin equation is

$$x(t) = xe^{-p(\theta)t} + \sigma \int_0^t e^{-p(\theta)(t-s)}\, dW(s).$$

Therefore, by the properties of stochastic integrals,

$$\mathbb{E}_x^{\theta, f_\theta^*}[e^{-\alpha t}x^2(t)] = \left[x^2 - \frac{\sigma^2}{2p(\theta)}\right]e^{-(\alpha+2p(\theta))t} + \frac{\sigma^2}{2p(\theta)}e^{-\alpha t}.$$

Finally, from the definitions of $p(\theta)$ and $h(\theta)$, we have $\alpha + 2p(\theta) = \alpha + 2h(\theta) + 2\theta > 0$. Hence

$$\lim_{t\to\infty} \mathbb{E}_x^{\theta, f_\theta^*}[e^{-\alpha t}x^2(t)] = 0, \quad \text{so} \quad \lim_{t\to\infty} \mathbb{E}_x^{\theta, f_\theta^*}[e^{-\alpha t}V_\theta(x(t))] = 0 \quad \forall x \in \mathbb{R}^n.$$

This last result implies that f_θ^* satisfies (3.8). Therefore, we conclude that f_θ^* minimizes (6.3) within the class $\mathbb{F}_D$ of admissible stationary policies.

(b) The above results are easily extended to n-dimensional linear systems with quadratic cost $x^T(t)\cdot Qx(t) + u^T(t)Ru(t)$ with Q and R being symmetric matrices, Q nonnegative definite, and R positive definite. This is the case of the applications presented in Sections 7 and 8.

6.1. Numerical results. In order to implement the optimal control law (6.5) we estimate the unknown parameter with the least likelihood function LSE in (5.3)–(5.4). Replacing the discrete version of $b(x,u,\theta) = -u - \theta x$ in (5.3) a straightforward calculation gives the following estimator:

$$\theta_{\mathrm{LSE}_m} := \frac{\sum_{i=1}^m x_{t_i}\left(\frac{x_{t_{i+1}} - x_{t_{i-1}}}{2(t_{i+1}-t_{i-1})}\right) - \sum_{i=1}^m u_{t_i}x_{t_i}}{\sum_{i=1}^m x_{t_i}^2}.$$

To illustrate our results, let us assume that in the θ-optimal control problem (6.1)–(6.2) the true parameter value is $\theta = 2$, $u(t) = 8$, $\alpha = 0.1$, and λ is a linear function, say, $\lambda(\theta) = \theta + 1$ for $\theta \in [0, \infty)$. We next obtain discrete observations of the stochastic differential equation (6.1) simulating the equa-

tion by the Euler–Maruyama method in the interval $[0, 1.5]$. Based on this information, we obtained $m = 4097$ observations varying $\sigma = 0.1, 0.01, 0.001$.

Table 1 shows the information on the different values of θ_{LSE_m} for $m = 820, 1025, 1366, 2049$ and 4097. As can be seen, as m increases the estimator approaches the true parameter value $\theta = 2$. In Table 1 we can also see that the RMSE between the predicted process $x(t)^{\theta_{\mathrm{LSE}_m}, f_{\theta_{\mathrm{LSE}_m}}}$ and the real process $x(t)^{\theta, f_\theta}$ decreases as the number of data increases, implying a good fit. See Figure 1. The diffusion process (6.1) with $\sigma = 0.001$ showed a best fit because with 4097 data $\theta_{\mathrm{LSE}_m} = 1.9996$ and its RMSE is 0.0003, which suggests that the lower the noise in the measured data, the more accurate is the least square estimator.

Table 1. RMSE between $x(t)^{\theta_{\mathrm{LSE}_m}, f_{\theta_{\mathrm{LSE}_m}}}(\cdot)$ and $x^{\theta, f_\theta}(\cdot)$

	$\sigma = 0.1$		$\sigma = 0.01$		$\sigma = 0.001$	
m	θ_{LSE_m}	RMSE	θ_{LSE_m}	RMSE	θ_{LSE_m}	RMSE
4097	1.9794	0.02045	1.9982	0.0018	1.9996	0.0003
2049	2.2639	0.2399	2.2808	0.2543	2.2821	0.2554
1366	2.3588	0.3180	2.3750	0.3307	2.3763	0.3318
1025	2.4062	0.3546	2.4221	0.3675	2.4234	0.3685
820	2.4347	0.3757	2.4504	0.3889	2.4516	0.3899

Table 2 shows the information on the RMSE between the predicted optimal control $f_{\theta_{\mathrm{LSE}_m}}(x^{\theta_{\mathrm{LSE}_m}}(t))$ and the real optimal control $f_\theta((x^\theta(t)))$ de-

Table 2. RMSE between the estimated processes and the real processes ($\theta = 2$)

	$\sigma = 0.1$			$\sigma = 0.01$		
m	θ_{LSE_m}	RMSE_f	RMSE_V	θ_{LSE_m}	RMSE_f	RMSE_V
4097	1.9794	0.0438	2.2246	1.9982	0.0039	0.1951
2049	2.2639	0.6352	40.9422	2.2808	0.5067	31.6512
1366	2.3588	0.4766	29.6627	2.3750	0.6603	42.8962
1025	2.4062	0.7096	46.6922	2.4221	0.7358	48.6476
820	2.4347	0.7499	50.1668	2.4504	0.7806	52.1409
	$\sigma = 0.001$					
m	θ_{LSE_m}	RMSE_f	RMSE_V			
4097	1.9996	0.0006	0.0382			
2049	2.2821	0.5083	31.8036			
1366	2.3763	0.6624	43.0529			
1025	2.4234	0.7374	48.8016			
820	2.4516	0.7824	52.2962			

noted by RMSE_f. Table 2 also shows the RMSE between the predicted optimal discount cost $V_{\theta_{\mathrm{LSE}_m}}(x)$ and the real optimal discount cost $V^*_\theta(x)$ (RMSE_V). As can be seen in Figure 2, as m increases, both errors decrease and the estimator approaches the true parameter value $\theta = 2$.

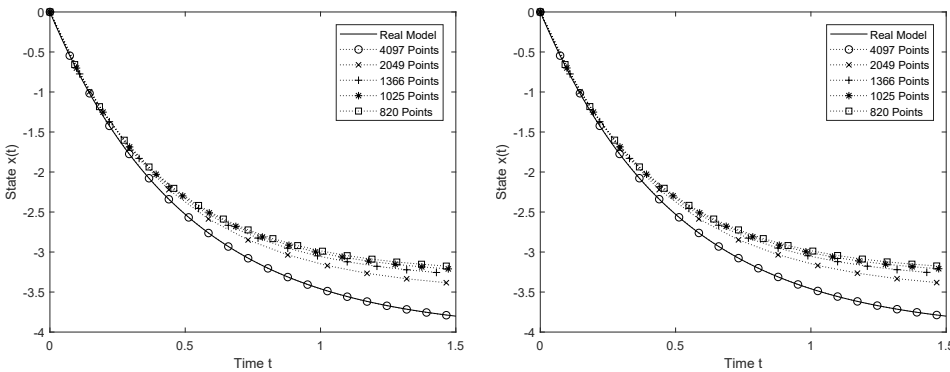

Fig. 1. Asymptotic behavior of $x^{\theta_{\mathrm{LSE}_m}, f_{\theta_{\mathrm{LSE}_m}}}(t)$ with $\sigma = 0.01$ (left) and $\sigma = 0.001$ (right)

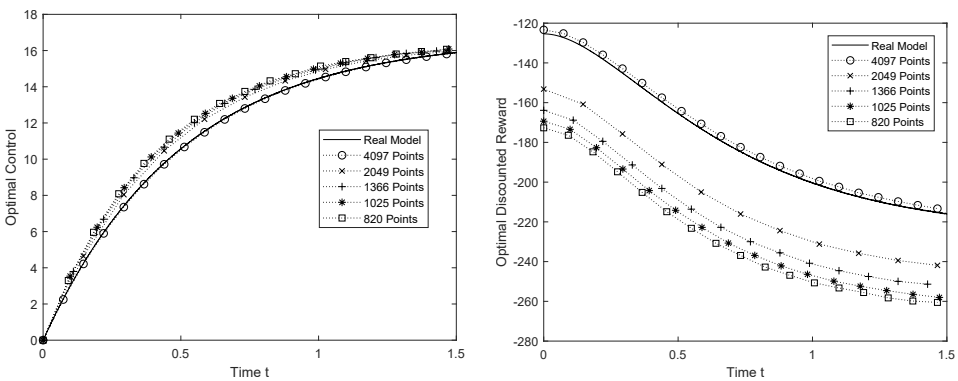

Fig. 2. Asymptotic behavior of the predicted optimal control and optimal discount cost for $\sigma = 0.1$ and $\alpha = 0.1$

7. Angular position and angular velocity from a DC motor. The angular acceleration in a DC motor evolves according to the third-order differential equation

$$V_{\mathrm{in}} = \frac{JL}{K_T}\frac{d^3w}{dt} + \frac{RJ}{K_T}\frac{d^2w}{dt^2} + K_b\frac{dw}{dt}, \tag{7.1}$$

where V_{in} is the supply voltage of the rotor, R is the winding resistance of the rotor, L is the inductance of the winding of the rotor, dw/dt the angular

velocity of rotation of the rotor, J the moment of inertia of the rotor shaft, w the angular position, and K_T, K_b proportionality constants. For many motors, the inductance can be neglected ($L \approx 0$). This implies that (7.1) can be rewritten in matrix form as

$$\frac{d^2w}{dt^2} = -\frac{K_b K_T}{RJ}\frac{dw}{dt} + \frac{K_T V_{\text{in}}}{RJ}, \quad \text{so} \quad \frac{dx(t)}{dt} = Ax(t) + b, \tag{7.2}$$

where

$$A = \begin{bmatrix} 0 & 1 \\ 0 & \frac{-K_T K_b}{RJ} \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ \frac{K_T V_{\text{in}}}{RJ} \end{bmatrix}, \quad x(t) := \begin{bmatrix} w(t) \\ \frac{dw(t)}{dt} \end{bmatrix}.$$

Experimental procedure to obtain discrete observations of the $x(t)$ process. A motor with Encoder CN5003-6006 was used to carry out the experiment. The data description for this motor are: 32 lines of code, 2 output channels: A and B, approximate weight: 25 grm, voltage: 6 V/12 V (4000/8000 RPM), current: 30 mA, arrow diameter: 2 mm, measure: 15 mm thick × 21 mm wide × 64 mm total length including auger and card. The electrical variables values of the motor are $R = 0.07\,\Omega$, $J = 2.05932971\text{e}$ $-06\,\text{kg}\cdot\text{m}^2$, $K_T = 0.002559499\,\text{N}\cdot\text{mA}^{-1}$, $K_b = K_T V\,(\text{Rad}\cdot\text{s}^{-1})^{-1}$, $V_{\text{in}} = 8.5\,\text{V}$. The experiment consists in connecting the direct current (DC) source to the DC motor, measuring the angular position $w(t)$, and calculating the angular velocity $dw(t)/dt$. This process was repeated 20 times with a duration of $T = 0.25$ seconds. To obtain the measurements a Compact Rio 9068, a digital inputs module 9375, an encoder connected to the motor shaft and the Labview software were used.

The most currently used DC motors are connected switched sources. These sources are more efficient than regulated sources, but generate high frequency noise. Trying to capture the noise affecting the system of the DC motor, we add a "white noise" in the differential equation (7.2) to obtain the SDE

$$dx(t) = (Ax(t) + b)dt + \sigma\xi(t)dt, \tag{7.3}$$

where $\xi(t)$ is the white noise of mean zero and variance one, and σ is a positive constant known as the noise intensity. Its magnitude determines the deviation of the stochastic case from the deterministic one. Formally, the white noise $\xi(t)$ is the time derivative in a distributional sense of Brownian motion $W(t)$, that is, $dW(t) = \xi(t)dt$. Therefore, strictly speaking, equation (7.3) can be rewritten as a controlled stochastic differential equation in terms of the two-dimensional standard Brownian motion, i.e.,

$$dx(t) = (Ax(t) + Bu(t))dt + \sigma dW(t), \tag{7.4}$$

where

$$B = \begin{bmatrix} 0 \\ \frac{K_T}{RJ} \end{bmatrix}, \quad \sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_1 \end{bmatrix},$$

σ_1 is a positive constant and $u(t) := V_{\text{in}}$ is the control.

Table 3. RMSE between the experimental data and the analytical solutions of (7.2) and (7.4)

Analytical solutions vs experimental measurements	RMSE w (Rev)	RMSE dw/dt (Rev/min)
ODE solution vs Experimental	1.8754	1.3257
SDE solution vs Experimental	1.8765	1.3258

Taking $\sigma_1 = 1$, Table 3 shows the RMSE between the experimental data (measurements) and the analytical solutions of the ODE (7.2) and the SDE (7.4). These RMSE show that the SDE (7.4) proposed in this work presents a good fit to the experimental data (real system). See Figure 3.

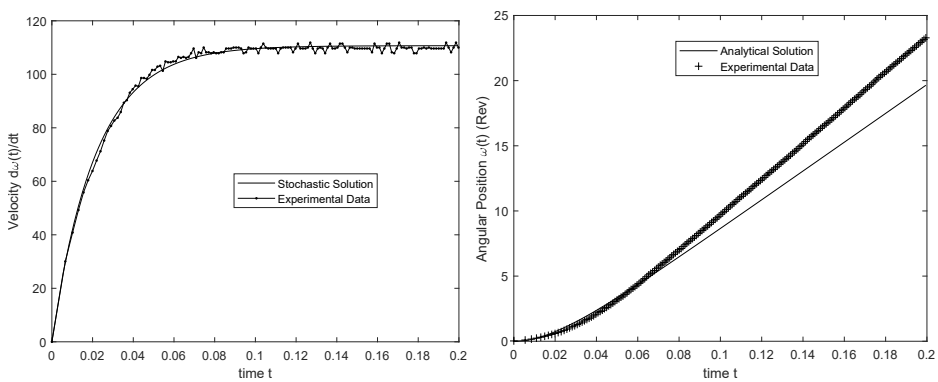

Fig. 3. Asymptotic behavior of w (right) and dw/dt (left) in (7.4) and experimental data

(θ_1, θ_2)-control problem (real model). In this application we assume that $\theta_1 := \frac{-k_T K_b}{RJ}$ and $\theta_2 := \frac{K_T}{RJ}$ are the unknown parameters. This implies that the pair of matrices $(A, B) := (A(\theta_1), B(\theta_2))$ in the stochastic differential equation (7.4) are unknown. Once the stochastic differential equation for the system under study is proposed, the objective of the adaptive control problem is to design an admissible control process so that the following α-discounted cost for system (7.4) is minimized:

$$V(x, u, \theta_1, \theta_2) = \mathbb{E}_x^{\theta_1, \theta_2, u}\Big[\int_0^\infty e^{-\alpha t}[x(t)^T Q x(t) + u(t)^T R u(t)]\, dt\Big] \tag{7.5}$$

with Q a positive definite 2×2 matrix, and $R > 0$ a constant. The HJB

equation associated to the optimal control problem (7.4)–(7.5) is

$$\begin{aligned}(7.6)\qquad &\alpha V_{\theta_1,\theta_2}(x)\\ &= \min_{u\in U}\Big\{x(t)^T Q x(t) + u(t)^T R u(t) + (A(\theta_1)x(t) + B(\theta_2)u(t))\partial_x V_{\theta_1,\theta_2}(x)\\ &\quad + \tfrac{1}{2}\operatorname{Tr}[\sigma\sigma^T \partial^2_{xx} V_{\theta_1,\theta_2}(x)]\Big\}.\end{aligned}$$

Here, we take $U = \mathbb{R}^2$ and $\Theta := (-\infty, 0] \times [0, \infty)$. To solve (7.6), we propose a solution $V_{\theta_1,\theta_2} \in C^2(\mathbb{R}^2) \cap \mathcal{B}_w(\mathbb{R}^2)$ of the form

$$V_{\theta_1,\theta_2}(x) = x^T K(\theta_1,\theta_2) x + g \tag{7.7}$$

with $K(\theta_1,\theta_2)$ a positive symmetric 2×2 matrix, and g is a constant, both to be determined. So, substituting the derivatives of V_{θ_1,θ_2} in (7.6), and letting $K \equiv K(\theta_1,\theta_2)$, we obtain the optimal control

$$f^*_{\theta_1,\theta_2}(x(t)) = -R^{-1}B(\theta_2)^T K^T x(t), \tag{7.8}$$

where K satisfies the algebraic Riccati equation

$$Q - KB(\theta_2)(R^{-1})^T B^T K + KA(\theta_1) + A(\theta_1)^T K - \alpha K = 0, \tag{7.9}$$

and $g = \operatorname{Tr}(\sigma\sigma^T K/\alpha)$.

7.1. Numerical results. In this application the discrete approximate likelihood function (5.1) depends on three variables, i.e., $\mathrm{MLR}(X_T, \theta_1, \theta_2)$. By matching to zero the partial derivatives of $\mathrm{MLR}(X_T, \theta_1, \theta_2)$ with respect to θ_1 and θ_2, a straightforward calculation gives the estimator

$$\theta_{\mathrm{LR}_m} := \begin{bmatrix} \theta_{1m} \\ \theta_{2m} \end{bmatrix} := [\mathrm{Aux}]^{-1} \cdot \mathrm{baux} \tag{7.10}$$

where

$$\mathrm{Aux} := \begin{bmatrix} \sum_{i=1}^m x_{2t_i}^2 & V_{\mathrm{in}} \sum_{i=1}^m x_{2t_i} \\ V_{\mathrm{in}} \sum_{i=1}^m x_{2t_i} & V_{\mathrm{in}}^2 \end{bmatrix},$$

$$\mathrm{baux} := \begin{bmatrix} \sum_{i=1}^m x_{2t_i}[x_{2t_i+1} - x_{2t_i-1}] \\ V_{\mathrm{in}} \sum_{i=1}^m [x_{2t_i+1} - x_{2t_i-1}] \end{bmatrix} \frac{1}{t_{i+1} - t_{i-1}},$$

and

$$x(t) := \left[w(t), \frac{dw}{dt}\right]^T = [x_1, x_2]^T.$$

Now, taking into account that the true values of the parameters are $\theta_1 := \frac{-k_T K_b}{RJ} = -45.4450$ and $\theta_2 := \frac{K_T}{RJ} = 17755.4$, we compare the optimal control problem with known parameters and the adaptive control problem in which the estimator (7.10) is used as the correct value of the unknown parameters.

Table 4 displays the RMSE for the state variables $w(t)$, $dw(t)/dt$ and their optimal controls ($u_w \equiv 0$, $u_{dw} \equiv f_{\theta_1,\theta_2}$), respectively. Moreover, Table 4 also shows the RMSE between the θ_{LR_m}-optimal cost and the true (θ_1, θ_2)-optimal cost taking as discount factor $\alpha = 0.02$. Note that the best approximation is with $m = 594$ data. We note that the principle of estimation and control (PEC) gives good results for this application because the RMSE values decrease as the number of observations increases. See Table 4 and Figure 4.

Table 4. RMSE between the estimated and the real values of $w(t)$, $dw(t)/dt$, $u(t)$, and $V^*_{\theta_1,\theta_2}$

m	θ_{1m}	θ_{2m}	RMSE_w	RMSE_{u_w}	m	RMSE_{dw}	$\mathrm{RMSE}_{u_{dw}}$	RMSE_{V^*}
594	−45.4450	17755.4	0.0527	0	594	0.0261	0.0417316	0.0198
297	−45.5676	17806.6	0.2675	0	297	12.1074	11.8778	2.8167
198	−45.9121	17951	0.8383	0	198	27.2737	27.0925	7.09924
149	−46.3383	18094.2	1.0446	0	149	40.6889	40.1953	13.2269
119	−46.6966	18254.3	1.8782	0	119	56.9997	56.5847	17.6605

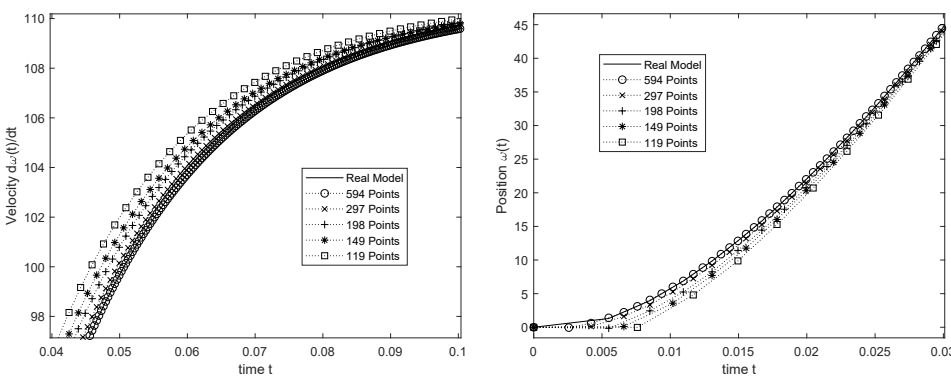

Fig. 4. Asymptotic behavior of estimates of w (right) and dw/dt (left) in (7.4)

8. Series RLC circuit, [28]. The charge in a series RLC circuit evolves according to the second-order differential equation

$$\frac{d^2q(t)}{dt^2} = -\frac{1}{LC}q(t) - \frac{R}{L}\frac{dq(t)}{dt} + \frac{V}{L}, \tag{8.1}$$

with initial conditions $q(0) = 0$ and $q'(0) = 0$. Here, R is the resistance, L the inductance, C the capacitance, and V the source voltage. Considering a stochastic effect in the source voltage, it is possible to get the following matrix stochastic differential equation for the charge in a series RLC circuit:

$$dQ(t) = (AQ(t) + Bu(t))dt + adW(t), \tag{8.2}$$

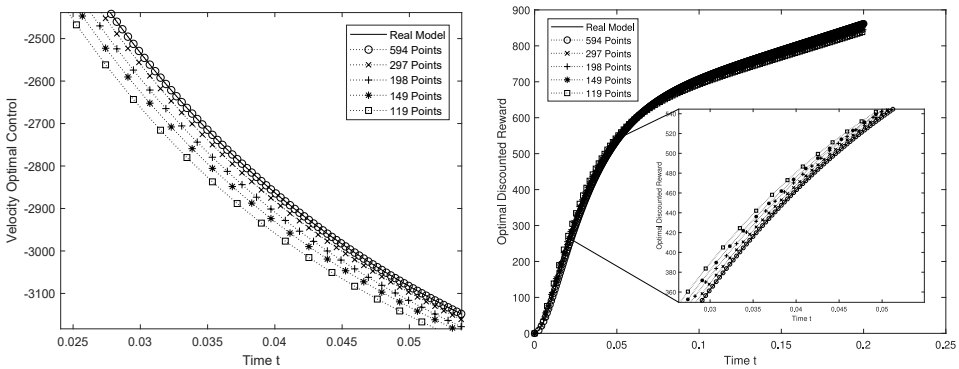

Fig. 5. Asymptotic behaviour of the estimated optimal control $f_{\theta_{1m},\theta_{2m}}(x^{\theta_{1m},\theta_{2m}})$ for dw/dt (left) and the θ_m-optimal discount costs in (7.7) (right)

with

$$Q(t) := \begin{pmatrix} q(t) \\ dq(t)/dt \end{pmatrix}, \; A := \begin{bmatrix} 0 & 1 \\ -1/LC & -R/L \end{bmatrix}, \; B := \begin{bmatrix} 0 \\ 1/L \end{bmatrix}, \; a = \begin{pmatrix} 0 \\ \sigma/L \end{pmatrix}$$

and $u(t) = V(t)$. See [28, 23] for details.

Experimental procedure. A DC voltage source connected to a circuit was used to carry out the experiment of a series RLC electric circuit, with $V = 5\,\mathrm{V}$, $R = 160\,\mathrm{k}\Omega$, $L = 2.85\,\mathrm{mH}$, and $C = 100\mathrm{e}{-6}\,\mu\mathrm{F}$. A National Instrument (NI) data acquisition card USB 6003 was used with a resolution of 16 bits connected to Labview software to measure the voltage across the capacitor and store data. The experiment consists of connecting the DC source to the RLC circuit and measuring the voltage in the capacitor, the source is disconnected and the capacitor is discharged; this process was repeated 50 times. Using the relation $V(t) = Cq(t)$ we can obtain the charge data. Figure 6 shows the graph of the analytical solutions of the ODE ($\sigma = 0$), SDE with $\sigma = 0.0277$, and experimental data. The RMSE between the analytical solutions of ODE, SDE, and the experimental data is displayed in Table 5. The RMSE values present higher variation with respect to the applications

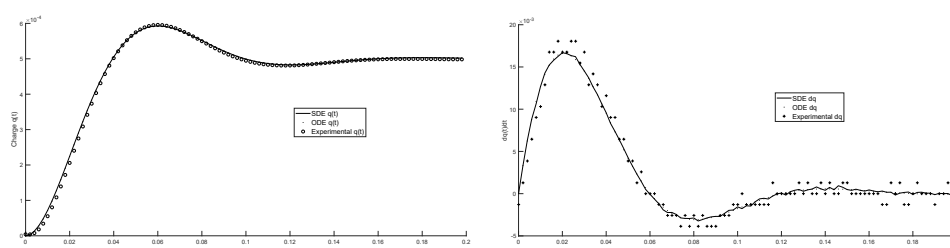

Fig. 6. Experimental data together with the solution of (8.2), $\frac{dq}{dt}$ (right) and $q(t)$ (left)

Table 5. RMSE between the experimental data and analytical solutions of (8.1) and (8.2)

Analytical solutions vs experimental measurements	RMSE $q(t)$	RMSE $dq(t)/dt$
ODE solution vs Experimental	6.2922	9.1703e−04
SDE solution vs Experimental	6.4040	9.3316e−04
ODE vs SDE	2.4456e−07	1.4084e−04

presented in previous sections. This is because the series RLC circuit, in addition to white noise, has thermal noise, which is not considered in this work. Even so, the proposed SDE (8.2) serves our purposes.

In this application we study (8.1)–(8.2) assuming that the pair of matrices (A, B) in the stochastic differential equation (8.2) are unknown. That is, $(A, B) = (A(\theta_1, \theta_2), B(\theta_3))$, where $\theta_1 := -\frac{1}{LC}$, $\theta_2 := -\frac{R}{L}$, and $\theta_3 := \frac{1}{L}$ are unknown parameters. Moreover, we assume that $U = \mathbb{R}^2$ and $\Theta = (-\infty, 0] \times (-\infty, 0] \times [0, \infty)$.

8.1. Numerical results. The discrete approximate likelihood estimator θ_{LR} in (5.2) is used to estimate the unknown parameters. Observe that in this application the discrete approximate likelihood function depends on four variables, i.e., $\mathrm{MLR}(X_T, \theta_1, \theta_2, \theta_3)$. Table 6 displays the RMSE for the state variables $q(t)$, $dq(t)/dt$ and their optimal controls ($u_q \equiv 0$, $u_{dq} \equiv f_{\theta_1,\theta_2,\theta_3}$), respectively, whereas Table 7 shows the RMSE between the θ_{LR_m}-optimal cost and the true $(\theta_1, \theta_2, \theta_3)$-optimal cost for discount factor $\alpha = 0.02$. Note that in both tables the best approximation is with $m = 1998$. We note that the principle of estimation and control gives good results for this application. See Table 6 and Figure 7.

Table 6. RMSE between the estimated and the real values of $q(t), dq(t)/dt$, and $u(t)$

m	θ_1	θ_2	θ_3	RMSE_q
True value	−3508.77	−56.1404	0.3508	0
1998	−3378.66	−50.0797	0.3346	7.9430e−06
1332	−3118.99	−45.8068	0.3075	1.3207e−05
998	−3247.59	−48.0138	0.3210	1.0850e−05
798	−3128.17	−46.0119	0.3085	1.4178e−05

m	RMSE_{u_q}	RMSE_{dq}	$\mathrm{RMSE}_{u_{dq}}$
1998	0	0.00042	1.17307e−05
1332	0	0.00078	1.62389e−05
998	0	0.00064	1.43059e−05
798	0	0.00077	1.61938e−05

Table 7. RMSE between $V^*_{\theta_{\mathrm{LR}_m}}(x^{\theta_{\mathrm{LR}_m}}(t))$ and $V^*_\theta(x^\theta(t))$

m	θ_1	θ_2	θ_3	RMSE_{V^*}
1998	−3378.66	−50.0797	0.334635	0.00269323
1332	−3118.99	−45.8068	0.307551	0.0240681
998	−3247.59	−48.0138	0.321076	0.0138565
798	−3128.17	−46.0119	0.308563	0.0234434

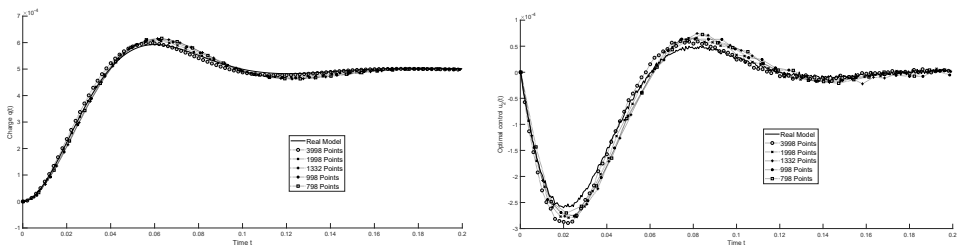

Fig. 7. Asymptotic behavior of the estimated charges (left) and the corresponding optimal control (right)

9. Appendix: Interchange of limits. Throughout this section we assume that $\mathcal{O} \subset \mathbb{R}^n$ is an open, bounded, connected set. We denote its closure by $\bar{\mathcal{O}}$. For every $x \in \mathbb{R}^n$, $u \in U$, $\alpha > 0$, and h in $W^{2,p}(\bar{\mathcal{O}})$, let

$$\hat{\Psi}(x,u,\theta;h) := r(x,u,\theta) + \sum_{i=1}^{n} b_i(x,u,\theta)\partial_i h(x) - \alpha h(x), \tag{9.1}$$

where r is the reward rate given in Assumption 2.3 and b_i is the ith component of the drift function b in (2.1). We also define

$$L^{\theta,u}h(x) := \hat{\Psi}(x,u,\theta;h) + \frac{1}{2}\sum_{i,j=1}^{n} a^{ij}(x)\partial^2_{ij}h(x), \tag{9.2}$$

with a as in Assumption 2.1(c). For each $\pi \in \mathit{\Pi}$, let

$$\hat{\Psi}(x,\pi,\theta;h) := \int_U \hat{\Psi}(x,u,\theta;h)\,\pi(du|x),$$

$$L^{\theta,\pi}h(x) := \hat{\Psi}(x,\pi,\theta;h) + \frac{1}{2}\sum_{i,j=1}^{n} a^{ij}(x)\partial^2_{ij}h(x). \tag{9.3}$$

Our approach in this paper requires the following result on the interchange of limits, which is an extension to the adaptive case of [24, Theorem 6.1]. We omit the proof because, except for obvious notational changes, it is the same as [24, Theorem 6.1].

THEOREM 9.1. *Let $\mathcal{O}$ be a bounded $\mathcal{C}^2$ domain, and suppose that Assumptions* 2.1–2.3 *hold. In addition, assume that there exist sequences $\{h_m\} \subset W^{2,p}(\mathcal{O})$, $\{\xi_m\} \subset L^p(\mathcal{O})$, with $p > n$ (n is the dimension of* (2.1)*), $\{\pi_m\} \subset \Pi$, and $\theta_m \in \Theta$, satisfying the following:*

(a) $L^{\theta_m,\pi_m} h_m = \xi_m$ *in* $\mathcal{O}$ *for* $m = 1, 2, \ldots$.
(b) *There exists a constant* $\tilde{M}_1$ *such that* $\|h_m\|_{W^{2,p}(\mathcal{O})} \le \tilde{M}_1$ *for* $m = 1, 2, \ldots$.
(c) ξ_m *converges in* $L^p(\mathcal{O})$ *to some function* ξ.
(d) θ_m *converges to some* θ $\mathbb{P}^{\theta,\pi}$*-a.s.*
(e) $\pi_m \xrightarrow{w} \pi$ *in the topology of relaxed controls* (*Definition* 4.1).

Then there exist a function $h \in W^{2,p}(\mathcal{O})$ and a subsequence $\{m_k\} \subset \{1, 2, \ldots\}$ such that $h_{m_k} \to h$ in the norm of $\mathcal{C}^{1,\eta}(\mathcal{O})$ for $\eta < 1 - n/p$ as $k \to \infty$. Moreover,

$$L^{\theta,\pi} h = \xi \quad \text{in } \mathcal{O} \ \mathbb{P}^{\theta,\pi}\text{-a.s.}$$

10. Concluding remarks. This paper concerns controlled stochastic differential equations (2.1) in which the drift coefficient depends on an unknown parameter $\theta \in \Theta$. The optimal control problem (OCP) is to maximize, for every initial state $x \in \mathbb{R}^n$, the expected total discounted reward $V(x, \pi, \theta)$ over all control policies $\pi \in \Pi$, given that the true parameter value is θ. Then, defining $V_\theta^*(x) := \sup_\pi V(x, \pi, \theta)$ the optimal discounted reward, our main results can be summarized as follows: If θ_m is a sequence of uniformly strongly consistent (USC) estimators of θ, then, under suitable conditions:

(1) For each m there is an optimal control policy π_{θ_m} for the θ_m-OCP.
(2) For each initial state x, the optimal reward $V_{\theta_m}^*(x)$ converges to $V_\theta^*(x)$ almost surely as $m \to \infty$.
(3) There is a subsequence $\{m_k\}$ of $\{m\}$ and a policy $\pi_\theta^* \in \Pi$ such that $\pi_{\theta_{m_k}}$ converges to π_θ^* in two different ways (see (4) below), and moreover π_θ^* is optimal for the θ-OCP.
(4) The convergence $\pi_{\theta_{m_k}}^* \to \pi_\theta^*$ in (3) is both in the topology of relaxed controls (Definition 4.1), and in the sense of Schäl (Definition 4.2). The main difference between these forms of convergence is that the former has "better" properties (see Remark 4.3 for instance).

In view of these remarks, it is evident that a crucial step in our approach is to obtain a sequence of USC estimators of the unknown parameter θ. There are, in principle, many ways to do this; see [1, 2, 6, 14, 22, 31, 32, 33, 34, 35, 36, 37, 23, 42]. Our experience is, however, that in some applications, to use the MLR and the LSQ functions one needs to check the type of the required numerical approximation of the derivative $dx(t)$. Indeed, in our case, we

replaced $dx(t)$ by its central difference, instead of the backward difference, because for our applications it yields more accurate approximations.

Finally, it is important to note that our assumptions in Section 2 are *sufficient* for the results in Sections 3 and 4, but in many cases they are not necessary. For instance, for LQ problems (linear systems with quadratic costs), the compactness of U and Θ is not necessary. Similarly, some of our results using the PEC hold for *deterministic* systems (with $\sigma(\cdot) \equiv 0$ in (2.1)), even though the uniform ellipticity condition in Assumption 2.1(c) does not hold.

Acknowledgements. We wish to give a special mention to the Control Laboratory of the Faculty of Engineering, University of Veracruz for allowing us to use their equipment and facilities for the completion of the experiments. The research of O. Hernández-Lerma was partially supported by the Fondo SEP-CINVESTAV grant FIDSC 2018/196 and by CONACYT grant 263963.

References

[1] Y. Aït-Sahalia, *Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach*, Econometrica 70 (2002), 223–262.

[2] Y. Aït-Sahalia, *Closed-form likelihood expansions for multivariate diffusions*, Ann. Statist. 36 (2008), 906–937.

[3] A. Arapostathis, V. S. Borkar and M. K. Ghosh, *Ergodic Control of Diffusion Processes*, Encyclopedia Math. Appl. 143, Cambridge Univ. Press, 2012.

[4] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, Dover, New York, 2013.

[5] T. R. Bielecki, *Adaptive control of continuous-time linear stochastic systems with discounted cost criterion*, J. Optim. Theory Appl. 68 (1991), 379–383.

[6] V. S. Borkar and A. Bagchi, *Parameter estimation in continuous-time stochastic processes*, Stochastics 8 (1982), 193–212.

[7] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions I: The existence results*, SIAM J. Control Optim. 26 (1986), 112–126.

[8] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions II: Adaptive control*, Appl. Math. Optim. 21 (1990), 191–220.

[9] H. E. Chen, T. E. Duncan and B. Pasik-Duncan, *Stochastic adaptive control for continuous-time linear systems with quadratic cost*, Appl. Math. Optim. 34 (1996), 113–138.

[10] G. B. Di Masi and Ł. Stettner, *Bayesian ergodic adaptive control of diffusion processes*, Stoch. Stoch. Reports 60 (1997), 155–183.

[11] T. E. Duncan, L. Guo and B. Pasik-Duncan, *Adaptive continuous-time linear quadratic Gaussian control*, IEEE Trans. Automatic Control 44 (1999), 1653–1662.

[12] T. E. Duncan and B. Pasik-Duncan, *Adaptive control of continuous time linear stochastic systems*, Math. Control Signals Systems 3 (1990), 45–60.

[13] T. E. Duncan, B. Pasik-Duncan and Ł. Stettner, *Almost self-optimizing strategies for the adaptive control of diffusion processes*, J. Optim. Theory Appl. 81 (1994), 479–507.

[14] G. B. Durham and A. R. Gallant, *Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes*, J. Business Economic Statistics 20 (2002), 297–316.

[15] W. H. Fleming and M. Nisio, *On the stochastic relaxed control for partially observed diffusions*, Nagoya Math. J. 93 (1984), 71–108.

[16] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd ed., Springer, New York, 2006.

[17] M. K. Ghosh, A. Arapostathis and S. I. Marcus, *Optimal control of switching diffusions with applications to flexible manufacturing systems*, SIAM J. Control Optim. 30 (1992), 1–23.

[18] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Heildelberg, 1998.

[19] O. Hernández-Lerma and T. E. Govindan. *Nonstationary continuous-time Markov control process with discounted costs on infinite horizon*, Acta Appl. Math. 67 (2001), 277–293.

[20] O. Hernández-Lerma and S. I. Marcus, *Adaptive control of discounted Markov Decision chains*, J. Optim. Theory Appl. 46 (1985), 227–235.

[21] N. Hilgert and A. Minjárez-Sosa, *Adaptive control of stochastic systems with unknown disturbance distribution: discounted criteria*, Math. Methods Oper. Res. 63 (2006), 443–460.

[22] M. Huzak, *Estimating a class of diffusions from discrete observations via approximate maximum likelihood method*, Statistics 52 (2018), 239–272.

[23] E. B. Jamkhaneh, *The effect of different noise perturbations on the parameter of the RC circuit using stochastic differential equation (SDE)*, World Appl. Sci. J. 13 (2011), 2198–2202.

[24] H. Jasso-Fuentes, B. A. Escobedo-Trujillo and A. Mendoza-Pérez, *The Lagrange and the vanishing discount techniques to controlled diffusions with cost constraints*, J. Math. Anal. Appl. 437 (2016), 999–1035.

[25] H. Jasso-Fuentes and O. Hernández-Lerma, *Characterizations of overtaking optimality for controlled diffusion processes*, Appl. Math. Optim. 57 (2007), 349–369.

[26] H. Jasso-Fuentes and G. Yin, *Advanced Criteria for Controlled Markov-Modulated Diffusions in an Infinite Horizon: Overtaking, Bias, and Blackwell Optimality*, Science Press, Beijing, 2013.

[27] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[28] T. Kumar and H. Parthasarathy, *Modeling of an RC circuit using a stochastic differential equation*, Thammasat Int. J. Sci. Tech. 13 (2008), 40–47.

[29] M. Kurano, *Discrete-time Markovian decision processes with an unknown parameter. Average return criterion*, J. Oper. Res. Soc. Japan 15 (1972), 67–76.

[30] P. Mandl, *Estimation and control in Markov chains*, Adv. Appl. Probab. 6 (1974), 40–60.

[31] B. B. Martin and M. Sorensen, *Martingale estimation functions for discretely observed diffusion processes*, Bernoulli 1 (1995), 017–039.

[32] N. J. Nygaard, H. Madsen and C. P. Young, *Parameter estimation in stochastic differential equations: An overview*, Ann. Rev. Control 24 (2000), 83–94.

[33] C. R. Penha, *On the statistical estimation of diffusion processes: A partial survey*, Brazil. Rev. Econometrics 24 (2004), 273–301.

[34] A. R. Perdersen, *A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations*, Scand. J. Statist. 22 (1995), 55–71.

[35] A. R. Perdersen, *Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusions process*, Bernoulli 1 (1995), 257–279..

[36] B. L. S. Prakasa Rao, *Asymptotic theory for non-linear least squares estimator for diffusion processes*, Math. Operationsforsch. Statist. Ser. Statist. 14 (1983), 195–209.

[37] K. Ralchenko, *Asymptotic normality of discretized maximum likelihood estimator for drift parameter in homogeneous diffusion model*, Modern Stochastics Theory Appl. 2 (2015), 17–28.

[38] U. Rieder, *Measurable selection theorems for optimization problems*, Manuscripta Math. 24 (1978), 115–131.

[39] M. Schäl, *A selection theorem for optimization problems*, Arch. Math. (Basel) 25 (1974), 219–224.

[40] M. Schäl, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrsch. Verw. Gebiete 32 (1975), 179–196.

[41] M. Schäl, *Estimation and control in discounted stochastic dynamic programming*, Stochastics 20 (1987), 51–71.

[42] I. Shoji, *A note on asymptotic properties of the estimator derived from the Euler method for diffusion processes at discrete times*, Statist. Probab. Lett. 36 (1997), 153–159.

[43] D. Vrabie, O. Pastravanu, M. Abu-Khalaf and F. L. Lewis, *Adaptive optimal control for continuous-time linear systems based on policy iteration*, Automatica 45 (2009), 477–484.

[44] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

B. A. Escobedo-Trujillo
Facultad de Ingeniería
Universidad Veracruzana
Av. Universidad km 7.5
C.P. 96535, Coatzacoalcos, Veracruz, México
ORCID: 0000-0002-8937-3019
E-mail: bescobedo@uv.mx

O. Hernández-Lerma
Departamento de Matemáticas
CINVESTAV-IPN
A. Postal 14-740
Ciudad de México, 07000, México
ORCID: 0000-0003-3308-5218
E-mail: ohernand@math.cinvestav.mx

F. A. Alaffita-Hernández
Centro de Investigación en Recursos Energéticos y Sustentables
Universidad Veracruzana
Av. Universidad km 7.5
C.P. 96535, Coatzacoalcos, Veracruz, México
ORCID: 0000-0002-7971-6356
E-mail: falaffita@uv.mx