Systems & Control Letters 118 (2018) 101–108

Contents lists available at ScienceDirect

Systems & Control Letters

journal homepage: www.elsevier.com/locate/sysconle

Stochastic L^1-optimal control via forward and backward sampling

Ioannis Exarchos *, Evangelos A. Theodorou, Panagiotis Tsiotras

The Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, 30332 GA, USA

ARTICLE INFO

Article history:
Received 14 January 2018
Received in revised form 26 April 2018
Accepted 15 June 2018
Available online 6 July 2018

Keywords:
Stochastic L^1 optimal control
Forward and backward stochastic differential equations

ABSTRACT

The aim of this work is to present a sampling-based algorithm designed to solve a certain class of stochastic optimal control problems, utilizing forward and backward stochastic differential equations (FBSDEs). Specifically, we address the class of problems in which the running cost of the performance index involves an L^1-type minimization problem in terms of the control effort. Such problems are typically called *minimum fuel* problems in optimal control literature. By means of a nonlinear version of the Feynman–Kac lemma, we obtain a probabilistic representation of the solution to the nonlinear Hamilton–Jacobi–Bellman equation, expressed in the form of a system of decoupled FBSDEs. This system of FBSDEs can be solved by employing linear regression techniques.

1. Introduction

Solving an optimal control problem in a stochastic setting has been typically associated with the solution of a second-order, nonlinear partial differential equation (PDE), called the Hamilton–Jacobi–Bellman (HJB) equation. By and large, the literature on optimal control deals with the minimization of a performance index which penalizes control *energy*, since the input appears in quadratic form as part of the running cost. Such problems are typically referred to as *minimum energy* or L^2 problems in optimal control theory. While such L^2 formulation can be useful in many engineering problems (e.g., preventing engine overheating, avoiding high frequency control input signals etc.), there are practical applications in which the control input is bounded (e.g. due to actuation constraints), and the L^1 norm is a more suitable choice to penalize. These problems are also called *minimum fuel* problems, due to the nature of the running cost, which involves an integral of the absolute value of the input signal. Minimum fuel control appears as a necessity in several settings, especially in spacecraft guidance and control [1,2], in which fuel is a limited resource. Indeed, in such applications, using the L^2-norm results in significantly more propellant consumption as well as undesirable continuous thrusting. In some illustrative examples, this fuel penalty can be as high as 50% [3].

The notion of L^1-optimal control is also tightly related to *Maximum Hands-Off control* [4,5]. The distinguishing characteristic of a hands-off controller is that it tries to retain a zero control input value over an extended time interval. Thus, the objective of "maximum hands-off" control is to accomplish a specific task while applying zero input for the longest time duration possible. Applications of this type of control range from the automotive industry (engine stop–start systems [6], hybrid vehicles [7]) to networked and embedded systems [8,9]. The "hands-off" property, especially in a discrete context, is equivalent to the *sparsity* of a signal, i.e., minimizing the total length of intervals over which the signal takes non-zero values. The relationship between L^1-optimality and the "hands-off" property, or sparsity, is shown in [4,5]. Specifically, if an L^1-optimal control problem is *normal* (see Remark 1 of Section 2, as well as [10]), then its optimal solution is also the optimal sparse, "hands-off" solution.

Despite the aforementioned advantages, investigation of L^1-optimal control in the literature is not as widespread as L^2 problems, since it leads to significantly more complicated optimal control structures. These structures are usually a combination of *bang–bang* control (i.e, the control signal switches between its extrema) and *singular* control, in which the control input receives intermediate values. Moreover, the particular structure often depends on the specific initial condition or other parameter values, and neither existence, nor uniqueness of solutions, can always be guaranteed [10].

In this paper, we present a sampling-based algorithm designed to solve L^1 stochastic optimal control problems, utilizing forward and backward stochastic differential equations (FBSDEs). By means of a nonlinear version of the Feynman–Kac lemma, we obtain a probabilistic representation of the solution to the nonlinear Hamilton–Jacobi–Bellman equation, expressed in the form of a system of decoupled FBSDEs. This system of FBSDEs can be solved by employing linear regression techniques. The scheme is enhanced with importance sampling and trajectory blending, resulting in an iterative algorithm that learns the optimal control without requiring an initial guess. We validate the algorithm by applying

* Corresponding author.
E-mail address: exarchos@gatech.edu (I. Exarchos).

https://doi.org/10.1016/j.sysconle.2018.06.005
0167-6911/

it on a well-known minimum fuel problem, and demonstrate its superiority against deterministic control laws in the presence of stochastic disturbances.

The contributions of this paper are as follows:

- Theoretic: It is shown that L^1-optimal control problems of the form considered within this paper can be cast as a specific FBSDE problem, by virtue of the nonlinear Feynman–Kac lemma, and through a particular decomposability condition presented herein. This FBSDE problem can then be solved in lieu of the original PDE problem. Our previous work on L^2-optimal control [11] does not cover topic, as it requires separate derivation.
- Algorithmic: We present an algorithm utilizing a simplified discretization scheme and importance sampling. In order to address the numerical difficulties arising in L^1-optimal problems in particular, the algorithm in this paper is further enhanced with a trajectory blending technique to improve its convergence properties. The algorithm is validated on a problem for which a known solution is available, and its performance is also tested on a nonlinear system. To the best of our knowledge, this paper is the first to present an algorithm addressing stochastic L^1-optimal control problems.

2. Problem statement

Let $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P})$ be a complete, filtered probability space on which a p-dimensional standard Brownian motion W_t is defined, such that $\{\mathscr{F}_t\}_{t\geq 0}$ is the natural filtration of W_t augmented by all $\mathbb{P}$-null sets. Our goal is to minimize the expected cost defined by

$$J(\tau, x_\tau; u(\cdot)) = \mathbb{E}\Bigg[g(x(T)) + \int_\tau^T q_0(t, x(t)) + q_1(t, x(t))^\top |u(t)|\, \mathrm{d}t\Bigg], \tag{1}$$

wherein T is a fixed time of termination, $x \in \mathbb{R}^n$ is the state vector and $u \in U \subset \mathbb{R}^\nu$ is the control vector. The dynamics are assumed to obey an Itô drift–diffusion process, given by the stochastic differential equation (SDE)

$$\mathrm{d}x(t) = f(t, x(t))\,\mathrm{d}t + G(t, x(t))u(t)\,\mathrm{d}t + \Sigma(t, x(t))\,\mathrm{d}W_t,$$
$$t \in [\tau, T], \qquad x(\tau) = x_\tau. \tag{2}$$

Here, the control is restricted to $U = [-u_1^{\min}, u_1^{\max}] \times [-u_2^{\min}, u_2^{\max}] \times \cdots \times [-u_\nu^{\min}, u_\nu^{\max}]$, with $u_i^{\min} \geq 0$, $u_i^{\max} > 0$. Note that the assumption about the signs of $u_i^{\min}$ and $u_i^{\max}$ is without loss of generality. Furthermore, $|\cdot|$ denotes the element-wise absolute value, $q_1 : [0, T] \times \mathbb{R}^n \to \mathbb{R}^\nu_+$ is a (possibly time/state dependent) vector of nonnegative weights, and $q_0 : [0, T] \times \mathbb{R}^n \to \mathbb{R}$ is the state-depended part of the running cost. If the "fuel consumption" penalty is to be applied on all control channels equally, independently of time or state, then q_1 reduces to a constant vector of ones. Finally, all aforementioned functions, as well as $g : \mathbb{R}^n \to \mathbb{R}$, $f : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$, $G : [0, T] \times \mathbb{R}^n \to \mathbb{R}^{n\times\nu}$, and $\Sigma : [0, T] \times \mathbb{R}^n \to \mathbb{R}^{n\times p}$, are deterministic functions, continuous w.r.t. time t (in case there is explicit dependence), uniformly bounded w.r.t time t, and Lipschitz (uniformly in t) with respect to the state variables. These standard assumptions [12] guarantee that the SDE solution is unique and does not have a finite escape time, similar to the case of ordinary differential equations, in addition to a well-defined cost functional (1). Furthermore, the square-integrable[1] process $u : [0, T] \times \Omega \to U \subseteq \mathbb{R}^\nu$ is $\{\mathscr{F}_t\}_{t\geq 0}$-adapted, which essentially requires the control input to be non-anticipating, i.e., to rely only on past and present information.

Our goal is to minimize (1) for any given initial condition (τ, x_τ), and under all admissible functions $u(\cdot)$. The value function V is defined as

$$V(\tau, x_\tau) = \inf_{u(\cdot)\in\mathscr{U}[0,T]} J(\tau, x_\tau; u(\cdot)),\ \forall (\tau, x_\tau) \in [0, T) \times \mathbb{R}^n,$$
$$V(T, x) = g(x), \qquad \forall x \in \mathbb{R}^n. \tag{3}$$

Through Bellman's principle of optimality, it is shown [12,13] that if the value function is in $C^{1,2}([0, T] \times \mathbb{R}^n)$, then it satisfies the Hamilton–Jacobi–Bellman (HJB) equation, which – omitting function arguments for brevity – assumes for the case at hand the following form

$$v_t + \inf_{u\in U}\Bigg\{\frac{1}{2}\mathrm{tr}(v_{xx}\Sigma\Sigma^\top) + v_x^\top f + \big(v_x^\top G + q_1^\top \mathrm{D}(\mathrm{sgn}(u))\big)u + q_0\Bigg\} = 0, \quad (t, x) \in [0, T) \times \mathbb{R}^n, \quad v(T, x) = g(x), \tag{4}$$

wherein v_x and v_{xx} denote the gradient and the Hessian of v, respectively, $\mathrm{D}(x) \in \mathbb{R}^{n\times n}$ denotes the diagonal matrix with the elements of $x \in \mathbb{R}^n$ in its diagonal, and $\mathrm{sgn}(\cdot)$ denotes the signum function. This result can also be extended to include cases in which the smoothness condition of the value function is not satisfied. Specifically, if one also considers viscosity solutions of (4), then the value function is proven to be a viscosity solution of (4), which is furthermore equal to the classical solution, if such a classical solution exists. For the chosen forms of cost integrand and dynamics at hand, we may carry out the infimum operation over u explicitly. To this end, letting u_i be the ith element of u, it is easy to show that the optimal control law is given by

$$u_i^* = \begin{cases} u_i^{\max}, & (v_x^\top G)_i < -(q_1^\top)_i \\ -u_i^{\min}, & (v_x^\top G)_i > (q_1^\top)_i, \\ 0, & -(q_1^\top)_i < (v_x^\top G)_i < (q_1^\top)_i, \end{cases} \qquad i = 1, \ldots, \nu, \tag{5}$$

namely, the optimal control law turns out to be *bang-off-bang* control.

Remark 1. Notice that in the control law given by (5), we do not assign a value for u^* whenever $(v_x^\top G)_i = -(q_1^\top)_i$ or $(v_x^\top G)_i = (q_1^\top)_i$, because in those two cases the control input is not uniquely defined. In fact, any value in $[0, u_i^{\max}]$ and $[-u_i^{\min}, 0]$, respectively, attains the same infimum value in (4). A problem in which either one of these equalities is satisfied over a nontrivial time interval is a *singular* fuel-optimal problem [10]. In what follows, we shall assume that the minimum fuel problem is *normal*, in the sense that the aforementioned equalities are not satisfied over a nontrivial time interval, $\mathbb{P}$-almost surely.

Substituting the control law given by (5), the HJB equation (4) assumes the equivalent form

$$\begin{cases} v_t + \frac{1}{2}\mathrm{tr}(v_{xx}\Sigma\Sigma^\top) + v_x^\top f + q_0 + \\ \sum_{i=1}^{\nu} \min\Big\{\big(v_x^\top G + q_1^\top\big)_i u_i^{\max},\ 0,\ -\big(v_x^\top G - q_1^\top\big)_i u_i^{\min}\Big\} = 0, \\ (t, x) \in [0, T) \times \mathbb{R}^n, \qquad v(T, x) = g(x), \quad x \in \mathbb{R}^n. \end{cases} \tag{6}$$

3. A solution representation via a Feynman–Kac formula

The cornerstone of the proposed approach is the nonlinear Feyman–Kac lemma, which establishes a close relationship between stochastic differential equations (SDEs) and second-order partial differential equations (PDEs) of parabolic or elliptic type.

[1] A process H_s is called square-integrable if $\mathbb{E}\big[\int_t^T H_s^2 \mathrm{d}s\big] < \infty$ for any $T > t$.

Specifically, this lemma demonstrates that solutions to a certain class of PDEs can be represented by solutions to SDEs or systems of forward and backward stochastic differential equations (FBSDEs). There is a plethora of similar theoretic results in the literature, all of which are referred to as Feynman–Kac-type formulas, since the earliest result of this type was due to Feynman and Kac [14]. In this work, we propose to employ a nonlinear Feynman–Kac-type formula, which links the solution of a nonlinear PDE to a system of FBSDEs. In what follows, we will briefly review the theory of forward and backward processes, and then present the nonlinear Feynman–Kac formula.

3.1. The forward and backward process

The forward process is defined as the square-integrable, $\{\mathscr{F}_s\}_{s\geq 0}$-adapted process $X(\cdot)$, which satisfies the Itô FSDE

$$\begin{cases} \mathrm{d}X_s = b(s, X_s)\mathrm{d}s + \Sigma(s, X_s)\mathrm{d}W_s, & s \in [t, T], \\ X_t = x, \end{cases} \tag{7}$$

wherein $(t, x) \in [0, T] \times \mathbb{R}^n$ is a given initial condition. In the literature, the forward process (7) is referred to as the *state process*. We denote the solution to the FSDE (7) as $X_s^{t,x}$, wherein (t, x) are the initial condition parameters.

The associated backward process is the square-integrable, $\{\mathscr{F}_s\}_{s\geq 0}$-adapted pair $(Y(\cdot), Z(\cdot))$ defined via a BSDE satisfying a terminal condition

$$\begin{cases} \mathrm{d}Y_s = -h(s, X_s^{t,x}, Y_s, Z_s)\mathrm{d}s + Z_s^\top \mathrm{d}W_s & s \in [t, T], \\ Y_T = g(X_T). \end{cases} \tag{8}$$

The function $h(\cdot)$ is referred to as *generator* or *driver*. The initial condition parameters (t, x) implicitly define the solution of the BSDE due to the terminal condition $g(X_T^{t,x})$, and thus we will similarly use the notation $Y_s^{t,x}$ and $Z_s^{t,x}$ for the solution associated with a particular initial condition parameter (t, x). Whenever the forward SDE does not depend on Y_s or Z_s, the resulting FBSDEs are *decoupled*.

The difficulty in dealing with BSDEs is that, in contrast to FSDEs, and due to the presence of a terminal condition, integration must be performed backwards in time, i.e., in a direction opposite to the evolution of the filtration. If we do not impose the solution to be adapted (i.e, non-anticipating, obeying the evolution direction of the filtration), we must require new definitions such as the *backward Itô integral* or, more generally, the so-called *anticipating stochastic calculus* [15]. In this work we will restrict the analysis to adapted –non-anticipating – solutions. As shown in [15], an adapted solution is obtained if the *conditional expectation* of the process is back-propagated, by setting $Y_s \triangleq \mathbb{E}[Y_s|\mathscr{F}_s]$. In a sense, systems of FBSDEs describe two-point boundary value problems involving SDEs, with the extra requirement that their solution is adapted to the forward filtration.

3.2. The nonlinear Feynman–Kac lemma

The following lemma, which links the solution of a class of PDEs to that of FBSDEs, can be proven by an application of Itô's formula (see [12,16,15]):

Lemma 1 (*Nonlinear Feynman–Kac*)**.** *Consider the Cauchy problem*

$$\begin{cases} v_t + \dfrac{1}{2}\mathrm{tr}(v_{xx}\Sigma(t, x)\Sigma^\top(t, x)) + v_x^\top b(t, x) \\ \quad + h(t, x, v, \Sigma^\top(t, x)v_x) = 0, \quad (t, x) \in [0, T) \times \mathbb{R}^n, \\ v(T, x) = g(x), \quad x \in \mathbb{R}^n, \end{cases} \tag{9}$$

wherein the functions Σ, b, h and g satisfy mild regularity conditions (see Remark 2). Then (9) admits a unique (viscosity) solution $v : [0, T] \times \mathbb{R}^n \to \mathbb{R}$, which has the following probabilistic representation:

$$v(t, x) = Y_t^{t,x}, \qquad \forall (t, x) \in [0, T] \times \mathbb{R}^n, \tag{10}$$

wherein $(X(\cdot), Y(\cdot), Z(\cdot))$ is the unique adapted solution of the FBSDE system (7),(8). Furthermore,

$$(Y_s^{t,x}, Z_s^{t,x}) = \left(v(s, X_s^{t,x}),\ \Sigma^\top(s, X_s^{t,x})v_x(s, X_s^{t,x})\right), \tag{11}$$

for $s \in [t, T]$, and if (9) admits a classical solution, then (10) provides that classical solution.

Remark 2. Concerning the regularity conditions of Lemma 1, [12] requires the functions Σ, b, h and g to be continuous, Σ and b to be uniformly Lipschitz in x, and h to be Lipschitz in (y, z), uniformly with respect to (t, x). However, the nonlinear Feynman–Kac lemma has been extended to cases imposing less restrictions; see for example [17,18].

Remark 3. The viscosity solution is to be understood in the sense of $v(t, x) = \lim_{\varepsilon\to 0} v^\varepsilon(t, x)$, uniformly in (t, x) over any compact set, where v^ε is the classical solution of the nondegenerate PDE satisfying the form of (9), with Σ, b, h and g replaced by Σ_ε, b_ε, h_ε and g_ε, the latter being smooth functions that converge to Σ, b, h and g uniformly over compact sets, respectively, and $\Sigma_\varepsilon(t, x)\Sigma_\varepsilon^\top(t, x) \geq \varepsilon I + \Sigma(t, x)\Sigma^\top(t, x)$ for all (t, x).

By comparing Eqs. (6) and (9) we conclude that Lemma 1 can be applied to the HJB equation given by (6) under a certain decomposability condition, stated in the following assumption:

Assumption 1. There exists a matrix-valued function $\Gamma : [0, T] \times \mathbb{R}^n \to \mathbb{R}^{p\times\nu}$ such that $G(t, x) = \Sigma(t, x)\Gamma(t, x)$ for all $(t, x) \in [0, T] \times \mathbb{R}^n$.

This assumption implies that the range of G must be a subset of the range of Σ, thus excluding the case of a channel containing control input but no noise, although the converse is allowed. Under this assumption, the HJB equation given by (6) satisfies the format of (9) with

$$b(t, x) \equiv f(t, x), \tag{12}$$

$$h(t, x, z) \equiv q_0(t, x) + \sum_{i=1}^{\nu} \min\left\{\left(z^\top\Gamma + q_1^\top\right)_i u_i^{\max}, 0, -\left(z^\top\Gamma - q_1^\top\right)_i u_i^{\min}\right\}. \tag{13}$$

We thus obtain the (viscosity) solution of (6) by simulating the system of FBSDEs given by (7) and (8) using the definitions (12) and (13).

4. Obtaining a numerical solution to FBSDE systems

FBSDEs have received a lot of attention in the literature, and their solution has been studied independently, by and large, from their connection to PDEs. Several results appear within the field of mathematical finance, and a few generic numerical schemes have been proposed [19–21]. In this paper we employ a scheme proposed in previous work by the authors [11], which capitalizes on the regularity present whenever systems of FBSDEs are linked to PDEs.

4.1. Time discretization of FBSDEs

On a time grid $\{t = t_0 < \cdots < t_N = T\}$ for the interval $[t, T]$, we denote by $\Delta t_i \triangleq t_{i+1} - t_i$ the $(i + 1)$th interval of the grid (which can be selected to be constant) and $\Delta W_i \triangleq W_{t_{i+1}} - W_{t_i}$

the $(i+1)$th Brownian motion increment; here, ΔW_i is simulated as $\sqrt{\Delta t_i}\xi_i$, where $\xi_i \sim \mathcal{N}(0, I)$. We also denote $X_i \triangleq X_{t_i}$ for notational brevity. The simplest scheme for the FSDE is the *Euler–Maruyama* scheme [22]:

$$\begin{cases} X_{i+1} \approx X_i + b(t_i, X_i)\Delta t_i + \Sigma(t_i, X_i)\Delta W_i, \\ \quad i = 0, \dots, N-1, \\ X_0 = x. \end{cases} \tag{14}$$

Note that several higher order schemes do exist that can be used alternatively. The interested reader is referred to [22] for a detailed presentation.

The backward process is discretized by introducing the notation $Y_i \triangleq Y_{t_i}$ and $Z_i \triangleq Z_{t_i}$. Recalling that adapted BSDE solutions impose a back-propagation of conditional expectations, i.e., $Y_s \triangleq \mathbb{E}[Y_s|\mathscr{F}_s]$ and $Z_s \triangleq \mathbb{E}[Z_s|\mathscr{F}_s]$, Eq. (8) is approximated by

$$Y_i = \mathbb{E}[Y_i|\mathscr{F}_{t_i}] \approx \mathbb{E}[Y_{i+1} + h(t_{i+1}, X_{i+1}, Y_{i+1}, Z_{i+1})\Delta t_i|X_i], \tag{15}$$

for $i = N-1, \dots, 0$. We note that the term $Z_i^\top \Delta W_i$ in (8) vanishes in the last equality because of the conditional expectation (ΔW_i is zero mean). Furthermore, $\mathscr{F}_{t_i}$ is replaced by X_i in (15) by virtue of Lemma 1. In light of Eq. (11), the Z-process in (8) corresponds to the term $\Sigma^\top(s, X_s^{t,x})v_x(s, X_s^{t,x})$. Thus, we may write

$$\begin{aligned} Z_i &= \mathbb{E}[Z_i|\mathscr{F}_{t_i}] = \mathbb{E}[\Sigma^\top(t_i, X_i)\nabla_x v(t_i, X_i)|X_i] \\ &= \Sigma^\top(t_i, X_i)\nabla_x v(t_i, X_i). \end{aligned} \tag{16}$$

The above expression requires knowledge of the solution on a neighborhood x at time t_i, that is, $V(t_i, x)$. We initialize the back-propagation at

$$Y_T = g(X_T), \qquad Z_T = \Sigma(T, X_T)^\top \nabla_x g(X_T), \tag{17}$$

assuming that $g(\cdot)$ is a.e. differentiable. Several ways exist to numerically approximate the conditional expectation in (15). Here, we employ the Least Squares Monte Carlo (LSMC) method, which is briefly reviewed in what follows.

4.2. Conditional expectation approximation and its application

The Least Squares Monte Carlo (LSMC) method for approximating conditional expectations was initially introduced within the field of financial mathematics by Longstaff and Schwartz in 2001 [23]. Given two square-integrable random variables X and Y for which we can sample M independent copies of (X, Y) pairs, the method enables estimating conditional expectations of the form $\mathbb{E}[Y|X]$, based on the principle that the conditional expectation of a random variable is a function of the variable on which it is conditioned on: $\mathbb{E}[Y|X] = \phi^*(X)$, where ϕ^* is the solution to the infinite-dimensional minimization problem

$$\phi^* = \arg\min_{\phi} \mathbb{E}[|\phi(X) - Y|^2], \tag{18}$$

and ϕ ranges over all measurable functions with $\mathbb{E}[|\phi(X)|^2] < \infty$. In practice, this problem is substituted by a finite-dimensional approximation by decomposing $\phi(\cdot) \approx \sum_{i=1}^k \varphi_i(\cdot)\alpha_i = \varphi(\cdot)\alpha$, with $\varphi(\cdot)$ being a row vector of k predetermined basis functions and α a column vector of constants, and replacing the expectation operator with its empirical estimator [24]. Thus, one obtains the following least squares problem:

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^k} \frac{1}{M} \sum_{j=1}^{M} |\varphi(X^j)\alpha - Y^j|^2, \tag{19}$$

wherein (X^j, Y^j), $j = 1, \dots, M$ are independent copies of (X, Y). The final expression for the LSMC estimator is simply $\mathbb{E}[Y|X = x] = \phi^*(x) \approx \varphi(x)\alpha^*$. In order to numerically solve FBSDEs, the LSMC method is used to approximate the conditional expectation in Eq. (15) for each time step. We begin by sampling M independent trajectories of the FSDE, $\{X_i^m\}_{i=1,\dots,N}$, $m = 1, \dots, M$. The BSDE numerical scheme is initialized at the terminal time T and is iterated backwards along the time grid. At any given time step t_i, M pairs of data (Y_i^m, X_i^m)[2] are available, on which linear regression is performed to estimate the conditional expectation of Y_i as a function of x at the time step t_i. Thus we obtain an approximation of the Value function V at time t_i, which is valid for the neighborhood of the state space that has been explored by the sample trajectories at that time instant, since $V(t_i, x) = \mathbb{E}[Y_i|X_i = x] \approx \varphi(x)\alpha_i$. The Y_i^m sample values calculated before the regression are then replaced by their projection: $Y_i^m = \varphi(X_i^m)\alpha_i$. Finally, the associated values for Z_i are obtained by taking the gradient with respect to x on the choice of basis functions (assuming that they are differentiable almost everywhere), as follows:

$$Z_i^m \approx \Sigma(t_i, X_i^m)^\top \nabla_x \varphi(X_i^m)\alpha_i. \tag{20}$$

The aforementioned process is repeated for $t_{i-1}, \dots, t_0$. The proposed algorithm can be summarized as

$$\begin{cases} \text{Initialize}: Y_T = g(X_T), \quad Z_T = \Sigma(T, X_T)^\top \nabla_x g(X_T), \\ \alpha_i = \arg\min_{\alpha} \dfrac{1}{M} \\ \quad \times \left\| \Phi(X_i)\alpha - \left(Y_{i+1} + \Delta t_i h(t_{i+1}, X_{i+1}, Y_{i+1}, Z_{i+1}) \right) \right\|^2, \\ Y_i = \Phi(X_i)\alpha_i, \quad Z_i^m = \Sigma(t_i, X_i^m)^\top \nabla_x \varphi(X_i^m)\alpha_i, \end{cases} \tag{21}$$

where $m = 1, \dots, M$ and the matrix Φ contains all basis function evaluations for all data points X_i. The minimizer in (21) can be obtained by directly solving the normal equation, or by performing gradient descent. The algorithm output is essentially the collection of α_i's, i.e., the basis function coefficients for each time instant, allowing for an approximation of the Value function which is valid for the area of the state space explored by the FSDE. We note that the number of basis functions needed for an accurate solution largely depends on the application, the spread of the sample trajectories in the state space, and the type of basis functions used; the PDE boundary condition $v(x, T) = g(x)$ is perhaps the most helpful clue as to what basis functions should be chosen.

5. Iterative schemes: Importance sampling and trajectory blending

The method presented thus far suffers from a significant practical limitation. Specifically, an approximation to the HJB PDE solution is obtained, which is accurate for those areas of the state space that are visited by trajectories of uncontrolled dynamics (Eq. (7)), i.e., trajectories in which no control input is applied. Nevertheless, the optimal trajectory may lie on a different area of the state space, an area which uncontrolled trajectories might not access. In this case, simply extrapolating the locally obtained solution to cover the rest of the state space may be inaccurate. This is a practical limitation, because in theory, if one could sample infinitely many trajectories, they would cover the entire state space, thus eliminating this issue. While generating infinitely many sample trajectories is not a practical solution, the issue can be effectively addressed if one is given the ability to alter the drift term of these sampled trajectories. By changing the drift during sampling, we can essentially select the area of the state space for which the obtained solution is accurate.

In previous work by the authors [11], a scheme involving a drift term modification has been constructed through Girsanov's

[2] Here, Y_i^m denotes the quantity $Y_{i+1}^m + \Delta t_i h(t_{i+1}, X_{i+1}^m, Y_{i+1}^m, Z_{i+1}^m)$, which is the Y_i^m sample value before applying the conditional expectation operator.

theorem on the change of measure [14,25]. Indeed, the system of FBSDEs given by Eqs. (7) and (8) is equivalent, in the sense explained below, to one with modified drift

$$\begin{cases} \mathrm{d}\tilde{X}_s = [b(s,\tilde{X}_s) + \Sigma(s,\tilde{X}_s)K_s]\mathrm{d}s + \Sigma(s,\tilde{X}_s)\mathrm{d}W_s, \quad s \in [t,T], \\ \tilde{X}_t = x, \end{cases} \tag{22}$$

along with the compensated BSDE

$$\begin{cases} \mathrm{d}\tilde{Y}_s = [-h(s,\tilde{X}_s,\tilde{Y}_s,\tilde{Z}_s) + \tilde{Z}_s^\top K_s]\mathrm{d}s + \tilde{Z}_s^\top \mathrm{d}W_s, \quad s \in [t,T], \\ \tilde{Y}_T = g(\tilde{X}_T), \end{cases} \tag{23}$$

for any measurable, bounded and adapted process $K_s : [0,T] \to \mathbb{R}^p$. We note that this equivalence is not path-wise, since different paths are realized by both the forward as well as the backward process, under the modified drift dynamics. Nevertheless, the solution at starting time t, i.e., (Y_t, Z_t), remains the same. The formal proof, which involves Girsanov's theorem on the change of measure can be found in [11]. Additionally, one can explain why the modified system of FBSDEs (22) and (23) may substitute the original FBSDE system by examining the associated PDEs. Indeed, the FBSDE problem defined by (22) and (23) corresponds to the PDE problem

$$\begin{cases} v_t + \frac{1}{2}\mathrm{tr}(v_{xx}\Sigma\Sigma^\top) + v_x^\top(b + \Sigma K) + h(t,x,V,\Sigma^\top V_x) \\ - v_x^\top \Sigma K = 0, \\ (t,x) \in [0,T) \times \mathbb{R}^n, \qquad v(T,x) = g(x), \end{cases} \tag{24}$$

which is identical to the PDE (9), since the term $v_x^\top \Sigma K$ is both added and subtracted. Therefore, the FBSDEs, while being different, are associated with the same PDE problem.

To implement importance sampling for the problem at hand, we may apply any nominal control $\bar{u}$, in the dynamics given by (2), which, in light of the definition of $\Gamma(\cdot)$ in Assumption 1, exhibit the form

$$\begin{aligned} \mathrm{d}x(t) = {} & [f(t,x(t)) + \Sigma(t,x(t))\Gamma(t,x(t))\bar{u}(t)]\mathrm{d}t \\ & + \Sigma(t,x(t))\mathrm{d}W_t. \end{aligned} \tag{25}$$

By comparison with the forward process (22), we thus have

$$K_s = \Gamma(s,X_s)\bar{u}(s), \qquad s \in [t,T], \tag{26}$$

while $b(s,X_s) \equiv f(s,X_s)$ per (12). We note that, while the nominal control $\bar{u}$ may be any control, we are especially interested in the case in which we assign a control calculated at a previous run of the algorithm. Thus, one arrives at an iterative scheme, in which a more accurate approximation of the solution is obtained at each iteration. On the time grid of Section 4, we define $K_i = K_{t_i}$, and use the Euler–Maruyama scheme. Similarly, the most straightforward way to embed importance sampling in the backward process is to simply use the definition

$$\tilde{h}(s,x,y,z,k) \triangleq h(s,x,y,z) - z^\top k, \tag{27}$$

and then utilize the discretized scheme of Section 4 using $\tilde{h}$ instead of h.

Convergence of this procedure is typical in an L^2 setting [11], but harder to achieve in the L^1. The underlying cause seems to be the algorithm's sensitivity to changes in the control law between iterations. However, good performance can be achieved through blending of the sample trajectories used by the algorithm. Specifically, instead of generating all sample trajectories for the next iteration using solely the obtained control law, we may sample only a short percentage of the total number of them (typically 2%–5%). Thus, the new pool of sampled trajectories consists mainly (95%–98%) of the same trajectories as in previous iterations, with only a few new, resulting from the newly obtained control law. Furthermore, we may choose the old trajectories to correspond to lowest cost realizations, thereby discarding the least favorable ones in favor of new realizations generated using a new control law. It can be easily shown that the computational complexity per iteration is $\mathcal{O}(NMn)$ for the forward process and $\mathcal{O}(NMk^2)$ for the backward process.

6. Simulation results

The aim of the simulations presented in this section is twofold. First, the proposed algorithm is validated by means of an application to a linear problem for which an open loop control law is available in closed-form for the deterministic setting of that problem. This is the double integrator problem in Section 6.1. We demonstrate that the algorithm is able to recover the optimal control sequence, using only importance sampling. For this problem, sample trajectory blending is not necessary. Furthermore, the obtained stochastic feedback control law is shown to outperform both the deterministic open loop as well as the deterministic closed-loop control law in the presence of noise. Finally, in Section 6.2, the ability of the algorithm to handle nonlinear dynamics, as well as the significance of the sample trajectory blending technique, are demonstrated through simulations on an inverted pendulum system.

6.1. The double integrator

To validate the proposed algorithm, we tested it on the fuel-optimal control problem of a stochastic double integrator plant. The deterministic case offers a closed form solution; see [10], Ch. 8–6. Specifically, the deterministic problem reads: Given the system equations

$$\dot{x}_1(t) = x_2(t), \quad \dot{x}_2(t) = u(t), \quad |u(t)| \le 1, \tag{28}$$

we wish to find the control which forces the system from an initial state (x_{10}, x_{20}) to the goal state $(0,0)$, and which minimizes the fuel

$$J = \int_0^T |u(t)|\,\mathrm{d}t, \tag{29}$$

where T is a fixed (i.e., prespecified) response time. Existence of solutions is guaranteed if T satisfies a number of conditions depending on the values of the initial state. For an initial state (x_{10}, x_{20}) in the upper right quadrant of the plane, the condition reads $T \ge x_{20} + \sqrt{4x_{10} + 2x_{20}^2}$, and guarantees the existence of a unique solution. The corresponding fuel-optimal control sequence is $\{-1, 0, +1\}$, in which the control switching times t_1 and t_2 are

$$t_1 = 0.5\left(T + x_{20} - \sqrt{(T - x_{20})^2 - 4x_{10} - 2x_{20}^2}\right), \tag{30}$$

$$t_2 = 0.5\left(T + x_{20} + \sqrt{(T - x_{20})^2 - 4x_{10} - 2x_{20}^2}\right), \tag{31}$$

that is,

$$u^*(t) = \begin{cases} -1, & t \in [0,t_1), \\ 0, & t \in [t_1,t_2), \\ 1, & t \in [t_2,T]. \end{cases} \tag{32}$$

We are interested in solving a stochastic version of this problem, in which the system equations are modeled in the following form:

$$\mathrm{d}x_1(t) = x_2(t)\,\mathrm{d}t, \quad \mathrm{d}x_2(t) = u(t)\,\mathrm{d}t + \sigma\,\mathrm{d}w(t), \tag{33}$$

i.e., modeling stochasticity in form of perturbations in the control input u. The introduction of the deterministic problem (28), (29) is done merely to demonstrate that the numerical solution of the stochastic problem obtained by the algorithm exhibits obvious

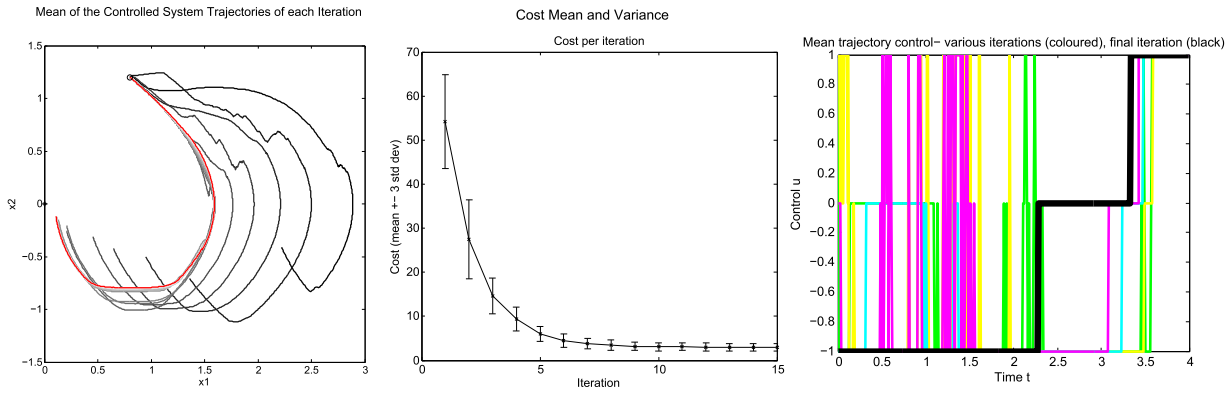

Fig. 1. (a) The mean of the controlled system trajectories of each iteration (grayscale) and after the final iteration (red). (b) Cost mean ± 3 standard deviations per iteration. (c) The control input for the mean system trajectory for each iteration (colored) and after the final iteration (black). We see that the optimal control sequence $\{-1, 0, +1\}$ is finally recovered. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

similarity to the closed form solution of the deterministic problem. An alternative stochastic counterpart could feature noise in the first channel as well. Terminal state conditions are not meaningful in a stochastic setting, since whenever the system dynamics are modeled by controlled diffusions, the probability of hitting a particular point in state space *exactly* is zero. Therefore, instead of the final condition $(x_1(T), x_2(T)) = (0, 0)$, we introduce a "soft" constraint in the cost function by adding a terminal cost:

$$J = \mathbb{E}\left[q(x_1^2(T) + x_2^2(T)) + \int_0^T |u(t)|\,\mathrm{d}t\right], \tag{34}$$

where q is a large enough constant, thus penalizing deviation from the origin at the time of termination. We simulated 2000 trajectories using $\Delta t = 0.01$, with $\sigma = 0.1$, $T = 4$ and $(x_{10}, x_{20}) = (0.8, 1.2)$, and selected $[1, x, x^2]$ for the basis of the value function approximation. The proposed algorithm was run for 15 iterations, using solely importance sampling. The use of sample trajectory blending was not necessary for the convergence of the algorithm. Fig. 1(a) depicts the mean of the controlled trajectories in phase-plane after each iteration of the algorithm (gray scale). The trajectory that corresponds to the final iteration is marked in red. Fig. 1(b) depicts the cost mean ±3 standard deviations per iterations of the algorithm. Lastly, Fig. 1(c) shows the corresponding controls for these mean trajectories in various colors, each color representing an algorithm iteration. The control that corresponds to the final algorithm iteration is marked in black and illustrates that the optimal control sequence $\{-1, 0, +1\}$ is indeed finally recovered.

Some interesting insights are obtained if one compares the performance of the proposed stochastic control law against the deterministic control law (32), if both laws are applied in a system influenced by noise. Specifically, for the same noise profile, we used the three following approaches:

- application of the deterministic control law (32), calculated once (at the initial condition) and applied in an open-loop fashion (D-OL),
- the same control law, applied in a feedback fashion, in which for each time instant and state (t_i, x_i) of the sampled trajectories, the controls are recalculated[3] using the current state as initial condition and $T - t_i$ as a new fixed final time (D-CL),
- the proposed stochastic feedback control law, obtained by our algorithm (S-CL).

[3] Note that the control law in (32) is valid for initial conditions in the upper right quadrant. See [10] for more details.

The results of each approach are depicted in Fig. 2(a), (b), and (c), respectively. As expected, in D-OL, noise results in large variation between trajectories, causing failure in reaching the goal state. Performance is improved in the case of D-CL, as the deterministic controls are recalculated at each iteration, however the improvement is rather minor. This is because in D-CL, even though the control law is applied in a feedback fashion, it does not account for the noise, and thus the resulting trajectories are allowed to drift to areas of the state space for which a new fixed final time $T - t_i$ no longer guarantees existence of a solution that leads to the goal state. The S-CL law obtained by the proposed algorithm does not suffer from this phenomenon. A comparison of the cost mean and variance of these three approaches is shown in Fig. 3. Specifically, D-OL, D-CL and S-CL result in a cost mean of 5.36, 4.75 and 2.97, respectively, and a cost variance of 14.00, 2.49 and 0.07, respectively. Note that here we evaluate the cost given by Eq. (34) for all approaches. In the deterministic setting, and in presence of the fixed final state conditions, the two costs (29) and (34) are equivalent. Finally, by assigning a higher terminal cost than control cost (and vice versa), the proposed framework allows for a systematic way of *shaping* optimal trajectories to achieve the desired robustness-to-fuel-cost trade-off.

6.2. The inverted pendulum

The equations of motion for the inverted pendulum are given, for angle and angular velocity $x_1 = \theta, x_2 = \dot{\theta}$, by

$$\mathrm{d}x_1 = x_2\mathrm{d}t, \qquad \mathrm{d}x_2 = -\left(\frac{bx_2}{m\ell^2} - \frac{g}{\ell}\sin x_1 + u\right)\mathrm{d}t + \sigma\,\mathrm{d}w,$$

i.e., stochasticity enters the system in form of perturbations in the torque u. The constraint $u^{\max} = u^{\min} = mg\ell$ makes this problem nontrivial, since the controller is forced to generate enough momentum by swinging back and forth to successfully invert the pendulum. We simulated 2000 trajectories, using $\Delta t = 0.005$, and $T = 1.5$. The blending ratio was set to $\gamma = 0.98$, and the system noise covariance was set to 0.1. No initial guess for the control input was necessary. We used the same polynomial basis functions as Section 6.1. The task is achieved after approximately 55 iterations, as shown in Fig. 4. These results highlight the importance of sample trajectory blending as a technique to smoothen changes in the optimal control between successive algorithm iterations.

7. Error analysis

The main sources of errors in the proposed numerical algorithm consist of (a) the time discretization scheme, and (b) the LSMC

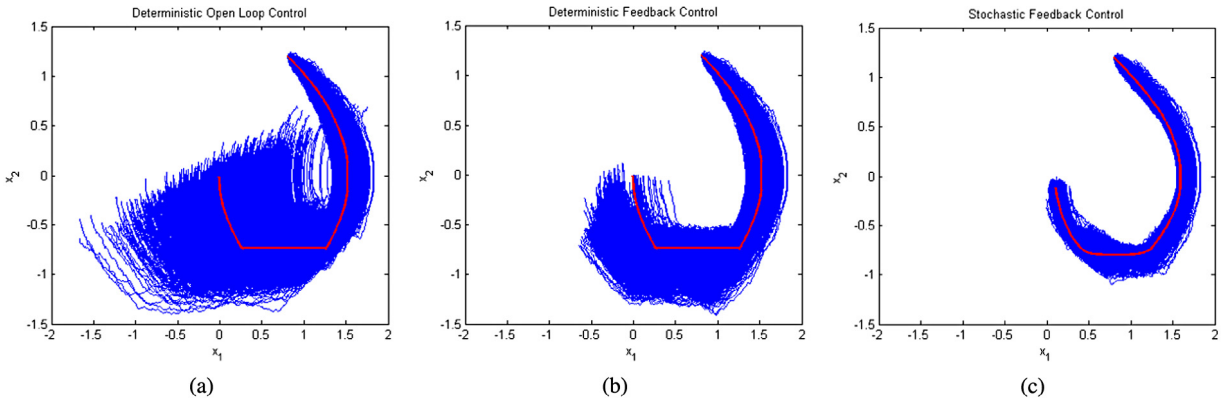

Fig. 2. Comparison between the deterministic open loop bang–bang control law (32) used in open loop fashion (a), in closed loop fashion (by recalculating the control at each time step) (b), and the stochastic feedback bang–bang control of the proposed algorithm (c). In (a) and (b), the red trajectory is the optimal trajectory of the deterministic system, while the blue trajectories result if this controlled system is influenced by noise. In (c), the red trajectory represents the mean controlled trajectory. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

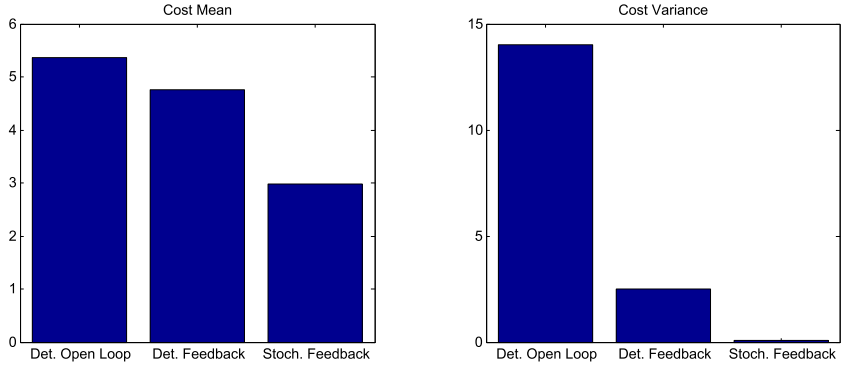

Fig. 3. Cost comparison between the deterministic open loop bang–bang control law (32) used in open loop, in closed loop, and the stochastic feedback bang–bang control of the proposed algorithm. Cost mean (left) and variance (right).

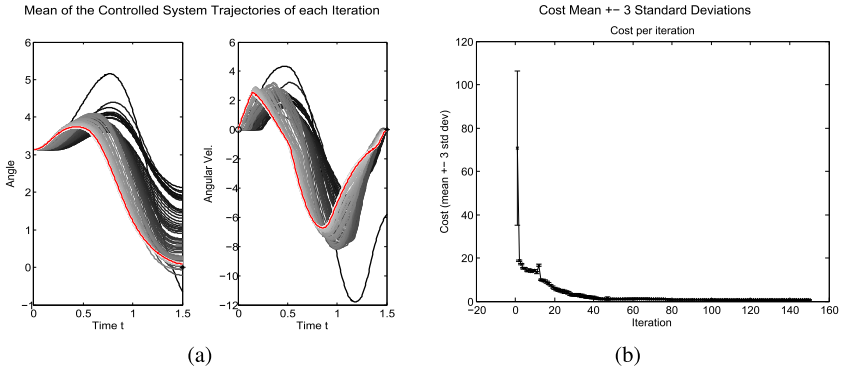

Fig. 4. The inverted pendulum system: The algorithm converges for a blending ratio of $\gamma = 0.98$. (a) The mean of the controlled system trajectories of each iteration (grayscale) and after the final iteration (red). (b) Cost mean $\pm$ 3 standard deviations per iteration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

method of approximating conditional expectations. The time discretization error in most schemes decreases at a rate $\sqrt{N}$, where N is the number of (equidistant) time steps [21]. The error due to the LSMC scheme can be reduced as the number of basis functions tends to infinity and is inversely proportional to the square root of the number of realizations, $\sqrt{M}$ [26]. Note that the PDE-FBSDE problem equivalence, illustrated by the nonlinear Feynman–Kac lemma (Section 3), being exact does not introduce any errors. Similarly, the importance sampling component, which is based on Girsanov's theorem (Section 5) is also exact; numerical differences appear only for finite number of samples. Proving the overall convergence of the proposed discretization scheme however is more involved, and is part of our on-going research.

8. Conclusion

In this paper we have developed a novel algorithm for nonlinear stochastic control problems in which the objective is to minimize a generalized L^1 norm of the control input. In light of a nonlinear version of the Feynman–Kac formula, and by utilizing the

connection between certain PDEs and SDEs, we were able to obtain a probabilistic representation of the solution to the Hamilton–Jacobi–Bellman equation, expressed as a system of FBSDEs. This system is then simulated using linear regression techniques.

Acknowledgments

This research was supported by ARO W911NF-16-1-0390 and NSF CMMI-1662523.

References

[1] M. Dixon, T. Edelbaum, J. Potter, W. Vandervelde, Fuel optimal reorientation of axisymmetric spacecraft, J. Spacecr. Rockets 7 (11) (1970) 1345–1351. http://dx.doi.org/10.2514/3.30168.
[2] H. Seywald, R.R. Kumar, S.S. Deshpande, M.L. Heck, Minimum fuel spacecraft reorientation, J. Guid. Control Dyn. 17 (1) (1994) 21–29. http://dx.doi.org/10.2514/3.21154.
[3] M.I. Ross, How to find minimum-fuel controllers, AIAA Guidance, Navigation, and Control Conference and Exhibit, Providence, RI, 16-19 August, http://dx.doi.org/10.2514/6.2004-5346.
[4] M. Nagahara, D.E. Quevedo, D. Nešić, Maximum hands-off control and L^1 optimality, in: 52nd IEEE Conference on Decision and Control, Florence, Italy, 2013, pp. 3825–3830.
[5] M. Nagahara, D.E. Quevedo, D. Nešić, Maximum hands-off control: a paradigm of control effort minimization, IEEE Trans. Automat. Control 61 (3) (2016) 735–747.
[6] B. Dunham, Automatic on/off switching gives 10-percent gas saving, Pop. Sci. 205 (4) (1974) 170.
[7] C.C. Chan, The state of the art of electric, hybrid, and fuel cell vehicles, Proc. IEEE 95 (4) (2007) 704–718.
[8] D.G. Jeong, W.S. Jeon, Performance of adaptive sleep period control for wireless communications systems, IEEE Trans. Wireless Commun. 5 (11) (2006) 3012–3016.
[9] L. Kong, G.K. Wong, D.H. Tsang, Performance study and system optimization on sleep mode operation in IEEE 802.16 e, IEEE Trans. Wireless Commun. 8 (9) (2009) 4518–4528.
[10] M. Athans, P. Falb, Optimal Control- An Introduction to the Theory and its Applications, Dover Publications, Inc., 2007.
[11] I. Exarchos, E. Theodorou, Stochastic optimal control via forward and backward stochastic differential equations and importance sampling, Automatica 87 (2018) 159–165. http://dx.doi.org/10.1016/j.automatica.2017.09.004.
[12] J. Yong, X.Y. Zhou, Stochastic Controls: Hamiltonian Systems and HJB Equations, Springer-Verlag New York Inc., 1999. http://dx.doi.org/10.1007/978-1-4612-1466-3.
[13] W. Fleming, H. Soner, Controlled Markov Processes and Viscosity Solutions, second ed., in: Stochastic Modelling and Applied Probability, Springer, 2006. http://dx.doi.org/10.1007/0-387-31071-1.
[14] I. Karatzas, S. Shreve, Brownian Motion and Stochastic Calculus, second ed., Springer-Verlag New York Inc., 1991.
[15] J. Ma, J. Yong, Forward-Backward Stochastic Differential Equations and Their Applications, Springer-Verlag Berlin Heidelberg, 1999. http://dx.doi.org/10.1007/978-3-540-48831-6.
[16] N. El Karoui, S. Peng, M.C. Quenez, Backward stochastic differential equations in finance, Math. Finance 7 (1) (1997). http://dx.doi.org/10.1111/1467-9965.00022.
[17] J.P. Lepeltier, J.S. Martìn, Existence for BSDE with superlinear–quadratic coefficient, Stochastics 63 (3–4) (1998) 227–240.
[18] M. Kobylanski, Backward stochastic differential equations and partial differential equations with quadratic growth, Ann. Probab. (2000) 558–602.
[19] B. Bouchard, N. Touzi, Discrete time approximation and Monte Carlo simulation of BSDEs, Stochastic Process. Appl. 111 (2) (2004) 175–206.
[20] C. Bender, R. Denk, A forward scheme for backward SDEs, Stochastic Process. Appl. 117 (12) (2007) 1793–1812.
[21] J.P. Lemor, E. Gobet, X. Warin, Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations, Bernoulli 12 (5) (2006) 889–916.
[22] P. Kloeden, E. Platen, Numerical Solution of Stochastic Differential Equations, third ed., in: Applications in Mathematics, Stochastic Modelling and Applied Probability, vol. 23, Springer-Verlag Berlin Heidelberg, 1999. http://dx.doi.org/10.1007/978-3-662-12616-5.
[23] F.A. Longstaff, R.S. Schwartz, Valuing American options by simulation: A simple least-squares approach, Rev. Financ. Stud. 14 (2001) 113–147.
[24] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer Series in Statistics, Springer-Verlag New York, Inc., 2002.
[25] B. Øksendal, Stochastic Differential Equations- An Introduction with Applications, sixth ed., Springer-Verlag Berlin Heidelberg, 2007.
[26] D. Xiu, Numerical Methods for Stochastic Computations- A Spectral Method Approach, Princeton University Press, 2010.