2016 American Control Conference (ACC)
Boston Marriott Copley Place
July 6-8, 2016. Boston, MA, USA

Learning Optimal Control via Forward and Backward Stochastic Differential Equations

Ioannis Exarchos[1] and Evangelos A. Theodorou[2]

Abstract— **In this paper we present a novel sampling-based numerical scheme designed to solve a certain class of 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 decoupled system of FBSDEs. This system of FBSDEs can then be simulated by employing linear regression techniques. To enhance the efficiency of the proposed scheme when treating more complex nonlinear systems, we then derive an iterative modification based on Girsanov's theorem on the change of measure, which features importance sampling. The modified scheme is capable of learning the optimal control without requiring an initial guess. We present simulations that validate the algorithm and demonstrate its efficiency in treating nonlinear dynamics.**

I. INTRODUCTION

The problem of obtaining an optimal control in a stochastic setting is typically associated with the solution of a generally nonlinear second-order partial differential equation (PDE), known as the Hamilton-Jacobi-Bellman (HJB) equation. Different methods can be distinguished based on whether they seek a solution over the entire domain, or locally around a nominal system trajectory. In the first case, several attempts have been made to address the inherent difficulty in solving such nonlinear PDEs, as well as the curse of dimensionality [1]–[5]. The latter case, on the other hand, includes methods such as Stochastic Differential Dynamic Programming [6], [7], which is based on linearization of the dynamics and a quadratic approximation of the cost functions around nominal trajectories, as well as sampling-based methods, which are distinguished for their ability to accommodate scalable iterative schemes. Various sampling-based methods appear in the literature under the names Path Integral control (PI) [8], [9], Kullback Leibler (KL) control, or Linearly Solvable Control [10], [11].

The fundamental characteristic of all aforementioned sampling-based methods is that they rely on the exponential transformation of the Value function [12], and the linear version of the Feynman-Kac lemma [13]. Under the exponential transformation, and by introducing certain restrictions between control authority and stochasticity, there exists a direct relationship between the Hamilton-Jacobi-Bellman PDE and the backward Chapman-Kolmogorov PDE. The latter, being a linear PDE, enables the use of the linear Feynman-Kac formula, which relates certain linear backward PDEs to forward stochastic differential equations (SDEs). Thus, the corresponding optimal control problem can be solved using forward sampling. While forward sampling-based methods exhibit several advantages against traditional methods of stochastic control, such as the mild conditions on the differentiability of the cost and the stochastic dynamics, there are also some key disadvantages which pertain to the nature of the exponential transformation. In particular, the effect of the exponential transformation can be identified as the mapping of the value function $v(t,x)$, which has range $[0,\infty)$, to the desirability function $\psi(t,x)$, whose range is $(0,1]$. This mapping leads to a drastic reduction in the ability to distinguish states with high cost (low desirability) from states with low cost (high desirability). This issue has been partially addressed with renormalization of the trajectory cost [7]. Finally, while the necessary constraint introduced between control authority and stochasticity can lead to symmetry breaking phenomena and delayed decision [14], it is a rather restrictive assumption whenever applications to engineered systems are considered.

In this paper we present a learning control algorithm which capitalizes on the innate relationship between certain *nonlinear* PDEs and Forward and Backward SDEs, demonstrated by a *nonlinear* version of the Feynman-Kac lemma. By means of this lemma, we obtain a probabilistic representation of the solution to the nonlinear HJB PDE, expressed in the form of a system of decoupled FBSDEs. This system of FBSDEs can then be simulatedby employing linear regression techniques. We wish to highlight the difference between the proposed approach and already existing sampling-based formulations: our approach addresses directly the nonlinear PDE, while the latter make use of the exponential transformation, which under certain conditions yields a linear PDE problem, and then use forward sampling to address that linear problem. Thus, the herein proposed framework relaxes these restrictive conditions. To enhance the efficiency of the proposed scheme when treating more complex nonlinear systems, we then derive an iterative algorithm based on Girsanov's theorem on the change of measure, which features importance sampling. The proposed scheme is capable of learning the optimal control without requiring an initial guess. We present simulations that validate the algorithm and demonstrate its efficiency in treating nonlinear dynamics.

II. PROBLEM STATEMENT

Let $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P})$ be a complete, filtered probability space on which is defined a p-dimensional standard Brown-

[1]Ioannis Exarchos is a PhD candidate at the Department of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0150 exarchos@gatech.edu

[2]Evangelos A. Theodorou is Assistant Professor at the Department of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0150 evangelos.theodorou@gatech.edu

ian motion W_t, such that $\{\mathscr{F}_t\}_{t\geq 0}$ is the natural filtration of W_t augmented by all $\mathbb{P}$-null sets. Consider the problem of minimizing the expected cost defined by the cost functional

$$J(\tau, x_\tau; u(\cdot)) = \mathbb{E}\left[g(x(T)) + \int_\tau^T q(t, x(t)) + \frac{1}{2}u^\top(t)Ru(t)\mathrm{d}t\right], \tag{1}$$

associated with the stochastic controlled system which is represented by the Itô stochastic differential equation (SDE)

$$\begin{aligned} &\mathrm{d}x(t) = f(t, x(t))\mathrm{d}t + G(t, x(t))u(t)\mathrm{d}t + \Sigma(t, x(t))\mathrm{d}W_t, \\ &t \in [\tau, T], \quad x(\tau) = x_\tau, \end{aligned} \tag{2}$$

where $T > \tau \geq 0$, T is a fixed time of termination, $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^\nu$ is the control vector, R is a $\nu \times \nu$ positive definite matrix, and $g : \mathbb{R}^n \to \mathbb{R}$, $q : [0, T]\times\mathbb{R}^n \to \mathbb{R}$, $f : [0, T]\times\mathbb{R}^n \to \mathbb{R}^n$, $G : [0, T]\times\mathbb{R}^n \to \mathbb{R}^{n\times\nu}$, and $\Sigma : [0, T] \times \mathbb{R}^n \to \mathbb{R}^{n\times p}$ are deterministic functions, that is, they do not depend explicitly on $\omega \in \Omega$. We assume that all standard technical conditions [15] which pertain to the filtered probability space and the regularity of functions are met, in order to guarantee existence, uniqueness of solutions to (2), and a well defined cost functional (1). These impose for example that the functions g, q, f, G and Σ are continuous w.r.t. time t (in case there is explicit dependence), Lipschitz (uniformly in t) with respect to the state variables and satisfy standard growth conditions over the domain of interest. Furthermore, the square-integrable process $u : [0, T] \times \Omega \to U \subseteq \mathbb{R}^\nu$ is $\{\mathscr{F}_t\}_{t\geq 0}$-adapted (also called *progressively measurable*), which essentially translates into the control input being non-anticipating, i.e., relying only on past and present information. We denote the set of all admissible U-valued functions as $\mathscr{U}[0, T]$. For any given initial condition (τ, x_τ), we wish to minimize (1) under all admissible functions $u(\cdot) \in \mathscr{U}[0, T]$. We define the Value function V as

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

By applying the stochastic version of Bellman's principle of optimality, it is shown [15], [16] that if the Value function is in $C^{1,2}([0, T] \times \mathbb{R}^n)$, then it is a solution to the following terminal value problem of a second-order partial differential equation, known as the Hamilton-Jacobi-Bellman (HJB) equation, which, for the problem at hand, and suppressing function arguments for notational compactness, takes the form

$$\begin{aligned} &\left\{ v_t + \inf_{u\in U}\left\{\frac{1}{2}\mathrm{tr}(v_{xx}\Sigma\Sigma^\top) + v_x^\top(f + Gu) + q + \frac{1}{2}u^\top Ru\right\}\right. \\ &\left. = 0, \ (t, x) \in [0, T) \times \mathbb{R}^n, \quad v(T, x) = g(x), \quad x \in \mathbb{R}^n. \right. \end{aligned} \tag{4}$$

where v_x and v_{xx} denote the gradient and the Hessian of v, respectively. The term inside the brackets is the Hamiltonian, and is denoted by H. Note that this result can be extended to include cases where the Value function does not satisfy the smoothness condition. Then, if one also considers viscosity solutions of (4), the Value function is proven to be a viscosity solution of (4). Furthermore, the viscosity solution is equal to the classical solution, if a classical solution exists. For the chosen form of the cost integrand, and assuming that the optimal control lies in the interior of U, we may carry out the infimum operation by taking the gradient of the Hamiltonian with respect to u and setting it equal to zero to obtain

$$\frac{\partial H}{\partial u} = 0 \quad \text{or} \quad -Ru - G^\top(t, x)v_x(t, x) = 0. \tag{5}$$

Therefore, the optimal control is given by

$$u^*(t, x) = -R^{-1}G^\top(t, x)v_x(t, x), \quad (t, x) \in [0, T] \times \mathbb{R}^n. \tag{6}$$

Inserting the above expression back into the original HJB equation and suppressing again function arguments, we obtain the equivalent characterization

$$\begin{cases} v_t + \frac{1}{2}\mathrm{tr}(v_{xx}\Sigma\Sigma^\top) + v_x^\top f + q - \frac{1}{2}v_x^\top GR^{-1}G^\top v_x = 0, \\ (t, x) \in [0, T) \times \mathbb{R}^n, \quad v(T, x) = g(x), \quad x \in \mathbb{R}^n. \end{cases} \tag{7}$$

III. A FEYNMAN-KAC TYPE REPRESENTATION THROUGH FBSDES

There is an innate relation between stochastic differential equations and second-order partial differential equations of parabolic or elliptic type. Specifically, solutions to a certain class of nonlinear PDEs can be represented by solutions to forward-backward stochastic differential equations (FBSDEs), in the same spirit as demonstrated by the well-known Feynman-Kac formulas [13] for linear PDEs. We begin by briefly reviewing FBSDEs.

A. The Forward and Backward Process

As a forward process we shall define the square-integrable, $\{\mathscr{F}_s\}_{s\geq 0}$-adapted process $X(\cdot)$[1], which, for any given initial condition $(t, x) \in [0, T] \times \mathbb{R}^n$, satisfies the Itô FSDE

$$\begin{cases} \mathrm{d}X_s = b(s, X_s)\mathrm{d}s + \Sigma(s, X_s)\mathrm{d}W_s, & s \in [t, T], \\ X_t = x. \end{cases} \tag{8}$$

The forward process (8) is also called the *state process* in the literature. We shall denote the solution to the forward SDE (8) as $X_s^{t,x}$, wherein (t, x) are the initial condition parameters.

In contrast to the forward process, 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

$$\begin{cases} \mathrm{d}Y_s = -h(s, X_s, Y_s, Z_s)\mathrm{d}s + Z_s^\top \mathrm{d}W_s & s \in [t, T], \\ Y_T = g(X_T). \end{cases} \tag{9}$$

The function $h(\cdot)$ is called *generator* or *driver*. The solution is implicitly defined by the initial condition parameters (t, x) of the FSDE since it obeys the terminal condition $g(X_T^{t,x})$. We will similarly use the notation $Y_s^{t,x}$ and $Z_s^{t,x}$ to denote

[1]While X is a function of s and ω, we shall use X_s for notational brevity.

the solution for a particular initial condition parameter (t, x) of the associated FSDE.

While FSDEs have a fairly straightforward definition, in the sense that both the SDE and the filtration evolve forward in time, this is not the case for BSDEs. Indeed, since solutions to BSDEs need to satisfy a terminal condition, integration needs to be performed backwards in time in some sense, yet the filtration still evolves forward in time. It turns out [17] that a terminal value problem involving BSDEs admits an adapted (i.e., non-anticipating) solution if we back-propagate the *conditional expectation* of the process, that is, if we set $Y_s \triangleq \mathbb{E}[Y_s|\mathscr{F}_s]$.

Notice that the FSDE does not depend on Y_s or Z_s. Thus, the resulting system of FBSDEs is said to be *decoupled*. If, in addition, the functions b, Σ, h and g are deterministic, in the sense that they do not depend explicitly on $\omega \in \Omega$, then the adapted solution (Y, Z) exhibits the *Markovian* property; namely, it can be written as deterministic functions of solely time and the state process [18]:

Theorem 1: (The Markovian Property) – There exist deterministic functions $v(t, x)$ and $d(t, x)$[2] such that the solution $(Y^{t,x}, Z^{t,x})$ of the BSDE (9) is

$$Y_s^{t,x} = v(s, X_s^{t,x}), \quad Z_s^{t,x} = \Sigma^\top(s, X_s^{t,x})d(s, X_s^{t,x}), \quad (10)$$

for all $s \in [t, T]$.

B. The Nonlinear Feynman-Kac Lemma

We now proceed to state the nonlinear Feynman-Kac type formula, which links the solution of a class of PDEs to that of FBSDEs. Indeed, the following theorem can be proven [15], [17], [18] by an application of Itô's formula:

Theorem 2: (Nonlinear Feynman-Kac) – Let $v : [0, T] \times \mathbb{R}^n \to \mathbb{R}$ be a function of class $C^{1,2}$ which satisfies the Cauchy problem

$$\begin{cases} v_t + \frac{1}{2}\mathrm{tr}(v_{xx}\Sigma\Sigma^\top) + v_x^\top b(t, x) + h(t, x, v, \Sigma^\top v_x) = 0, \\ (t, x) \in [0, T) \times \mathbb{R}^n, \quad v(T, x) = g(x), \quad x \in \mathbb{R}^n, \end{cases} \quad (11)$$

wherein the functions Σ, b, h and g satisfy mild regularity conditions. Then (11) admits a unique viscosity solution $v : [0, T] \times \mathbb{R}^n \to \mathbb{R}$, which has the following probabilistic representation:

$$v(t, x) = Y_t^{t,x}, \qquad \forall (t, x) \in [0, T] \times \mathbb{R}^n, \quad (12)$$

where $(X(\cdot), Y(\cdot), Z(\cdot))$ is the unique adapted solution of the FBSDE system (8)-(9). Furthermore,

$$(Y_s^{t,x}, Z_s^{t,x}) = \Big(v(s, X_s^{t,x}),\ \Sigma^\top(s, X_s^{t,x})v_x(s, X_s^{t,x})\Big), \quad (13)$$

for all $s \in [t, T]$, and if (11) admits a classical solution, then (12) provides that classical solution.

A comparison of equations (7) and (11) indicates that the nonlinear Feynman-Kac representation can be applied to the HJB equation given by (7) 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 $(t, x) \in [0, T] \times \mathbb{R}^n$, satisfying the same mild regularity conditions.

This assumption implies that the range of G must be a subset of the range of Σ, and thus excludes the case of a channel containing control input but no noise, although the converse is allowed. Under this assumption, and suppressing arguments for brevity, the HJB equation given by (7) becomes

$$\begin{cases} v_t + \frac{1}{2}\mathrm{tr}(v_{xx}\Sigma\Sigma^\top) + v_x^\top f + q - \frac{1}{2}v_x^\top \Sigma\Gamma R^{-1}\Gamma^\top\Sigma^\top v_x \\ = 0, \\ (t, x) \in [0, T) \times \mathbb{R}^n, \quad v(T, x) = g(x), \quad x \in \mathbb{R}^n, \end{cases} \quad (14)$$

which satisfies the format of (11) with

$$b(t, x) \equiv f(t, x) \quad (15)$$

$$h(t, x, z) \equiv q(t, x) - \frac{1}{2}z^\top\Gamma(t, x)R^{-1}\Gamma^\top(t, x)z. \quad (16)$$

We may thus obtain the (viscosity) solution of (14) by simulating the system of FBSDE given by (8) and (9). Notice that (8) corresponds to the uncontrolled ($u = 0$) system dynamics.

IV. Approximating the Solution of FBSDEs

The solution of FBSDEs has been studied to a great extent independently from its connection to PDEs, mainly within the field of mathematical finance. Though several generic schemes exist [19]–[21], in this paper we propose a scheme which exploits the regularity present in FBSDEs that arise from the application of the nonlinear Feynman-Kac lemma.

We begin by selecting a time grid $\{t = t_0 < \ldots < t_N = T\}$ for the interval $[t, T]$, and 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[3]. For notational brevity, we also denote $X_i \triangleq X_{t_i}$. The simplest discretized scheme for the forward process is the Euler scheme, which is also called *Euler-Maruyama* scheme [22]:

$$\begin{cases} X_{i+1} \approx X_i + b(t_i, X_i)\Delta t_i + \Sigma(t_i, X_i)\Delta W_i, \\ i = 1, \ldots, N, \qquad X_0 = x. \end{cases} \quad (17)$$

Several alternative, higher order schemes exist that can be selected in lieu of the Euler scheme [22]. To discretize the backward process, we further introduce the notation $Y_i \triangleq Y_{t_i}$ and $Z_i \triangleq Z_{t_i}$. Then, recalling that adapted BSDE solutions impose $Y_s \triangleq \mathbb{E}[Y_s|\mathscr{F}_s]$ and $Z_s \triangleq \mathbb{E}[Z_s|\mathscr{F}_s]$ (i.e., a back-propagation of the conditional expectations), we approximate equation (9) by

$$Y_i = \mathbb{E}[Y_i|\mathscr{F}_{t_i}] \approx \mathbb{E}[Y_{i+1} + h(t_{i+1}, X_{i+1}, Y_{i+1}, Z_{i+1})\Delta t_i|X_i]. \quad (18)$$

Notice that in the last equality the term $Z_i^\top \Delta W_i$ in (9) vanishes because of the conditional expectation (ΔW_i is

[2] By abuse of notation, here (t, x) are symbolic arguments of the functions v and d, and not the initial condition parameters as in $(Y^{t,x}, Z^{t,x})$. Throughout this work, it should be clear from the context whether (t, x) are to be understood as initial condition parameters or symbolic arguments.

[3] Here, ΔW_i would be simulated as $\sqrt{\Delta t_i}\xi_i$, where $\xi_i \sim \mathcal{N}(0, I)$.

zero mean), and we replace $\mathscr{F}_{t_i}$ with X_i in light of the Markovian property presented in Section III-A. By virtue of equation (13), the Z-process in (9) corresponds to the term $\Sigma^\top(s, X_s^{t,x})v_x(s, X_s^{t,x})$. Therefore we can write

$$\begin{aligned} Z_i &= \mathbb{E}[Z_i|\mathscr{F}_{t_i}] = \mathbb{E}[\Sigma^\top(t_i, X_i)\nabla_x v(t_i, X_i)|X_i] \\ &= \Sigma^\top(t_i, X_i)\nabla_x v(t_i, X_i), \end{aligned} \tag{19}$$

which naturally requires knowledge of the solution at time t_i on a neighborhood x, $v(t_i, x)$. The backpropagation is initialized at $Y_T = g(X_T)$ and $Z_T = \Sigma(T, X_T)^\top \nabla_x g(X_T)$, for a $g(\cdot)$ which is differentiable almost everywhere. There are several ways to approximate the conditional expectation in (18), however in this work we shall employ the Least Squares Monte Carlo (LSMC) method[4], which we shall briefly review in what follows.

The LSMC method addresses the general problem of numerically estimating conditional expectations of the form $\mathbb{E}[Y|X]$ for square integrable random variables X and Y, if one is able to sample M independent copies of pairs (X, Y). The method itself is based on the principle that the conditional expectation of a random variable can be modeled as a function of the variable on which it is conditioned on, that is, $\mathbb{E}[Y|X] = \phi^*(X)$, where ϕ^* solves the infinite dimensional minimization problem

$$\phi^* = \arg\min_{\phi} \mathbb{E}[|\phi(X) - Y|^2], \tag{20}$$

and ϕ ranges over all measurable functions with $\mathbb{E}[|\phi(X)|^2] < \infty$. A finite-dimensional approximation of this problem can be obtained if one decomposes $\phi(\cdot) = \sum_{i=1}^k \varphi_i(\cdot)\alpha_i = \varphi(\cdot)\alpha$, with $\varphi(\cdot)$ being a row vector of predetermined basis functions and α a column vector of constants, thus solving $\alpha^* = \arg\min_{\alpha\in\mathbb{R}^k} \mathbb{E}[|\varphi(X)\alpha - Y|^2]$, with k being the dimension of the basis. Finally, this problem can be simplified to a linear least-squares problem if one substitutes the expectation operator with its empirical estimator [24], thus obtaining

$$\alpha^* = \arg\min_{\alpha\in\mathbb{R}^k} \frac{1}{M}\sum_{j=1}^{M} |\varphi(X^j)\alpha - Y^j|^2, \tag{21}$$

wherein (X^j, Y^j), $j = 1, \dots, M$ are independent copies of (X, Y). Introducing the notation

$$\Phi(X) = \begin{bmatrix} \varphi(X^1) \\ \vdots \\ \varphi(X^M) \end{bmatrix} \in \mathbb{R}^{M\times k}, \tag{22}$$

the solution to this least-squares problem can be obtained by directly solving the normal equation, i.e.,

$$a^* = \Big(\Phi^\top(X)\Phi(X)\Big)^{-1}\Phi^\top(X)\left(\begin{bmatrix} Y^1 \\ \vdots \\ Y^M \end{bmatrix}\right), \tag{23}$$

or by performing gradient descent. The LSMC estimator for the conditional expectation assumes then the form $\mathbb{E}[Y|X = x] = \phi^*(x) \approx \varphi(x)a^*$.

[4]Treating conditional expectations by means of linear regression was made popular in the field of mathematical finance by [23].

Returning to our problem, we may apply the LSMC method to approximate the conditional expectation in equation (18) for each time step. To this end, we require a vector of basis functions φ for the approximation of $\mathbb{E}[Y_i|X_i]$. Although the basis functions can be different at each time step, we shall use the same symbol for notational simplicity. Then, Monte Carlo simulation is performed by sampling M independent trajectories $\{X_i^m\}_{i=1,\dots,N}$, in which the index $m = 1, \dots, M$ specifies a particular Monte Carlo trajectory. Whenever this index is not present, the entirety with respect to this index is to be understood. The numerical scheme is initialized at the terminal time T and is iterated backwards along the entire time grid, until the starting time instant has been reached. At each time step t_i, we are given M pairs of data (Y_i^m, X_i^m)[5] on which we perform linear regression to estimate the conditional expectation of Y_i as a function of x at the time step t_i. This provides us an approximation of the Value function v at time t_i for the neighborhood of the state space that has been explored by the sample trajectories at that time instant, since $v(t_i, x) = \mathbb{E}[Y_i|X_i = x] \approx \varphi(x)\alpha_i$. We then replace $Y_i^m = \mathbb{E}[Y_i^m|X_i^m] \approx \varphi(X_i^m)\alpha_i$, thereby treating the conditional expectation as a projection operator. Finally, the approximation of the conditional expectation of Z_i is obtained by taking the gradient with respect to x on $v(t_i, x)$, evaluating it at X_i^m, and scaling it with Σ

$$Z_i^m \approx \Sigma(t_i, X_i^m)^\top \nabla_x \varphi(X_i^m)\alpha_i. \tag{24}$$

This process is repeated for $t_{i-1}, \dots, t_1$. Note that this approach requires the basis functions $\varphi(\cdot)$ of our choice to be differentiable almost everywhere, so that $\nabla_x\varphi(x)$ is available in analytical form for almost any x. The proposed algorithm is then summarized as

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

where $m = 1, \dots, M$ and the matrix Φ defined in (22). Again, the minimizer can be obtained by directly solving the normal equation, i.e.,

$$\begin{aligned} a_i = \Big(\Phi^\top(X_i)\Phi(X_i)\Big)^{-1}\Phi^\top(X_i) \\ \Big(Y_{i+1} + \Delta t_i h(t_{i+1}, X_{i+1}, Y_{i+1}, Z_{i+1})\Big), \end{aligned} \tag{25}$$

or by performing gradient descent. The essential algorithm output is the collection of a_i's, that is, the basis function coefficients at each time instant, which are needed to recover the Value function approximation for the particular area of the state space that is explored by the forward process. This is in contrast with methods that calculate the solution over

[5]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 the conditional expectation operator has been applied.

an entire pre-specified grid (and thus typically exhibit bad scalability), but also differs from local trajectory optimization methods which consider only infinitesimal variations around a nominal trajectory. Furthermore, an important difference between the proposed method and forward sampling based methods is that the latter provide a solution only for the point of the initial condition (t, x), while the solution of this method covers an area starting from the initial condition (t, x), expanding in state space until time T is reached.

V. LEARNING CONTROL: AN ITERATIVE SCHEME BASED ON IMPORTANCE SAMPLING

The proposed method, as it has been presented so far, suffers from a significant limitation. Namely, its ability to provide approximations to the value function is restricted to only those areas of the state space that are reachable by unforced dynamics (eq. (8)). Indeed, there are several cases of systems in which the goal state practically cannot be reached by the uncontrolled system dynamics (consider, for example, an inverted pendulum). Furthermore, even in the case in which the target state is indeed reached by unforced trajectories, as the dimensionality of the state space increases, the density of sample trajectories along any given path from the initial state to the target state reduces quickly, thus increasing the demand for available samples. These issues can be eliminated if one is given the ability to modify the drift term of the sampled trajectories. Specifically, by changing the drift, we can direct the exploration of the state space towards the goal state, or any other state of interest, reachable by control. As will be shortly demonstrated, such a scheme can indeed be constructed by means of a careful application of Girsanov's theorem on the change of measure. Applying importance sampling on FBSDEs is not an entirely new concept, as it was first introduced as a variance reduction technique [25]. Through Girsanov's theorem, one may alter the drift of the forward process if this modification is appropriately compensated for in the backward process. That is, the system of FBSDEs given by equations (8) and (9) is in some certain sense equivalent to one with modified drift

$$\begin{cases} \mathrm{d}\tilde{X}_s = [b(s,\tilde{X}_s) + \Sigma(s,\tilde{X}_s)K_s]\mathrm{d}s + \Sigma(s,\tilde{X}_s)\mathrm{d}W_s, \\ s \in [t,T], \quad \tilde{X}_t = x. \end{cases} \tag{26}$$

along with the compensated BSDE

$$\begin{cases} \mathrm{d}\tilde{Y}_s = [-h(s,\tilde{X}_s,\tilde{Y}_s,\tilde{Z}_s) + \tilde{Z}_s^\top K_s]\mathrm{d}s + \tilde{Z}_s^\top \mathrm{d}W_s, \\ s \in [t,T], \quad \tilde{Y}_T = g(\tilde{X}_T), \end{cases} \tag{27}$$

for any measurable, bounded and adapted process $K : [0,T] \to \mathbb{R}^p$. Equivalence is not path-wise of course, since the paths realized by both the forward and the backward processes will be different under the modified drift dynamics. However, the solution at starting time t, that is (Y_t, Z_t), will remain unaffected. In other words, the estimate of the Value function at the initial condition (t, x) is independent of the drift term modification, as will be proven shortly. Indeed, following Girsanov's Theorem [13], [26], we define a new measure $\mathbb{Q}$ with $\mathrm{d}\mathbb{Q}(\omega) = M(T;t,\omega)\mathrm{d}\mathbb{P}(\omega)$, where

$$M_s \triangleq \exp\left(-\int_t^s K_\tau^\top \mathrm{d}W_\tau - \frac{1}{2}\int_t^s |K_\tau|^2 \mathrm{d}\tau\right), \quad s \in [t,T],$$

is the process of Radon-Nikodym derivatives $\mathrm{d}\mathbb{Q}^{(s)}/\mathrm{d}\mathbb{P}^{(s)}$ with $\mathbb{Q}^{(s)}$ and $\mathbb{P}^{(s)}$ being the restrictions of $\mathbb{Q}$ and $\mathbb{P}$ to $\mathscr{F}_s$, respectively. Then, M_s is a $\mathbb{P}$-martingale, the $\mathbb{P}$-law of (X, Y, Z) is the same as the $\mathbb{Q}$-law of $(\tilde{X}, \tilde{Y}, \tilde{Z})$, and

$$\tilde{W}_s \triangleq \int_t^s K_\tau \mathrm{d}\tau + W_s, \qquad s \in [t,T],$$

is a Brownian motion under $\mathbb{Q}$. In fact, defining the $\mathbb{Q}$-Brownian increment $\mathrm{d}\tilde{W}_s = K_s\mathrm{d}t + \mathrm{d}W_s$, it becomes evident that equations (26) and (27) are simply copies of the dynamics of equations (8) and (9), if one substitutes $\mathrm{d}W_s$ in the latter with $\mathrm{d}\tilde{W}_s$. Now, notice that since at the time of initialization, t, M_t is by construction equal to one with probability one (in both $\mathbb{P}$ and $\mathbb{Q}$-measure), the measures $\mathbb{P}$ and $\mathbb{Q}$ restricted to $\mathscr{F}_t$ are equal, and therefore the pairs (Y_t, Z_t) and $(\tilde{Y}_t, \tilde{Z}_t)$ are equal in expectation as well. This proves that the Value function at the initial condition (t, x) is independent of the drift term modification. An additional intuitive, albeit informal, explanation of why the modified system of FBSDEs (26), (27) can be used in lieu of the original FBSDE system is readily obtained if one examines the associated PDEs. Indeed, the FBSDE problem defined by (26) and (27) corresponds to the PDE problem

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

which of course is identical to the PDE problem (11), as we have merely added and subtracted the term $v_x^\top \Sigma K$. Thus, although the FBSDEs are different, they are associated with the same PDE problem.

Returning to the original problem formulation and recalling the definition of $\Gamma(\cdot)$ in Assumption 1, we may apply any nominal control $\bar{u}$ to the state dynamics in order to obtain the modified drift system, which exhibits the form

$$\mathrm{d}x(t) = [f(t,x(t)) + \Sigma(t,x(t))\Gamma(t,x(t))\bar{u}]\mathrm{d}t + \Sigma(t,x(t))\mathrm{d}W_t. \tag{28}$$

Thus, the controlled system trajectories are samples from the forward process (26) with

$$K_s = \Gamma(s,X_s)\bar{u}(s,X_s), \qquad s \in [t,T], \tag{29}$$

while $b(s,X_s) \equiv f(s,X_s)$ as per (15). Notice that the nominal control $\bar{u}$ may be any open-loop control, a random control, or even a control calculated by a previous run of the algorithm. In the latter case, one obtains a more refined solution, thus arriving at an iterative scheme. For the discrete representation on the time grid of Section IV, we define $K_i = K_{t_i}$. The forward process can again be sampled using the Euler-Maruyama scheme. There are several equivalent ways to incorporate importance sampling in the backward process, however the most straightforward way is to simply define

$$\tilde{h}(s,x,y,z,k) \triangleq h(s,x,y,z) - z^\top k, \tag{30}$$

and utilize the discretized scheme presented in Section IV using $\tilde{h}$ instead of h.

VI. SIMULATION RESULTS

To evaluate the algorithm's performance, we simulated the algorithm on an inverted pendulum and a cart-pole system. These simulations demonstrate that the nonlinearity in the dynamics is handled efficiently, and furthermore illustrate the significance of the iterative scheme which features importance sampling.

A. The Inverted Pendulum

The equations of motion for the inverted pendulum are given by

$$m\ell^2\ddot{\theta} + b\dot{\theta} - mg\ell\sin\theta = u, \tag{31}$$

and stochasticity enters the system in form of perturbations in the torque u. For the purposes of this simulation, two thousand trajectories were generated on a time grid of 0.004 with time horizon T=2. The system noise covariance was set on 0.1. No initial guess for the control input was necessary, though a white noise signal has been injected in the control input during the sampling stages to increase variation in the trajectories, since the system noise intensity is low. For the basis of the Value function approximation, modified Chebyshev polynomials [27] up to second order have been selected. The scheme was repeated for 15 iterations, with the algorithm successfully learning the optimal control to invert and stabilize the pendulum. Figure 1 depicts the mean of the controlled trajectories for each algorithm iteration (gray scale). The trajectories after the final iteration are shown in red. Finally, Figure 2 depicts the convergence of the cost mean and standard deviation as the iterative scheme progresses.

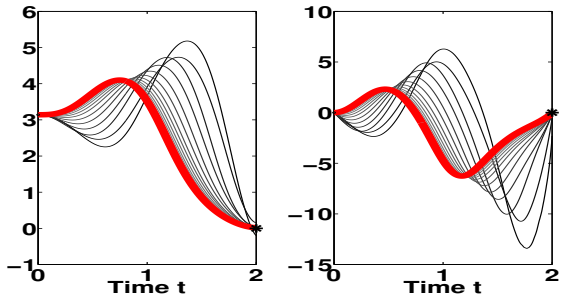

Fig. 1. Inverted Pendulum: Trajectory mean for the position (left) and velocity (right) of the controlled system for each iteration (gray scale) and after the final iteration (red). The black dots represent the target states.

B. The Cart-Pole system

To assess the efficiency of the proposed scheme in under-actuated systems, we simulated the algorithm on a cart-pole system (see Figure 3). The equations of motion are given by

$$\ddot{x} = \frac{1}{m_c + m_p\sin^2\theta}\Big(u - m_p\sin\theta(\ell\dot{\theta}^2 + g\cos\theta)\Big), \tag{32}$$

$$\ddot{\theta} = \frac{1}{\ell(m_c + m_p\sin^2\theta)}\Big(u\cos\theta - m_p\ell\dot{\theta}^2\cos\theta\sin\theta + (m_c + m_p)g\sin\theta\Big), \tag{33}$$

and stochasticity enters the system in form of perturbations in u. To this end, five thousand trajectories were generated on a time grid of 0.004 with time horizon T=3. The system noise covariance was set on 1. Again, no initial guess for the control input was necessary, and a white noise signal has been injected in the control input during the sampling stages to increase variation in the trajectories, since the system noise intensity is low. For the basis of the Value function approximation, modified Chebyshev polynomials up to second order have been selected. The scheme was repeated for 35 iterations. Figure 4 depicts the mean of the controlled trajectories for each algorithm iteration (gray scale). The trajectories after the final iteration are shown in red. Finally, Figure 5 depicts the convergence of the cost mean and standard deviation as the iterative scheme progresses.

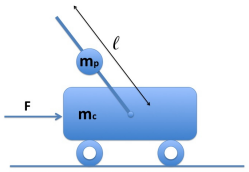

Fig. 3. Cart pole: m_c denoted the mass of the cart, m_p denotes the mass of the pole and ℓ is the length of the pole.

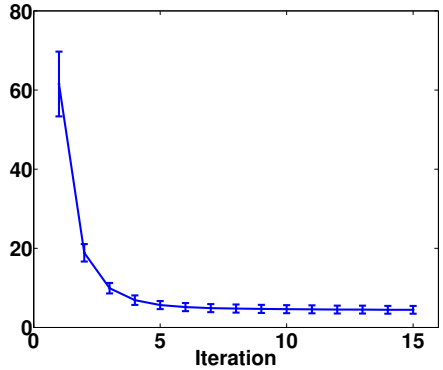

Fig. 2. Inverted Pendulum: Cost mean +- 3 standard deviations per iteration.

VII. CONCLUSIONS

In this paper we proposed a new algorithm for nonlinear stochastic control problems with dynamics affine in control and cost functions that are non-quadratic in the state and

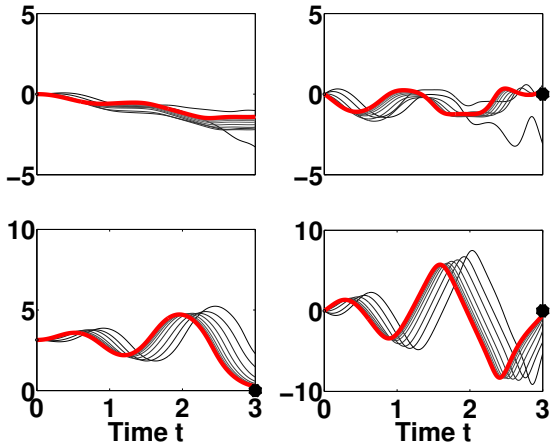

Fig. 4. Cart-pole: Clockwise starting at the top left– cart position, cart velocity, pole velocity, pole position. Trajectory mean of the controlled system for each iteration (gray scale) and after the final iteration (red). The black dots represent the target states.

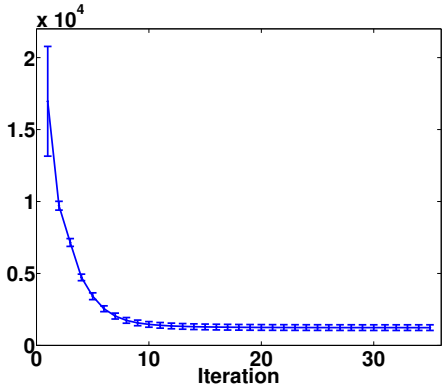

Fig. 5. Cart-pole: Cost mean +- 3 standard deviations per iteration.

quadratic in controls. In light of a nonlinear Feynman-Kac lemma which establishes a connection between certain PDEs and SDEs, we were able to obtain a probabilistic representation of the solution to the nonlinear HJB PDE, expressed as a system of FBSDEs. This system is then simulated using linear regression. Finally, in order to enhance the algorithm's efficiency in treating more complex nonlinear systems, we proposed an iterative scheme based on Girsanov's theorem on the change of measure, which features importance sampling. We demonstrated the ability of the proposed iterative algorithm to learn the optimal controls without an initial guess in an inverted pendulum system and a cart-pole system.

ACKNOWLEDGMENT

The first author gratefully acknowledges support from the A. S. Onassis Foundation.

REFERENCES

[1] I. M. Mitchell and C. J. Tomlin, "Overapproximating reachable sets by Hamilton-Jacobi projections," *Journal of Scientific Computing*, vol. 19, no. 1-3, pp. 323–346, 2003.

[2] C. O. Aguilar and A. J. Krener, "Numerical solutions to the Bellman equation of optimal control," *Journal of Optimization Theory and Applications*, vol. 160, no. 2, pp. 527–552, 2014.

[3] M. B. Horowitz and J. W. Burdick, "Semidefinite relaxations for stochastic optimal control policies," in *American Controls Conference (ACC)*, pp. 3006–3012, 2014.

[4] M. B. Horowitz, A. Damle, and J. W. Burdick, "Linear Hamilton Jacobi Belman equations in high dimensions," in *53rd IEEE Conference on Decision and Control, Los Angeles, California, USA*, December 15-17 2014.

[5] A. Gorodetsky, S. Karaman, and Y. Marzouk, "Efficient high-dimensional stochastic optimal motion control using tensor-train decomposition," in *Robotics: Science and Systems (RSS)*, 2015.

[6] E. Todorov and W. Li, "A generalized iterative LQG method for locally optimal feedback control of constrained nonlinear stochastic systems," *American Control Conference*, pp. 300–306, 2005.

[7] E. A. Theodorou, Y. Tassa, and E. Todorov, "Stochastic differential dynamic programming," *American Control Conference*, pp. 1125–1132, 2010.

[8] H. J. Kappen, "Linear theory for control of nonlinear stochastic systems," *Physical Review Letters*, vol. 95, November 2005.

[9] E. A. Theodorou, J. Buchli, and S. Schaal, "A generalized path integral control approach to reinforcement learning," *The Journal of Machine Learning Research*, vol. 11, pp. 3137–3181, January 2010.

[10] K. Dvijotham and E. Todorov, "Linearly solvable optimal control," *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, pp. 119–141, 2012.

[11] E. Todorov, "Efficient computation of optimal actions," *Proceedings of the National Academy of Sciences*, vol. 106, no. 28, pp. 11478–11483, 2009.

[12] E. A. Theodorou, "Nonlinear stochastic control and information theoretic dualities: Connections, interdependencies and thermodynamic interpretations," *Entropy*, vol. 17, no. 5, pp. 3352–3375, 2015.

[13] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*. Springer-Verlag New York Inc., 2nd ed., 1991.

[14] H. J. Kappen, "Path integrals and symmetry breaking for optimal control theory," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 11, November 2005.

[15] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*. Springer-Verlag New York Inc., 1999.

[16] W. Fleming and H. Soner, *Controlled Markov Processes and Viscosity Solutions*. Stochastic Modelling and Applied Probability, Springer, 2nd ed., 2006.

[17] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*. Springer-Verlag Berlin Heidelberg, 1999.

[18] N. El Karoui, S. Peng, and M. C. Quenez, "Backward stochastic differential equations in finance," *Mathematical Finance*, vol. 7, January 1997.

[19] B. Bouchard and N. Touzi, "Discrete time approximation and Monte Carlo simulation of BSDEs," *Stochastic Processes and their Applications*, vol. 111, pp. 175–206, June 2004.

[20] C. Bender and R. Denk, "A forward scheme for backward SDEs," *Stochastic Processes and their Applications*, vol. 117, pp. 1793–1812, December 2007.

[21] J. P. Lemor, E. Gobet, and X. Warin, "Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations," *Bernoulli*, vol. 12, no. 5, pp. 889–916, 2006.

[22] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, vol. 23 of *Applications in Mathematics, Stochastic Modelling and Applied Probability*. Springer-Verlag Berlin Heidelberg, 3rd ed., 1999.

[23] F. A. Longstaff and R. S. Schwartz, "Valuing American options by simulation: A simple least-squares approach," *Review of Financial Studies*, vol. 14, pp. 113–147, 2001.

[24] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics, Springer-Verlag New York, Inc., 2002.

[25] T. Moseler and C. Bender, "Importance sampling for backward SDEs," *Stochastic Analysis and Applications*, vol. 28, no. 2, pp. 226–253, 2010.

[26] B. Øksendal, *Stochastic Differential Equations- An Introduction with Applications*. Springer-Verlag Berlin Heidelberg, 6th ed., 2007.

[27] J. T. King, *Introduction to Numerical Computation*. McGraw-Hill, Inc., 1984.