Game-Theoretic and Risk-Sensitive Stochastic Optimal Control via Forward and Backward Stochastic Differential Equations

Ioannis Exarchos[1] Evangelos A. Theodorou[2] Panagiotis Tsiotras[3]

***Abstract*— In this work we present a sampling-based algorithm designed to solve game-theoretic control problems and risk-sensitive stochastic optimal control problems. The cornerstone of the proposed approach is the formulation of the problem in terms of 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-Isaacs equation, expressed in the form of a decoupled system of FBSDEs. This system of FBSDEs can then be simulated by employing linear regression techniques. Utilizing the connection between stochastic differential games and risk-sensitive optimal control, we demonstrate that the proposed algorithm is also applicable to the latter class of problems. Simulation results validate the algorithm.**

I. Introduction

Game-theoretic or min-max extensions to optimal control are known to have a direct connection to robust and H^∞ nonlinear control theory, as well as to risk-sensitive optimal control [1], [2], [3]. The origin of game-theoretic control dates back to the work of Isaacs (1965) [4] on differential games for two strictly competitive players, which provided a framework for the treatment of such problems. Isaacs associated the solution of a differential game with the solution to a HJB-like equation, namely its min-max extension, also known as the Isaacs (or Hamilton-Jacobi-Isaacs, HJI) equation. This equation was derived heuristically by Isaacs under the assumptions of Lipschitz continuity of the dynamics and the cost, as well as the assumption that both of them are *separable* in terms of the minimizing and maximizing controls. A treatment of the stochastic extension to differential games was first provided in [5]. Despite the plethora of theoretic work in the area of differential games, the algorithmic part has received significantly less attention, due to the inherent difficulty of solving such problems. A few approaches have been suggested in the past, such as the Markov Chain approximation method [6], but these have found limited applicability due to the "curse of dimensionality." Only very recently, a specific class of minimax control trajectory optimization methods have been derived, all based on the foundations of *differential dynamic programming* (DDP) [7], [8], [9].

[1]Ph.D. candidate, School of Aerospace Engineering, Georgia Institute of Technology, Atlanta. GA 30332-0150. USA. Email: exarchos@gatech.edu

[2]Assistant Professor, Institute for Robotics and Intelligent Machines, School of Aerospace Engineering, Georgia Institute of Technology, Atlanta. GA 30332-0150. USA. Email: evangelos.theodorou@ae.gatech.edu

[3]Professor, Institute for Robotics and Intelligent Machines, School of Aerospace Engineering, Georgia Institute of Technology, Atlanta. GA 30332-0150. USA. Email: tsiotras@gatech.edu

There is an innate connection between min-max extensions of optimal control and risk-sensitive stochastic control formulations. This relationship was first investigated by Jacobson in [1]. References [10] and [11] investigate risk-sensitive stochastic control in an LQG setting and for nonlinear stochastic systems and infinite horizon control tasks respectively. Ever since the fundamental work of [1], [10], [11], the topic of risk sensitivity has been studied extensively. In a risk-sensitive setting, the control objective is to minimize a performance index, which is expressed as a function of the mean and variance of a given state- and control-dependent cost. Therefore, the element of risk sensitivity arises from the minimization of the variance of that cost. An application of the Dynamic Programming principle on the risk-sensitive optimization criterion results in a backward PDE that is similar to the HJI PDE. Thus, risk-sensitive optimal control problems are directly related to stochastic differential games [3].

In this paper we present an algorithm designed to solve stochastic differential games by using the nonlinear Feynman-Kac lemma. This algorithm is a sampling-based scheme which relies on the theory of forward and backward stochastic differential equations (FBSDEs) and their connection to backward PDEs [12], [13]. In particular, we first obtain a probabilistic representation of the solution to the HJI PDE, expressed in the form of a system of FBSDEs. This system of FBSDEs is then simulated by employing linear regression techniques. Since the HJI PDE appears in both stochastic differential games and risk-sensitive optimal control problems, the proposed scheme is applicable to both types of stochastic optimal control formulations.

The paper is organized as follows: In Section II we introduce the problem statement and present the associated HJI equation for the class of problems considered in this work. Section III provides the stochastic representation of the solution to the HJI equation using the nonlinear Feynman-Kac lemma through FBSDEs. Risk-sensitive control and its connection to game-theoretic control is treated in Section IV. Section V deals with the numerical approximation used in this paper to solve FBSDEs, whereas in Section VI we provide application examples of the proposed algorithm. Finally, conclusions are presented in the Section VII.

II. 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. Consider the game-theoretic setting in which the expected game payoff is defined by the functional

$$P(\tau, x_\tau; u(\cdot), v(\cdot)) = \mathbb{E}\Big[g(x_T) + \int_\tau^T \big[q(t, x_t) + \tfrac{1}{2}u_t^\top R u_t - \tfrac{1}{2}v_t^\top Q v_t\big]\mathrm{d}t\Big], \quad (1)$$

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

$$\begin{cases} \mathrm{d}x_t = f(t, x_t)\mathrm{d}t + G(t, x_t)u_t\mathrm{d}t + L(t, x_t)v_t\mathrm{d}t \\ \qquad\quad + \Sigma(t, x_t)\mathrm{d}W_t, \quad t \in [\tau, T], \quad x(\tau) = x_\tau, \end{cases} \quad (2)$$

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 minimizing control vector, and $v \in \mathbb{R}^\mu$ is the maximizing control vector. Furthermore, R and Q are respectively $\nu \times \nu$ and $\mu \times \mu$ positive definite matrices, $g : \mathbb{R}^n \to \mathbb{R}$, $q : [\tau, T] \times \mathbb{R}^n \to \mathbb{R}$, $f : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$, $G : [0, T] \times \mathbb{R}^n \to \mathbb{R}^{n\times\nu}$, $L : [0, T] \times \mathbb{R}^n \to \mathbb{R}^{n\times\mu}$ 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 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 payoff (1). These impose, for example, that the functions g, q, f, G, L 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 processes $u : [0, T]\times\Omega \to \mathcal{U} \subseteq \mathbb{R}^\nu$ and $v : [0, T] \times \Omega \to \mathcal{V} \subseteq \mathbb{R}^\mu$ are $\{\mathscr{F}_t\}_{t\geq 0}$-adapted, which essentially translates into the control inputs being non-anticipating, i.e., relying only on past and present information.

The intuitive idea behind the game-theoretic setting is the existence of two players of conflicting interests. The first player controls u and wishes to minimize the payoff P over all choices of v, while the second player wishes to maximize P over all choices of u of his opponent. At any given time, the current state, as well as each opponents' current control action is known to both players. Furthermore, instantaneous switches in both controls are permitted, rendering the problem difficult to solve in general.

A. The Value Function and the HJI Equation

For any given initial condition (τ, x_τ), we investigate the game of conflicting control actions u, v that minimize (1) under all admissible non-anticipating strategies assigned to $u(\cdot)$, while maximizing it over all admissible non-anticipating strategies assigned to $v(\cdot)$. For the structure imposed on this problem by the form of the cost and dynamics at hand, the Isaacs condition[1] [4], [15], [16] holds, and the payoff is a saddlepoint solution to the following terminal value problem of a second order partial differential equation, known as the Hamilton-Jacobi-Isaacs (HJI) equation, which herein takes the form

$$\begin{cases} V_t + \inf\limits_{u\in\mathcal{U}} \sup\limits_{v\in\mathcal{V}} \Big\{ \tfrac{1}{2}\mathrm{tr}(V_{xx}\Sigma\Sigma^\top) + V_x^\top(f + Gu + Lv) + q \\ \quad + \tfrac{1}{2}u^\top R u - \tfrac{1}{2}v^\top Q v \Big\} = 0, \qquad (t, x) \in [0, T) \times \mathbb{R}^n, \\ V(T, x) = g(x), \quad x \in \mathbb{R}^n. \end{cases} \quad (3)$$

In the above, function arguments have been suppressed for notational compactness, and V_x and V_{xx} denote the gradient and the Hessian of V, respectively. The term inside the brackets is the Hamiltonian. For the chosen form of the cost integrand, and assuming that the optimal controls lie in the interiors of $\mathcal{U}$ and $\mathcal{V}$, we may carry out the infimum and supremum operations in (3) explicitly by taking the gradient of the Hamiltonian with respect to u and v and setting it equal to zero to obtain

$$Ru + G^\top(t, x)V_x(t, x) = 0, \quad (4)$$

$$-Qv + L^\top(t, x)V_x(t, x) = 0. \quad (5)$$

Therefore, for all $(t, x) \in [0, T] \times \mathbb{R}^n$, the optimal controls are given by

$$u^*(t, x) = -R^{-1}G^\top(t, x)V_x(t, x), \quad (6)$$

$$v^*(t, x) = Q^{-1}L^\top(t, x)V_x(t, x). \quad (7)$$

Inserting the above expression back into the original HJI equation and suppressing function arguments for notational brevity, we obtain the equivalent characterization

$$\begin{cases} V_t + \tfrac{1}{2}\mathrm{tr}(V_{xx}\Sigma\Sigma^\top) + V_x^\top f + q - \tfrac{1}{2}V_x^\top\Big(GR^{-1}G^\top \\ \quad - LQ^{-1}L^\top\Big)V_x = 0, \quad (t, x) \in [0, T) \times \mathbb{R}^n, \\ V(T, x) = g(x), \quad x \in \mathbb{R}^n. \end{cases} \quad (8)$$

III. A Feynman-Kac Representation through FBSDEs

There is a close relationship between stochastic differential equations and second-order partial differential equations (PDEs) 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 [17] 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 $X(\cdot)$[2], which, for any given initial

[1] The Isaacs condition renders the viscosity solutions of the upper and lower value functions equal (see [14]), thus making the order of minimization/maximization inconsequential.

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

condition $(t,x) \in [0,T] \times \mathbb{R}^n$, satisfies the Itô FSDE

$$\begin{cases} \mathrm{d}X_s = b(s,X_s)\mathrm{d}s + \Sigma(s,X_s)\mathrm{d}W_s, & s \in [t,T], \\ X_t = x. \end{cases} \tag{9}$$

The forward process (9) is also called the *state process* in the literature. We shall denote the solution to the forward SDE (9) 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

$$\begin{cases} \mathrm{d}Y_s = -h(s,X_s,Y_s,Z_s)\mathrm{d}s + Z_s^\top \mathrm{d}W_s & s \in [t,T], \\ Y_T = g(X_T). \end{cases} \tag{10}$$

The function $h(\cdot)$ is called the *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 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 [12] 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 forward SDE 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 $d(t,x)$[3] such that the solution $(Y^{t,x}, Z^{t,x})$ of the BSDE (10) is

$$Y_s^{t,x} = V(s, X_s^{t,x}), \quad Z_s^{t,x} = \Sigma^\top(s, X_s^{t,x}) d(s, X_s^{t,x}), \tag{11}$$

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 by an application of Itô's formula (see [13], [18], [12]):

[3]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.

Theorem 2: (Nonlinear Feynman-Kac) – Consider the Cauchy problem

$$\begin{cases} V_t + \frac{1}{2}\mathrm{tr}(V_{xx}\Sigma\Sigma^\top) + V_x^\top b(t,x) + h(t,x,V,\Sigma^\top V_x) = 0, \\ (t,x) \in [0,T) \times \mathbb{R}^n, \quad V(T,x) = g(x), \quad x \in \mathbb{R}^n, \end{cases} \tag{12}$$

wherein the functions Σ, b, h and g satisfy mild regularity conditions[4]. Then (12) admits a unique (viscosity) solution $V : [0,T] \times \mathbb{R}^n \to \mathbb{R}$, which has the following probabilistic representation:

$$V(t,x) = Y_t^{t,x}, \qquad \forall (t,x) \in [0,T] \times \mathbb{R}^n, \tag{13}$$

wherein $(X(\cdot), Y(\cdot), Z(\cdot))$ is the unique adapted solution of the FBSDE system (9)-(10). Furthermore,

$$(Y_s^{t,x}, Z_s^{t,x}) = \Big(V(s, X_s^{t,x}),\ \Sigma^\top(s, X_s^{t,x}) V_x(s, X_s^{t,x}) \Big), \tag{14}$$

for all $s \in [t,T]$, and if (12) admits a classical solution, then (13) provides that classical solution.

A careful comparison between equations (8) and (12) indicates that the nonlinear Feynman-Kac representation can be applied to the HJI equation given by (8) under a certain decomposability condition, stated in the following assumption:

Assumption 1: There exist matrix-valued functions $\Gamma : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{p \times \nu}$ and $B : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{p \times \mu}$ such that $G(t,x) = \Sigma(t,x)\Gamma(t,x)$ and $L(t,x) = \Sigma(t,x)B(t,x)$ 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 and L 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, the HJI equation given by (8) becomes

$$\begin{cases} V_t + \frac{1}{2}\mathrm{tr}(V_{xx}\Sigma\Sigma^\top) + V_x^\top f + q - \frac{1}{2} V_x^\top \Sigma \Big(\Gamma R^{-1} \Gamma^\top \\ - B Q^{-1} B^\top \Big) \Sigma^\top V_x = 0, \quad (t,x) \in [0,T) \times \mathbb{R}^n, \\ V(T,x) = g(x), \quad x \in \mathbb{R}^n, \end{cases} \tag{15}$$

in which function arguments have been suppressed, and which satisfies the format of (12) with

$$b(t,x) \equiv f(t,x), \tag{16}$$

and

$$\begin{aligned} h(t,x,z) \equiv q(t,x) - \tfrac{1}{2} z^\top \big(\Gamma(t,x) R^{-1} \Gamma^\top(t,x) \\ - B(t,x) Q^{-1} B^\top(t,x) \big) z. \end{aligned} \tag{17}$$

We may thus obtain the (viscosity) solution of (15) by simulating the system of FBSDE given by (9) and (10).

[4]In fact, [13] requires the functions Σ, b, h and g to be continuous, Σ and b to be uniformly Lipschitz in x, and h to be uniformly – w.r.t (t,x)– Lipschitz in (y,z). However, the nonlinear Feynman-Kac lemma has been recently extended to cases in which the driver is only continuous, and satisfies quadratic growth in z– see References [**?**], [**?**], [**?**], [**?**]. Concerning existence of solutions to the HJI equation in this case, see [**?**].

Notice that (9) corresponds to the uncontrolled ($u = 0$, $v = 0$) system dynamics.

IV. Connection to Risk-Sensitive Control

The connection between dynamic games and risk-sensitive stochastic control is well-documented in the literature [1], [2], [3]. Specifically, the optimal controller of a stochastic control problem with exponentiated integral cost (a so-called risk-sensitive problem) turns out to be identical to the minimizing player's unique minimax controller in a stochastic differential game setting. Indeed, consider the problem of minimizing the expected cost given by

$$J(\tau, x_\tau; u(\cdot)) = \epsilon \ln \mathbb{E}\Bigg\{ \exp \frac{1}{\epsilon}\Bigg[g(x_T) + \int_\tau^T q(t, x_t) + \tfrac{1}{2} u_t^\top R u_t \, \mathrm{d}t \Bigg]\Bigg\}, \quad (18)$$

where ϵ is a small positive number. The state dynamics are described by the Itô SDE

$$\begin{cases} \mathrm{d}x_t = f(t, x_t)\mathrm{d}t + G(t, x_t)u_t \mathrm{d}t + \sqrt{\dfrac{\epsilon}{2\gamma^2}}\tilde{\Sigma}(t, x_t)\mathrm{d}W_t, \\ t \in [\tau, T], \qquad x(\tau) = x_\tau. \end{cases} \quad (19)$$

Suppressing function arguments for notational compactness, the associated Hamilton-Jacobi-Bellman PDE for this problem is [3]

$$\begin{cases} V_t + \inf\limits_{u \in \mathcal{U}} \Bigg\{ \dfrac{\epsilon}{4\gamma^2} \mathrm{tr}(V_{xx}\tilde{\Sigma}\tilde{\Sigma}^\top) + V_x^\top (f + Gu) + q \\ \quad + \tfrac{1}{2} u^\top R u + \dfrac{1}{4\gamma^2} V_x^\top \tilde{\Sigma}\tilde{\Sigma}^\top V_x \Bigg\} = 0, \quad (t, x) \in [0, T) \times \mathbb{R}^n, \\ V(T, x) = g(x), \quad x \in \mathbb{R}^n. \end{cases} \quad (20)$$

The infimum operation can be performed explicitly, and yields the optimal control $u^*(t, x) = -R^{-1}G^\top(t, x)V_x(t, x)$. Setting $\Sigma = \sqrt{\epsilon/2\gamma^2}\tilde{\Sigma}$ and substituting the optimal control in the PDE (20) we readily obtain the equivalent characterization

$$\begin{cases} V_t + \tfrac{1}{2}\mathrm{tr}(V_{xx}\Sigma\Sigma^\top) + V_x^\top f + q - \tfrac{1}{2} V_x^\top \Bigg(GR^{-1}G^\top \\ \quad - \dfrac{1}{\epsilon}\Sigma\Sigma^\top \Bigg) V_x = 0, \qquad (t, x) \in [0, T) \times \mathbb{R}^n, \\ V(T, x) = g(x), \quad x \in \mathbb{R}^n. \end{cases} \quad (21)$$

The above equation is merely a special case of equation (8) obtained for the game-theoretic version, if one substitutes $Q = (1/\epsilon)I$ and $L = \Sigma$. Notice that this special case of L automatically satisfies Assumption 1 with B being the identity matrix. Thus, imposing the same decomposability condition on G, the solution to the risk-sensitive stochastic optimal control problem can be obtained by simulating the system of FBSDEs given by (9) and (10) using the definitions (16) and (17).

V. 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], [20], [21], in this paper we employ a modification proposed in previous work by the authors [22], 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[5]. 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 [23]:

$$\begin{cases} X_{i+1} \approx X_i + b(t_i, X_i)\Delta t_i + \Sigma(t_i, X_i)\Delta W_i, \\ i = 1, \ldots, N, \qquad X_0 = x. \end{cases} \quad (22)$$

Several alternative, higher order schemes exist that can be selected in lieu of the Euler scheme [23]. 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 (10) by

$$Y_i = \mathbb{E}[Y_i|\mathscr{F}_{t_i}] \approx \mathbb{E}[Y_{i+1} + h(t_{i+1}, X_{i+1}, Y_{i+1}, Z_{i+1})\Delta t_i | X_i]. \quad (23)$$

Notice that in the last equality the term $Z_i^\top \Delta W_i$ in (10) vanishes because of the conditional expectation (ΔW_i is 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 (14), the Z-process in (10) corresponds to the term $\Sigma^\top(s, X_s^{t,x})v_x(s, X_s^{t,x})$. Therefore we can write

$$\begin{aligned} Z_i &= \mathbb{E}[Z_i|\mathscr{F}_{t_i}] = \mathbb{E}[\Sigma^\top(t_i, X_i)\nabla_x v(t_i, X_i)|X_i] \\ &= \Sigma^\top(t_i, X_i)\nabla_x v(t_i, X_i), \end{aligned} \quad (24)$$

which naturally requires knowledge of the solution at time t_i on a neighborhood x, $v(t_i, x)$. The backpropagation is initialized at

$$Y_T = g(X_T), \qquad Z_T = \Sigma(T, X_T)^\top \nabla_x g(X_T), \quad (25)$$

for a $g(\cdot)$ which is differentiable almost everywhere. There are several ways to approximate the conditional expectation in (23), however in this work we shall employ the Least Squares Monte Carlo (LSMC) method[6], 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

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

[6]Treating conditional expectations by means of linear regression was made popular in the field of mathematical finance by [24].

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

$$\phi^* = \arg\min_{\phi} \mathbb{E}[|\phi(X) - Y|^2], \tag{26}$$

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 $\phi(\cdot) \approx \sum_{i=1}^{k} \varphi_i(\cdot)\alpha_i = \varphi(\cdot)\alpha$, with $\varphi(\cdot)$ being a row vector of k predetermined basis functions and α a column vector of constants, 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 [25], thus obtaining

$$\alpha^* = \arg\min_{\alpha\in\mathbb{R}^k} \frac{1}{M} \sum_{j=1}^{M} |\varphi(X^j)\alpha - Y^j|^2, \tag{27}$$

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

$$\Phi(X) = \begin{bmatrix} \varphi(X^1) \\ \vdots \\ \varphi(X^M) \end{bmatrix} \in \mathbb{R}^{M\times k}, \tag{28}$$

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

$$a^* = \Big(\Phi^\top(X)\Phi(X)\Big)^{-1} \Phi^\top(X) \Bigg(\begin{bmatrix} Y^1 \\ \vdots \\ Y^M \end{bmatrix} \Bigg), \tag{29}$$

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

Returning to our problem, we may apply the LSMC method to approximate the conditional expectation in equation (23) 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 $\{X_i^m\}_{i=1,\ldots,N}$, in which the index $m = 1, \ldots, M$ 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)[7] 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

$$Z_i^m \approx \Sigma(t_i, X_i^m)^\top \nabla_x \varphi(X_i^m)\alpha_i. \tag{30}$$

Concluding one iteration, this process is repeated for $t_{i-1}, \ldots, t_1$. 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

$$\begin{cases} \text{Initialize} : Y_T = g(X_T), \quad Z_T = \Sigma(T, X_T)^\top \nabla_x g(X_T), \\ \alpha_i = \arg\min\limits_{\alpha} \dfrac{1}{M} \Big\| \Phi(X_i)\alpha - \Big(Y_{i+1} \\ \qquad\qquad + \Delta t_i h(t_{i+1}, X_{i+1}, Y_{i+1}, Z_{i+1}) \Big) \Big\|^2, \\ Y_i = \Phi(X_i)\alpha_i, \quad Z_i^m = \Sigma(t_i, X_i^m)^\top \nabla_x \varphi(X_i^m)\alpha_i, \end{cases} \tag{31}$$

where $m = 1, \ldots, M$ and the matrix Φ defined in (28). Again, the minimizer in (31) can be obtained by directly solving the normal equation, i.e.,

$$a_i = \Big(\Phi^\top(X_i)\Phi(X_i)\Big)^{-1} \Phi^\top(X_i) \Big(Y_{i+1} \\ + \Delta t_i h(t_{i+1}, X_{i+1}, Y_{i+1}, Z_{i+1}) \Big), \tag{32}$$

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.

VI. Simulation Results

To evaluate the algorithm's performance, two simulations were performed on scalar systems, for which, owing to their simplicity, we have the opportunity to evaluate the system behavior.

A. A Linear System Example

The first example used is a scalar linear system for which the analytic solution can be recovered. Specifically, for a very high maximizer control weight Q, we expect the solution to be almost identical to the LQR solution, which is available in closed form [26]. We simulate the algorithm for $\mathrm{d}x = (0.2x + u + 0.5xv)\mathrm{d}t + 0.5\mathrm{d}w$, with $q(t, x) = 0$, $R = 2$, $x(0) = 1$, $T = 1$ and $g(x_T) = 40x_T^2$, thus penalizing deviation from the origin at the time of termination, T. For Q, the maximizing control cost factor, we selected varying values ranging from 5 to 50,000. In the latter case, we expect to recover the LQR coefficients. For the purposes of

[7] 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.

comparison with the closed form solution, the set of basis functions for Y was selected to be $[1\ x\ x^2]^\top$. For the LQR controller, the coefficients correspond to the basis functions $[1\ x^2]^\top$. Two thousand trajectories were generated on a time grid of $\Delta t = 0.004$. Fig. 1 shows that, indeed, for very hight values of Q the algorithm recovers the correct theoretic LQR coefficients, while Fig. 2 depicts simulations for the case in which the maximizing control is allowed to act on the system when it is relatively cheap. In this case, we can see that because the maximizer has enough control authority, the equilibrium has moved away from the desired value of $x(T) = 0$, as expected.

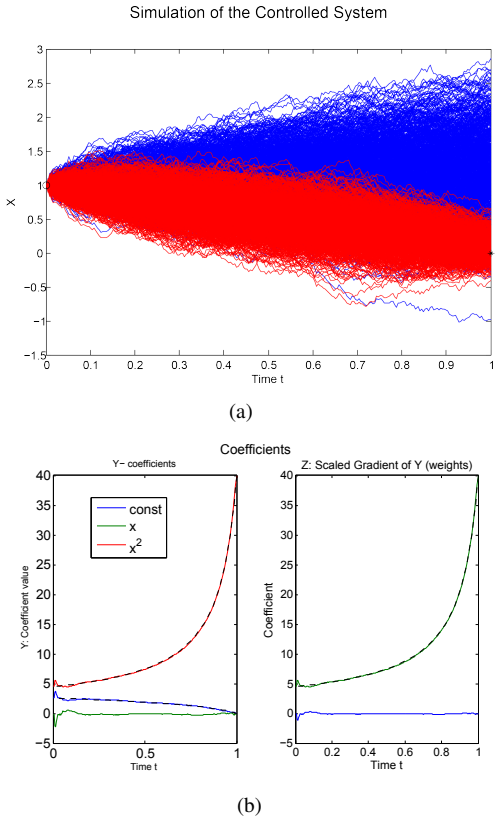

Fig. 1. Simulation of the system with very high maximizing control cost weight $Q = 50,000$. (a) Controlled trajectories (red) vs. uncontrolled (blue), (b) Y and Z coefficients, compared to those obtained by the closed form solution of the LQR if the maximizing control was not present (black dashed lines). We observe that for a high maximizing control cost, the obtained coefficients match those of the LQR as expected.

B. A Nonlinear System Example

To demonstrate that the scheme can accommodate nonlinearity in the dynamics, we also applied the algorithm to the same problem as in Section VI-A, by replacing the dynamics with $\mathrm{d}x = (4\cos x + u + 0.5xv)\mathrm{d}t + 0.5\mathrm{d}w$. The drift was replaced by a nonlinear term to introduce an additional behavior to the open-loop system trajectories. The results are depicted in Fig. 3. From the shape of the value function in Fig. 3(b) it is seen that the value is relatively flat at the beginning since there is no state-dependent running cost and

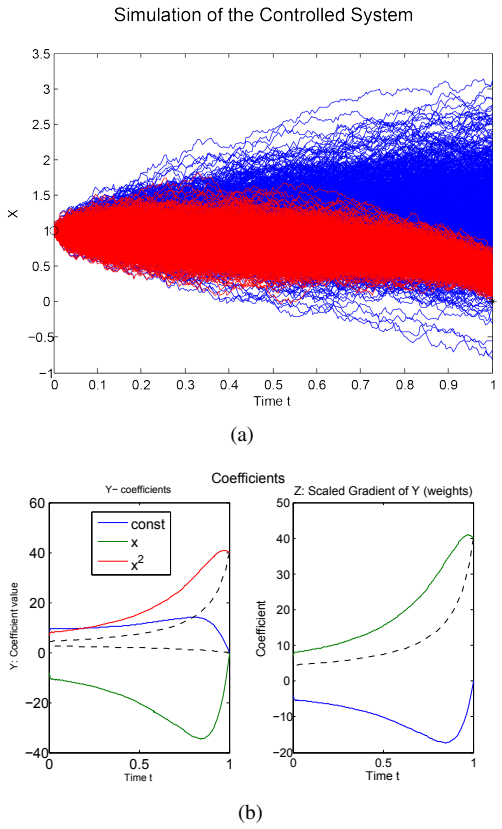

Fig. 2. Simulation of the system with small maximizing control cost weight $Q = 5$. (a) Controlled trajectories (red) vs. uncontrolled (blue), (b) Y and Z coefficients, compared to those obtained by the closed form solution of the LQR if the maximizing control was not present (black dashed lines).

becomes progressively quadratic at the final time owing to the boundary condition $V(T, x_T) = 40x_T^2$. Note, however, that Fig. 3(b) shows the value function over a rectangular grid. In fact, we have an accurate estimate of the value function only over the area of the state space visited by the sampled (open-loop) trajectories. In that sense, the areas not visited by the system are extrapolated based on the basis functions chosen to represent V.

VII. Conclusions

In this paper we presented a new algorithm for stochastic differential games and risk-sensitive stochastic optimal control problems. By utilizing a nonlinear version of the Feynman-Kac lemma, we obtained a probabilistic representation of the associated PDEs, by means of a system of FBSDEs. This system is then simulated using linear regression. We have demonstrated the applicability of the proposed algorithm by applying it on two scalar systems, including a linear system for which a closed-form solution is known. Future work will focus on alterative methods to perform regression, convergence and error properties of the scheme, as well as on the application of the proposed technique to more realistic systems.

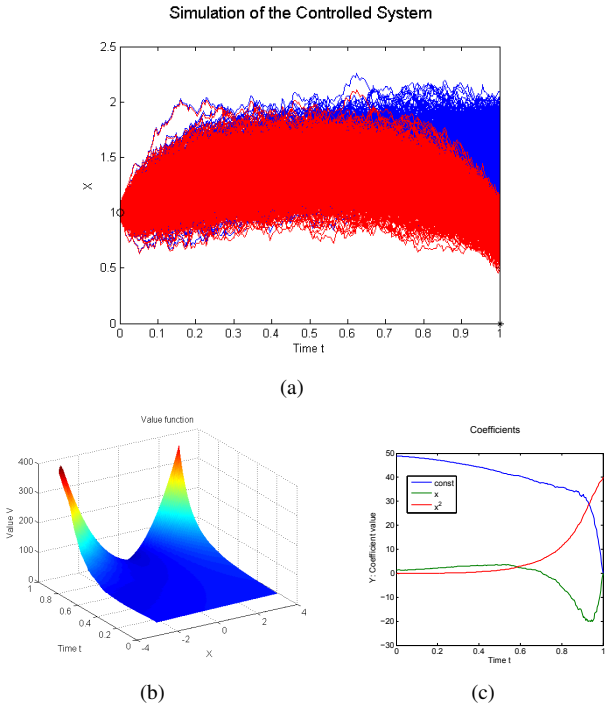

Fig. 3. Simulation of the nonlinear system with small maximizing control cost weight $Q = 5$. (a) Controlled trajectories (red) vs. uncontrolled (blue), (b) The Value function, (c) Y coefficients.

REFERENCES

[1] D. H. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Transactions on Automatic Control*, vol. 18, pp. 124–131, 1973.

[2] P. Dai Pra, L. Meneghini, and W. J. Runggaldier, "Connections between stochastic control and dynamic games," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 9, no. 4, pp. 303–326, 1996.

[3] T. Başar and P. Bernhard, H^∞- *Optimal Control and Related Minimax Design Problems*. Birkhäuser Boston, 2nd ed., 2008.

[4] R. Isaacs, *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. New York: Willey, 1965.

[5] H. Kushner and S. Chamberlain, "On stochastic differential games: Sufficient conditions that a given strategy be a saddle point, and numerical procedures for the solution of the game," *Journal of Mathematical Analysis and Applications*, vol. 26, pp. 560–575, 1969.

[6] H. Kushner, "Numerical approximations for stochastic differential games," *SIAM J. Control Optim.*, vol. 41, pp. 457–486, 2002.

[7] J. Morimoto, G. Zeglin, and C. Atkeson, "Minimax differential dynamic programming: Application to a biped walking robot," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, pp. 1927–1932, October 2003.

[8] J. Morimoto and C. Atkeson, "Minimax differential dynamic programming: An application to robust biped walking," *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[9] W. Sun, E. A. Theodorou, and P. Tsiotras, "Game-theoretic continuous time differential dynamic programming," (Chicago, IL), pp. 5593–5598, July 1–3, 2015.

[10] P. Whittle, "Risk-sensitive linear/quadratic/gaussian control," *Advances in Applied Probability*, vol. 13, no. 4, pp. 764–777, 1981.

[11] W. H. Fleming and W. M. McEneaney, "Risk-sensitive control on an infinite time horizon," *SIAM J. Control Optim.*, vol. 33, pp. 1881–1915, Nov. 1995.

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

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

[14] W. Fleming and P. Souganidis, "On the existence of value functions of two player zero-sum stochastic differential games," *Indiana University Mathematics Journal*, 1989.

[15] K. M. Ramachandran and C. P. Tsokos, *Stochastic Differential Games*. Atlantis Press, 2012.

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

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

[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] I. Exarchos and E. A. Theodorou, "Learning optimal control via forward and backward stochastic differential equations," *American Control Conference, Boston, MA, USA*, July 6-8, 2016.

[23] 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.

[24] 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.

[25] 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.

[26] R. F. Stengel, *Optimal Control and Estimation*. Dover Publications, Inc., 1994.