Value Function Estimators for Feynman-Kac Forward-Backward SDEs in Stochastic Optimal Control ⋆

Kelsey P. Hawkins [a], Ali Pakniyat [b], Panagiotis Tsiotras [a]

[a] *Georgia Institute of Technology, Atlanta, Georgia*

[b] *University of Alabama, Tuscaloosa, Alabama*

Abstract

Two novel numerical estimators are proposed for solving forward-backward stochastic differential equations (FBSDEs) appearing in the Feynman-Kac representation of the value function in stochastic optimal control problems. In contrast to the current numerical approaches, which are based on the discretization of the continuous-time FBSDE, we propose a converse approach, namely, we obtain a discrete-time approximation of the value function, and then we derive a discrete-time estimator that resembles the continuous-time counterpart. The proposed approach allows for the construction of higher accuracy estimators along with an error analysis. The approach is applied to the policy improvement step in a reinforcement learning framework. Numerical results, along with the corresponding error analysis, demonstrate that the proposed estimators show significant improvement in terms of accuracy over classical Euler-Maruyama-based estimators. In the case of LQ problems, we demonstrate that our estimators result in near machine-precision level accuracy, in contrast to previously proposed methods that can potentially diverge on the same problems.

Key words: Stochastic optimal control problems; Generalized solutions of Hamilton-Jacobi equations; Non-Linear Control Systems; Monte Carlo methods; Stochastic control and game theory; Parametric optimization.

1 Introduction

Feynman-Kac representation theory and its associated forward-backward stochastic differential equations (FBSDEs) have been gaining traction as a framework to solve nonlinear stochastic optimal control problems, including problems with quadratic cost [5], minimum-fuel (L_1-running cost) problems [5], differential games [6], as well as reachability problems [20]. Although FBSDE-based methods have seen growing attention in both the controls and robotics communities recently, much of the relevant research originated in the mathematical finance community [16,18].

The underlying foundation of Feynman-Kac-based FBSDE algorithms is the intrinsic relationship between the solution of a broad class of second-order parabolic or elliptic PDEs to the solution of FBSDEs (see, e.g., [24, Chapter 7]), brought to prominence in [19,4]. Both Hamilton-Jacobi-Bellman (HJB) and Hamilton-Jacobi-Isaacs (HJI) second order PDEs that are utilized for the solution of, respectively, stochastic optimal control and stochastic differential game problems, can thus be solved via FBSDE methods, even when the dynamics are nonlinear and the cost is non-quadratic. FBSDE methods thus provide an alternative to the grid-based solution of HJB/HJI-type PDEs, typically solved using finite-difference, finite-element, or level-set schemes, which are known for their poor scaling in high dimensional state spaces ($n \geq 4$).

Recently proposed methods [6,5] have suggested an *iterative*-FBSDE (iFBSDE) approach for solving stochastic optimal control problems, where alternating forward sampling passes and backward value function regression passes iteratively improve the approximation of the optimal value function. While initial results demonstrate promise in terms of flexibility and theoretical validity, iFBSDE methods have not yet matured. For even modest problems, iFBSDE methods can be unstable, producing value function approximations which quickly diverge. Thus, producing more robust numerical methods for solving FBSDEs is critical for the broader adoption of iFBSDE methods for real-world tasks.

⋆ This paper was not presented at any IFAC meeting. Corresponding author K. P. Hawkins. Tel. 919-280-5739. Email kphawkins@gatech.edu.

Email addresses: kphawkins@gmail.com (Kelsey P. Hawkins), apakniyat@ua.edu (Ali Pakniyat), tsiotras@gatech.edu (Panagiotis Tsiotras).

The iFBSDE numerical methods broadly consist of two steps: a forward pass, which generates Monte Carlo samples of the forward stochastic process, and a backward pass, which iteratively approximates the value function backwards in time. The value function approximation is performed using least-squares Monte Carlo (LSMC), which implicitly solves the backward SDE using parametric function approximation [16]. The approximate value function fit in the backward pass is then used to improve the sampling in an updated forward pass, leading to an iterative algorithm that improves the approximation till convergence.

Although at first glance iFBSDE methods seem similar to differential dynamic programming (DDP) techniques [14], the approach is significantly different. DDP methods require first and second order derivatives of the dynamics, and directly compute a quadratic approximation of the value function using constraints on the derivatives of the value function. By comparison, iFBSDE only uses approximations of the value function at a distribution of states, using the derivative of the value function to improve the accuracy of the estimator. The iFBSDE methods are more flexible, in the sense that they do not require derivatives of the dynamics and can be used with models of the value function that are not necessarily quadratic. Furthermore, for most DDP applications, a quadratic running cost with respect to the control is required for appropriate regularization whereas iFBSDE methods more easily accommodate non-quadratic running costs (e.g., of the class L_1 or zero-valued), lending to a variety of control applications [5].

In this work, we investigate the discrete-time approximation of the backward SDE in the context of solving for the value function in the backward pass in stochastic optimal control FBSDE methods. Although for some special cases analytic solutions of the backward SDEs over short intervals can be accommodated into the associated algorithms [16], for many nonlinear problems analytic solutions are not available and numerical integration based on time-discretization is necessary. In the currently available algorithms in the literature, Euler-Maruyama approximations are employed for discretizing the continuous-time FBSDEs [5], to solve for an approximation of the continuous-time value function.

Instead of the direct application of the Euler-Maruyama approximation on the Feynman-Kac FBSDEs, we formulate a discrete time problem with the Euler-Maruyama approximation of the dynamics, cost, and value function, and then we derive discrete-time relationships using Taylor expansions that resemble their continuous-time counterparts. By doing so, we arrive at a set of alternative estimators for the value function.

The primary contributions of this paper are as follows:

- We propose a pair of alternative estimators for the value function used in the backward pass of a Girsanov-drifted Feynman-Kac FBSDE numerical method.
- We characterize the theoretical bias and variance of these estimators and show their theoretic superiority to previously proposed estimators.
- We numerically confirm the theoretical results on representative stochastic optimal control problems.

This paper expands upon the authors' prior work in [12], first by providing more details into how the proposed estimators are constructed, and second, by providing detailed proofs for the stated theorems. In addition, we discuss how the methodology can be adapted to improve the policy in a reinforcement learning setting by computing a similar approximation of the Q-value function. Finally, we present new results of numerical experiments on a two-dimensional nonlinear problem and a four-dimensional LQ problem, verifying our theoretical claims about the accuracy of the proposed estimators.

2 Continuous-Time Feynman-Kac FBSDEs

In this section, we introduce the "on-policy" value function and show how its solution relates to the solution of a pair of continuous-time forward-backward stochastic differential equations (FBSDEs).

2.1 On-Policy Value Functions

Let $\mu(t,x)$ be a given bounded and measurable policy and let $f^\mu(t,x) := f(t,x,\mu(t,x))$ and $\ell^\mu(t,x) := \ell(t,x,\mu(t,x))$ refer to the dynamics and the running cost associated with some optimal control problem, respectively. The on-policy value function V^μ is defined as

$$V^\mu(t,x) = \mathbf{E}[\int_t^T \ell_s^\mu \,\mathrm{d}s + g(X_T)\,|X_t = x], \qquad (1)$$

with the process X_s satisfying the forward SDE (FSDE)

$$\mathrm{d}X_s = f_s^\mu \,\mathrm{d}s + \sigma_s \,\mathrm{d}W_s, \qquad (2)$$

with initial condition $X_0 = x_0$, where $f_s^\mu := f^\mu(s, X_s)$ and similarly for ℓ_s^μ and σ_s, and where W_s is an n-dimensional standard Brownian (Wiener) process. We assume that $f^\mu, \sigma, \ell^\mu, g$ are uniformly continuous in (t,x) and Lipschitz in x, and that σ^{-1} exists and is uniformly bounded on its domain. Furthermore, we assume that the PDE

$$\begin{aligned} \partial_t v + \frac{1}{2}\operatorname{tr}(\sigma\sigma^\top \partial_{xx} v) + f^{\mu\top}\partial_x v + \ell^\mu = 0, \\ g = v|_{t=T} \end{aligned} \qquad (3)$$

has a classical solution, that is, the solution is continuously differentiable in t, twice so in x, and satisfies equation (3) everywhere [1]. A Feynman-Kac-type theorem [24, Chapter 7, Theorem 4.1] establishes that V^μ in (1) is this classical solution to (3) and is the same for any

[1] The theory can be relaxed to the case where only viscosity solutions are available, at the cost of a more technical analysis. For more details, please see [10].

Brownian process W_s (i.e., the FSDE (2) has a unique strong solution).

2.2 Off-Policy Drifted FBSDE

If we sample from the trajectory distribution generated by the FSDE (2) with the *on-policy* drift term f_s^μ we can easily arrive at relationships which allow us to solve for V^μ either directly from (1) or via dynamic programming. Instead, we present a result that shows that we can sample from an FSDE with a different drift term K_s, and then solve a system of *drifted* FBSDEs to obtain the same value function V^μ.

Theorem 2.1 *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in[0,T]}, \mathsf{P})$ be a filtered probability space on which W_s^P is Brownian and let K_s be any $\mathcal{F}_s$-progressively measurable process on the interval $[0,T]$ such that $D_s := \sigma_s^{-1}(f_s^\mu - K_s)$ satisfies Novikov's criterion ($\mathbf{E}_\mathsf{P}[\exp(1/2\int_0^T \|D_s\|^2\,\mathrm{d}s)] < \infty$) [1, Theorem 15.4.2]*[2] *, and let*

$$\mathrm{d}X_s = K_s\,\mathrm{d}s + \sigma_s\,\mathrm{d}W_s^\mathsf{P}, \quad X_0 = x_0, \tag{4}$$

admit a unique square-integrable solution X_s (see, e.g., [24, Chapter 1, Theorem 6.16]). Then, the forward SDE (4) and the backward SDE

$$\mathrm{d}Y_s = -(\ell_s^\mu + Z_s^\top D_s)\,\mathrm{d}s + Z_s^\top \mathrm{d}W_s^\mathsf{P}, \quad Y_T = g(X_T), \tag{5}$$

have a unique, square-integrable solution (X_s, Y_s, Z_s) such that

$$\begin{aligned} Y_s &= V^\mu(s, X_s), && s \in [0,T], \\ Z_s &= \sigma_s^\top \partial_x V^\mu(s, X_s), && \text{a.e. } s \in [0,T], \end{aligned} \tag{6}$$

holds P-a.s. where V^μ is defined in (1). □

PROOF. The existence of a square-integrable solution to (4) allows the conditions of [24, Chapter 7, Theorem 3.2] to be satisfied for (5), guaranteeing a unique square-integrable solution (Y_s, Z_s). Defining the process

$$W_t^\mathsf{Q} := W_t^\mathsf{P} - \int_0^t D_s\,\mathrm{d}s, \quad t \in [0,T], \tag{7}$$

Girsanov's theorem guarantees that W_s^Q is Brownian in some measure Q [8, Chapter 5, Theorem 10.1]. With a simple algebraic reduction, Girsanov's theorem also guarantees that X_s solves the FSDE (2) (where $W_s = W_s^\mathsf{Q}$), and that (X_s, Y_s, Z_s) solves the BSDE $\mathrm{d}Y_s = -\ell_s^\mu\,\mathrm{d}s + Z_s^\top \mathrm{d}W_s^\mathsf{Q}$ with $Y_T = g(X_T)$. Moreover, Theorem 4.5 in [24, Chapter 7] establishes that (6) holds Q-a.s., since V^μ is the solution of (3). Novikov's condition on D_s yields that P and Q are equivalent measures [17], and thus we can conclude that (6) holds P-a.s. as well. ■

[2] The notation $\mathbf{E}_\mathsf{P}$ hereafter refers to the expectation taken in the measure P.

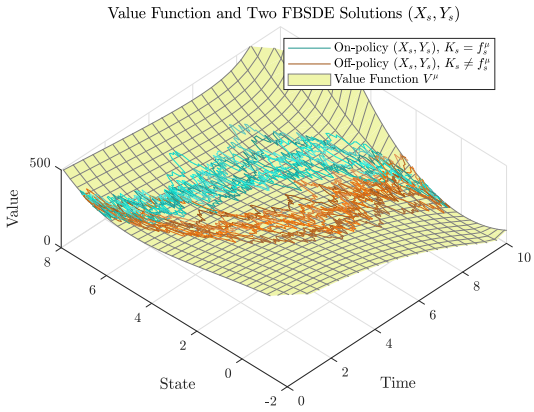


Fig. 1. Illustration of the result of (6) for two separate applications of Theorem 2.1 showing that the joint distribution (t, X_t, Y_t) lies on the surface $(t, x, V^\mu(t,x))$. This holds regardless of whether the drift term is on-policy ($K_s = f_s^\mu$) or off-policy ($K_s \neq f_s^\mu$).

When the samples of the FSDE are drawn using an arbitrary drift K_s instead of f_s^μ, the latter associated with the target policy μ, we say that the FBSDE samples "off-policy." Off-policy sampling is useful for numerical methods because one can arbitrarily sample in the forward pass, then solve for the value function V^μ associated with a target policy μ, where this policy can be established during the backward pass. Figure 1 illustrates Theorem 2.1. In the figure, V^μ is the optimal value function and the cyan trajectories depict the optimal trajectory distribution. When approximating the unknown optimal value function, we can begin with an approximate drift that generates the x component of the orange trajectory distribution [3]. As we solve the BSDE backwards along this distribution for the y component of the joint distribution (X_s, Y_s), we obtain new approximations for the optimal value function, and thus, new approximations for the optimal policy. At the end of the backward pass we have a direct estimate of the yellow surface around the distribution of the orange trajectories without ever having sampled from the optimal policy. A subsequent iteration samples forward utilizing a newly estimated policy.

Remark 2.1 *For any given process $\widehat{K}_s$ and some large constant $C > 0$, it is possible to construct a process K_s such that the corresponding process $D_s = \sigma_s^{-1}(f_s^\mu - K_s)$ is a bounded process and, thus, satisfying Novikov's condition and the assumption in Theorem 2.1. To be more specific, one can set*

$$K_s = \begin{cases} \widehat{K}_s, & \text{if} \quad \|f_s^\mu - \widehat{K}_s\| < C, \\ \widetilde{K}_s & \text{otherwise}, \end{cases} \tag{8}$$

with $\widetilde{K}_s$ an arbitrary process satisfying $\|f_s^\mu - \widetilde{K}_s\| < C$;

[3] Colors are best viewed in the electronic version.

e.g., $\widetilde{K}_s = f_s^\mu$ *or* $\widetilde{K}_s = -f_s^\mu + \left(\frac{C}{\|f_s^\mu - \widehat{K}_s\|} f_s^\mu - \widehat{K}_s \right)$, *etc.*

3 Forward-Backward Difference Equations

In [5] the results of the continuous-time FBSDE theory were reduced to a discrete-time approximation via the Euler-Maruyama method. In this section we propose the converse approach: we begin by forming a discrete-time approximation of the dynamics and the value function, then we derive relationships that resemble those arrived at by taking the Euler-Maruyama approximation of the FBSDE system (4)-(5). In doing so, we make two contributions: first, we arrive at better estimators compared to the direct discretization of the continuous time relations because we are able to exploit characteristics of the discrete-time formulation obscured by the continuous-time problem, and, secondly, we provide a discrete-time intuition for the continuous-time theory.

3.1 Discrete-Time On-Policy Value Function

The interval $[0, T]$ is partitioned into N subintervals of length Δt with the partition $\{t_0 = 0, t_1 = \Delta t, ..., t_{N-1} = T - \Delta t, t_N = T\}$. We abbreviate the variables $X_{t_i} =: X_i$ for brevity. Using the Euler-Maruyama method [15], let $F_i^\mu = f(t_i, X_i, \mu_i(X_i))\Delta t$, $\Sigma_i = \sigma(t_i, X_i)(\Delta t)^{1/2}$, and $L_i^\mu = \ell(t_i, X_i, \mu_i(X_i))\Delta t$, where $\mu_i(X_i) = \mu(t_i, X_i)$. The discrete-time on-policy value function is

$$V_i^\mu(x) = \mathbf{E}[\sum_{j=i}^{N-1} L_j^\mu + g(X_N) \,|X_i = x], \tag{9}$$

for $i = 0, \ldots, N$ where the discrete time process $\{X_j\}$ obeys the difference equation

$$X_{j+1} - X_j = F_j^\mu + \Sigma_j W_j, \tag{10}$$

with initial condition $X_i = x$, where $\{W_j\}_{j=i}^{N-1}$ is a standard discrete time Brownian increment process, that is, $W_j \sim \mathcal{N}(0, I_n)$ is normally distributed, is $\mathcal{F}_{j+1}$-measurable (for the given filtration $\{\mathcal{F}_j\}_{j\in\{i,\ldots,N\}}$), and $\{W_j\}$ are mutually independent.

3.2 Drifted Taylor-Expanded Backward Difference

We now offer a discrete-time approximation of the drifted off-policy FBSDEs.

3.2.1 FSDE Approximation

Overloading notation, let $(\Omega, \mathcal{F}, \{\mathcal{F}_i\}_{i\in\{0,\ldots,N\}}, \mathsf{P})$ be a discrete-time filtered probability space where W_i^P is the associated Brownian increment process. Define on this space the difference equation

$$X_{i+1} - X_i = K_i + \Sigma_i W_i^\mathsf{P}, \quad X_0 = x_0, \tag{11}$$

where the process $\{K_i\}_{i=0}^{N-1}$ is defined such that each K_i is $\mathcal{F}_i$-measurable and independent of W_i^P. For example, K_i can be constructed using the function $K_i(\omega) = \mathcal{K}_i(X_i(\omega), \xi_i(\omega))$, where $\{\xi_i\}$ is some random process where ξ_i is $\mathcal{F}_i$-measurable and independent of W_i^P (but not necessarily independent of W_{i-1}^P).

3.2.2 BSDE Approximation

We define the ideal discrete-time BSDE process as $\{Y_i := V_i^\mu(X_i)\}$ and the ideal backward difference as $\Delta Y_i := Y_{i+1} - Y_i$. For each backward step from $i+1$ to i we assume we have an approximation $\widetilde{V}_{i+1}^\mu \approx V_{i+1}^\mu$, twice differentiable, and we wish to produce an approximation $\widetilde{V}_i^\mu \approx V_i^\mu$ using least-squares Monte-Carlo (LSMC) function regression [16]. We use two separate estimators, $\widehat{Y}_{i+1} \approx Y_{i+1}$ and $\Delta\widehat{Y}_i \approx \Delta Y_i$, to obtain the combined estimator

$$\widehat{Y}_i := \widehat{Y}_{i+1} - \Delta\widehat{Y}_i, \tag{12}$$

with the interpretation that $\widehat{Y}_i$ estimates $\widetilde{V}_i^\mu(X_i) \approx V_i^\mu(X_i)$. Both $\widehat{Y}_{i+1}$ and $\Delta\widehat{Y}_i$ can be chosen according to different approximation schemes; these choices are investigated below.

3.2.3 Taylor-Based Backward Step Approximator

Similar to the definition (7) in the proof of Theorem 2.1, we define the process

$$W_i^\mathsf{Q} := W_i^\mathsf{P} - D_i, \quad i = 0, \ldots, N-1, \tag{13}$$

where $D_i := \Sigma_i^{-1}(F_i^\mu - K_i)$. A discrete-time version of Girsanov's theorem yields the existence of a measure Q under which the process $\{W_i^\mathsf{Q}\}$ is a Brownian increment process [3, Theorem 1]. By substituting this process into (11), note that $\{X_i\}$ always satisfies the difference equation in (10) where $\{W_i^\mathsf{Q}\}$ is the Brownian increment process. Since the choice of Brownian increment process is irrelevant to the definition of the on-policy value function, if we use the expectation $\mathbf{E}_\mathsf{Q}$ in (9), the solution to the off-policy drifted difference equation (11) can be used as the process in the definition of the on-policy value function. It is easy to show that the on-policy value function V_i^μ satisfies the Bellman equation [4]

$$V_i^\mu(X_i) = L_i^\mu + \mathbf{E}_\mathsf{Q}[V_{i+1}^\mu(X_{i+1})|X_i, K_i]. \tag{14}$$

The proposed backwards step estimator is a simplified form of

$$\Delta\widehat{Y}_i = \widetilde{Y}_{i+1} - (L_i^\mu + \mathbf{E}_\mathsf{Q}[\widetilde{Y}_{i+1}|X_i, K_i]), \tag{15}$$

[4] Although the rightmost term in the Bellman equation typically appears as $\mathbf{E}_\mathsf{Q}[V_{i+1}^\mu(X_{i+1})|X_i]$, we can substitute in $\mathbf{E}_\mathsf{Q}[V_{i+1}^\mu(X_{i+1})|X_i, K_i] = \mathbf{E}_\mathsf{Q}[V_{i+1}^\mu(X_{i+1})|X_i]$ because X_{i+1} is independent of K_i given X_i in the measure Q. Conditional independence can be demonstrated by noting that $\mathbf{E}_\mathsf{Q}[\mathbf{1}_{\{(X_{i+1}, K_i)\in A\times B\}}|X_i] = \mathbf{E}_\mathsf{Q}[\mathbf{1}_{\{X_i + F_i^\mu + \Sigma_i W_i^\mathsf{Q} \in A\}}|X_i]\mathbf{E}_\mathsf{Q}[\mathbf{1}_{\{K_i\in B\}}|X_i]$.

where $\widetilde{Y}_{i+1}$ is computed by a Taylor expansion to be introduced shortly. Specifically, using the second-order Taylor expansion of the function $\widetilde{V}^{\mu}_{i+1}(X_{i+1}) \approx V^{\mu}_{i+1}(X_{i+1}) = Y_{i+1}$ centered at the conditional mean of X_{i+1} taken in the measure P, yields $\overline{X}^{\mathsf{P}}_{i+1} := \mathbf{E}_{\mathsf{P}}[X_{i+1}|X_i, K_i] = X_i + K_i$. Furthermore, we have that

$$\widetilde{V}^{\mu}_{i+1}(X_{i+1}) = \widetilde{V}^{\mu}_{i+1}(\overline{X}^{\mathsf{P}}_{i+1} + \Sigma_i W^{\mathsf{P}}_i) = \widetilde{Y}_{i+1} + \delta^{\text{h.o.t.}}_{i+1}, \quad (16)$$

where,

$$\widetilde{Y}_{i+1} := \overline{Y}_{i+1} + \overline{Z}^{\top}_{i+1} W^{\mathsf{P}}_i + \frac{1}{2}(W^{\mathsf{P}}_i)^{\top} \overline{M}_{i+1} W^{\mathsf{P}}_i, \quad (17)$$

and $\overline{Y}_{i+1} := \widetilde{V}^{\mu}_{i+1}(\overline{X}^{\mathsf{P}}_{i+1})$, $\overline{Z}_{i+1} := \Sigma^{\top}_i \partial_x \widetilde{V}^{\mu}_{i+1}(\overline{X}^{\mathsf{P}}_{i+1})$, $\overline{M}_{i+1} := \Sigma^{\top}_i \partial_{xx} \widetilde{V}^{\mu}_{i+1}(\overline{X}^{\mathsf{P}}_{i+1})\Sigma_i$, and $\delta^{\text{h.o.t.}}_{i+1}$ includes the third and higher order terms in the Taylor series expansion. Substituting (13) into (17), then (17) into (15), and simplifying [5] yields the proposed estimator,

$$\begin{aligned} \Delta\widehat{Y}_i := &-L^{\mu}_i + \overline{Z}^{\top}_{i+1} W^{\mathsf{P}}_i - \overline{Z}^{\top}_{i+1} D_i \\ &+ \frac{1}{2} \operatorname{tr}\left(\overline{M}_{i+1}(W^{\mathsf{P}}_i (W^{\mathsf{P}}_i)^{\top} - I - D_i D_i^{\top})\right). \end{aligned} \quad (18)$$

Lemma 3.1 *The choice* (18) *yields the residual error*

$$\Delta Y_i - \Delta\widehat{Y}_i = \delta^{\Delta\widehat{Y}}_{i+1} - \mathbf{E}_{\mathsf{Q}}[\delta^{\Delta\widehat{Y}}_{i+1}|X_i, K_i], \quad (19)$$

where, $\delta^{\Delta\widehat{Y}}_{i+1} := V^{\mu}_{i+1}(X_{i+1}) - \widetilde{V}^{\mu}_{i+1}(X_{i+1}) + \delta^{\text{h.o.t.}}_{i+1}$ *is the sum of the error in approximation of* $V^{\mu}_{i+1}(X_{i+1})$ *and the residual due to the Taylor expansion.*

PROOF. The Taylor expansion (16) immediately gives $Y_{i+1} = \widetilde{Y}_{i+1} + \delta^{\Delta\widehat{Y}}_{i+1}$. Substituting into (15) yields $\Delta\widehat{Y}_i = -L^{\mu}_i + Y_{i+1} - \delta^{\Delta\widehat{Y}}_{i+1} - \mathbf{E}_{\mathsf{Q}}[Y_{i+1} - \delta^{\Delta\widehat{Y}}_{i+1}|X_i, K_i]$. If we substitute Y_i, Y_{i+1} into the Bellman equation (14) we have $Y_i = L^{\mu}_i + \mathbf{E}_{\mathsf{Q}}[Y_{i+1}|X_i, K_i]$. After substituting this expression into the previous equation and rearranging we arrive at (19). ∎

3.3 Estimators of $\widehat{Y}_{i+1}$

We propose two potential estimators for $\widehat{Y}_{i+1} \approx V^{\mu}_{i+1}(X_{i+1})$. First, we propose using the value function approximation associated with the previous backward step to re-estimate the $\widehat{Y}_{i+1}$ values,

$$\widehat{Y}^{\text{re-est}}_{i+1} := \widetilde{V}^{\mu}_{i+1}(X_{i+1}). \quad (20)$$

Alternatively, we can also use the estimator

$$\widehat{Y}^{\text{noiseless}}_{i+1} := \widetilde{Y}_{i+1}, \quad (21)$$

which ends up cancelling out the terms with W^{P}_i, so that (12) reduces to

$$\begin{aligned} \widehat{Y}^{\text{noiseless}}_i = &L^{\mu}_i + \overline{Y}_{i+1} + \overline{Z}^{\top}_{i+1} D_i \\ &+ \frac{1}{2} \operatorname{tr}\left(\overline{M}_{i+1}(I + D_i D_i^{\top})\right). \end{aligned} \quad (22)$$

3.3.1 Error Analysis

The following theorem establishes the error of the two estimators.

Theorem 3.2 *For the estimator* $\widehat{Y}_i := \widehat{Y}_{i+1} - \Delta\widehat{Y}_i$, *where* $\Delta\widehat{Y}_i$ *is defined in* (18) *and* $\widehat{Y}_{i+1}$ *is defined in* (20) *or* (21), *the bias is*

$$\begin{aligned} \mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}^{\text{re-est}}_i|X_i, K_i] = &\mathbf{E}_{\mathsf{Q}}[\delta^{\Delta\widehat{Y}}_{i+1}|X_i, K_i] \\ &- \mathbf{E}_{\mathsf{P}}[\delta^{\text{h.o.t.}}_{i+1}|X_i, K_i], \end{aligned} \quad (23)$$

$$\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}^{\text{noiseless}}_i|X_i, K_i] = \mathbf{E}_{\mathsf{Q}}[\delta^{\Delta\widehat{Y}}_{i+1}|X_i, K_i]. \quad (24)$$

Respectively, the variances of these estimators are

$$\operatorname{Var}_{\mathsf{P}}[\widehat{Y}^{\text{re-est}}_i|X_i, K_i] = \operatorname{Var}_{\mathsf{P}}[\delta^{\text{h.o.t.}}_{i+1}|X_i, K_i], \quad (25)$$

$$\operatorname{Var}_{\mathsf{P}}[\widehat{Y}^{\text{noiseless}}_i|X_i, K_i] = 0. \quad (26)$$

PROOF. See Appendix A. ∎

We call the estimation scheme used in [7] Euler-Maruyama-noiseless (EM-noiseless) because it is arrived at by applying EM to the continuous-time FBSDEs. The following proposition offers a comparative analysis.

Proposition 3.3 *The bias of the EM-noiseless estimator* $\widehat{Y}^{\text{em-nless}}_i := \widetilde{V}^{\mu}_{i+1}(X_{i+1}) + L^{\mu}_i + \widetilde{Z}^{\top}_{i+1} D_i$, *where* $\widetilde{Z}_{i+1} := \Sigma^{\top}_i \partial_x \widetilde{V}^{\mu}_{i+1}(X_{i+1})$, *has the following relationship with the Taylor re-estimate estimator bias,* $\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}^{\text{em-nless}}_i|X_i, K_i] = \mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}^{\text{re-est}}_i|X_i, K_i] + \frac{1}{2} D^{\top}_i \overline{M}_{i+1} D_i +$ h.o.t.. *Moreover, the variance of the EM-noiseless estimator is greater than the Taylor estimator,* $\operatorname{Var}_{\mathsf{P}}[\widehat{Y}^{\text{em-nless}}_i|X_i, K_i] \geq \operatorname{Var}_{\mathsf{P}}[\widehat{Y}^{\text{re-est}}_i|X_i, K_i] + \|\overline{Z}_{i+1} + \overline{M}_{i+1} D_i\|^2$.

PROOF. See Appendix B. ∎

The addition of the $\frac{1}{2} D^{\top}_i \overline{M}_{i+1} D_i$ term to the bias makes the EM estimator generally more biased than the Taylor estimator. This observation is made more precisely in the following proposition.

Proposition 3.4 *If the error in the approximation of* $V^{\mu}_{i+1}(X_{i+1})$ *and the third and higher order terms in the Taylor expansions of* $\widetilde{V}^{\mu}_{i+1}(X_{i+1})$ *and* $\partial_x \widetilde{V}^{\mu}_{i+1}(X_{i+1})$ *are all relatively small in magnitude compared to* $|\frac{1}{2} D^{\top}_i \overline{M}_{i+1} D_i|$, *the bias of the EM-noiseless estimator is greater than the bias of the Taylor estimator, that is,*

$$|\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}^{\text{em-nless}}_i|X_i, K_i]| \geq |\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}^{\text{re-est}}_i|X_i, K_i]|.$$

[5] Note that D_i, $\overline{Y}_{i+1}$, $\overline{Z}_{i+1}$, and $\overline{M}_{i+1}$, are (X_i, K_i)-measurable and thus come out of the conditional expectations $\mathbf{E}_{\mathsf{Q}}[\cdot|X_i, K_i]$.

PROOF. See Appendix C. ∎

It is worth remarking that neither of the two estimators are unbiased estimators but, as established in Proposition 3.3, the proposed Taylor estimator yields a smaller variance compared to the EM-noiseless estimator. Notice that $D_i := \Sigma_i^{-1}(F_i^\mu - K_i)$ is a consequence of the difference between the selection of K for forward sampling and the drift associated with the policy of interest μ. Therefore, if $D = 0$ (i.e., if K is always selected to be F^μ) the estimators have the same bias (while the proposed Taylor estimator always yields a smaller variance). However, in order to compare the two biases when $D \neq 0$, one needs to first fix other parameters of the underlying computational algorithm. In particular, the error in the approximation of $V_{i+1}^\mu(X_{i+1})$ and the third and higher order terms in the Taylor expansions of $\widetilde{V}_{i+1}^\mu(X_{i+1})$ and $\partial_x \widetilde{V}_{i+1}^\mu(X_{i+1})$ depend on several factors including the number of samples, the granularity of time discretization, and the selection of basis functions for the representation of $\widetilde{V}^\mu$. Notice also that selecting K_i different from F_i^μ can potentially improve numerical accuracy (see, e.g., [11]) and, hence, in the development of numerical algorithms $|\frac{1}{2} D_i^\top \overline{M}_{i+1} D_i|$ remains significant even at near convergence of the algorithm. In comparison, the error in the approximation of $V_{i+1}^\mu(X_{i+1})$ is expected to become small near convergence and, furthermore, with a proper selection of basis for the representation of $\widetilde{V}^\mu$ (see, e.g., Proposition 3.1 below) other errors can be suppressed in such a way that third and higher order derivatives are either zero or relatively small. Hence, the proposed Taylor estimator outperforms the EM estimator in both its bias and in its variance by Proposition 3.4. In particular, if we use a value function approximation representation that is always guaranteed to be quadratic, we have the following result.

Remark 3.1 *If the value function approximation $\widetilde{V}_{i+1}^\mu$ is quadratic, then $\delta_{i+1}^{\text{h.o.t.}} \equiv 0$.*

This is a consequence of the fact that if $\widetilde{V}_{i+1}^\mu$ is quadratic then its second order Taylor expansion is exact.

The magnitude of the error term $\delta_{i+1}^{\Delta\widehat{Y}}$ depends on the measure we use to interpret it. For numerical applications we sample from the measure P instead of Q, and thus $\mathbf{E}_\mathsf{Q}[\delta_{i+1}^{\Delta\widehat{Y}} | X_i, K_i]$ is difficult to interpret. We can use the following result to characterize the value exclusively in the measure P.

Proposition 3.5 *The bias term appearing in Theorem 3.2 is bounded as*

$$\begin{aligned} &|\mathbf{E}_\mathsf{Q}[\delta_{i+1}^{\Delta\widehat{Y}} | X_i, K_i]| \\ &\quad \leq \exp(\frac{1}{2}\|D_i\|^2)\ \mathbf{E}_\mathsf{P}[(\delta_{i+1}^{\Delta\widehat{Y}})^2 | X_i, K_i]^{1/2}. \end{aligned} \tag{27}$$

PROOF. See Appendix D. ∎

Although the error bound in Proposition 3.5 suggests that the bias grows rapidly with $\|D_i\|$, when this magnitude is small ($\|D_i\| \leq 1$) the first term in the product on the right hand side of the inequality is bounded by $\sqrt{e} \approx 1.65$. This suggests that in the selection of K_i, the magnitude of the difference $F_i^\mu - K_i$ should not be significantly higher than the magnitude of the diffusion as specified by Σ_i. This result justifies the assumption that for appropriately chosen K_i, the proposed estimators have relatively low bias and low variance. It also provides some guidance on how to select K_i.

Furthermore, note that if K_i is selected so that the difference $F_i^\mu - K_i$ is bounded, e.g., using the modification (8) to ensure that $\|F_i^\mu - K_i\| < C$ for some target drift $\widehat{K}_i \approx K_i$ and some (possibly, large) constant $C > 0$, then, as discussed in Remark 2.1, the continuous analog of the discrete-time problem will satisfy Novikov's condition, as required in Theorem 2.1.

4 Policy Improvement

In this section we discuss how we can improve the policy based on the value function parameters obtained from the backward passes in the context of reinforcement learning. According to the discussion in the previous section, we propose an alternative Taylor-based approach to policy improvement as follows. We begin with a discrete approximation of the continuous-time problem and form the Q-value function at time i, given the value function V_{i+1}^μ, as usual,

$$Q_i(x, u; V_{i+1}^\mu) := L_i(x, u) + \mathbf{E}[V_{i+1}^\mu(X_{i+1}^{x,u}) | X_i = x], \tag{28}$$

where $X_{i+1}^{x,u} := x + F_i(x, u) + \Sigma_i W_i$, corresponds to the forward difference step with $x_i = x, u_i = u$ and normally distributed W_i. For the optimal control problem defined by (F, L, Σ, g, N), let V^*, π^* refer to the optimal value function and the optimal policy, respectively. The Bellman equation states that the optimal policy satisfies $\pi_i^*(x) \in \arg\min_{u \in U} Q_i(x, u; V_{i+1}^*)$ and the optimal value function satisfies $V_i^*(x) = \min_{u \in U} Q_i(x, u; V_{i+1}^*)$ [21], so approximations of the Q-value function can be utilized to obtain improved policies, especially when the current approximation of the optimal value function is nearly optimal.

Performing the same Taylor expansion as in (16), but centered at $\overline{X}_{i+1}^{x,u} := \mathbf{E}[X_{i+1}^{x,u}] = x + F_i(x, u)$, we arrive at the approximation $\widetilde{Q}_i \approx Q_i$ given by

$$\widetilde{Q}_i(x, u; \widetilde{V}_{i+1}^\mu) := L_i(x, u) + \overline{Y}_{i+1}^{x,u} + \frac{1}{2} \operatorname{tr}(\overline{M}_{i+1}^{x,u}), \tag{29}$$

where $\overline{M}_{i+1}^{x,u} := \Sigma_i^\top \partial_{xx} \widetilde{V}_{i+1}^\mu(\overline{X}_{i+1}^{x,u}) \Sigma_i$ and $\overline{Y}_{i+1}^{x,u} := \widetilde{V}_{i+1}^\mu(\overline{X}_{i+1}^{x,u})$.

Proposition 4.1 *The error when using* (29) *to approximate the Q-value function is*

$$Q_i^\mu(x, u; V_{i+1}^\mu) - \widetilde{Q}_i^\mu(x, u; \widetilde{V}_{i+1}^\mu) = \mathbf{E}[\delta_{i+1}^{\Delta\widehat{Y}^{x,u}}], \tag{30}$$

where $\delta_{i+1}^{\Delta\widehat{Y}x,u} := V_{i+1}^{\mu}(X_{i+1}^{x,u}) - \widetilde{V}_{i+1}^{\mu}(X_{i+1}^{x,u}) + \delta_{i+1}^{\text{h.o.t. } x,u}$.

PROOF. The Taylor expansion of $\widetilde{V}_{i+1}^{\mu}(X_{i+1}^{x,u})$ centered at $\overline{X}_{i+1}^{x,u}$ is $\widetilde{Y}_{i+1}^{x,u} := \overline{Y}_{i+1}^{x,u} + (\overline{Z}_{i+1}^{x,u})^{\top} W_i + \frac{1}{2} W_i^{\top} \overline{M}_{i+1}^{x,u} W_i$, so the r.h.s. of (29) is $L_i(x,u) + \mathbf{E}[\widetilde{Y}_{i+1}^{x,u}]$. Substituting $\widetilde{V}_{i+1}^{\mu}(X_{i+1}^{x,u}) = \widetilde{Y}_{i+1}^{x,u} + \delta_{i+1}^{\text{h.o.t. } x,u}$ and subtracting both sides of (29) from (28) yields the desired result. ∎

In practice, we seek a policy π_i, improved over μ_i from the previous iteration, with smaller Q-value function, that is, $\widetilde{Q}_i(x, \pi_i(x); \widetilde{V}_{i+1}^{\mu}) \leq \widetilde{Q}_i(x, \mu_i(x); \widetilde{V}_{i+1}^{\mu})$. A potential method is to use the policy

$$\mu_i^*(x; \widetilde{V}_{i+1}^{\mu}) := \min_{u \in U} \widetilde{Q}_i(x, u; \widetilde{V}_{i+1}^{\mu}). \tag{31}$$

Similarly to the previous section, when $\widetilde{V}_{i+1}^{\mu}$ is quadratic the Taylor expansion used in this estimator is exact. Thus, this optimization will yield the exact optimal control solution for an LQ problem.

4.1 Iterative-FBSDE Numerical Method

The iFBSDE approach begins by approximating the distribution of $\{X_i^0\}_{i=0}^N$ in P^0 through Monte-Carlo techniques for some initial $\{K_i^0\}_{i=0}^N$. The initial target policy μ^0 can be specified in a variety of ways. One possibility is to use whatever policy was used to generate $\{K_i^0\}_{i=0}^N$, such that $K_i^0 \equiv F_i^{\mu^0}$, making the first backwards pass an on-policy pass. Another possibility is to generate μ_i^0 during the backward pass as $\mu_i^0 = \mu_i^*(x; \widetilde{V}_{i+1}^{\mu^0})$, as in (31). This is allowable because μ_i^0 is not needed during the forward sampling pass and only needed after $\widetilde{V}_{i+1}^{\mu^0}$ is already estimated. The drift of the forward pass in the subsequent iteration $\{K_i^1\}_{i=0}^N$ can be informed by the latest optimizing policy $\mu_i^*(x; \widetilde{V}_{i+1}^{\mu^0})$. Alternatively, the estimators and policy improvement techniques presented here can be employed in methods such as those presented in [11], which allow for the broad exploration of the state space without a prior.

5 Numerical Results

In this section, we numerically evaluate and compare the proposed Taylor estimators to the naïve Euler-Maruyama estimators on three problems: two nonlinear problems of state dimension $n = 1$ and $n = 2$, and an LQ 4-dimensional problem. The estimators evaluated in this section are summarized in Table 1.

It is worth noting that while the first two examples do not enjoy the guarantees for the existence of classical solutions, they are guaranteed to possess unique viscosity solutions [24, Chapter 7, Theorem 4.4] and regardless of the smoothness of the value function, the use of smooth basis functions to produce function estimators is justified by the fact that a viscosity solution is an upper-

Table 1
Expressions for the proposed noiseless and re-estimate estimators, as well as the competing Euler-Maruyama estimators. The Euler-Maruyama Noisy estimator is an application of Euler-Maruyama to (5), where its noiseless counterpart is a variance-reduced version of the same, proposed in [5].

Estimator	$\widehat{Y}_i$
Taylor Noiseless	$L_i^{\mu} + \overline{Y}_{i+1} + \overline{Z}_{i+1}^{\top} D_i + \frac{1}{2} \operatorname{tr}(\overline{M}_{i+1}(I + D_i D_i^{\top}))$
Taylor Re-estimate	$\widetilde{V}_{i+1}^{\mu}(X_{i+1}) + L_i^{\mu} - \overline{Z}_{i+1}^{\top} W_i^{\mathsf{P}} + \overline{Z}_{i+1}^{\top} D_i + \frac{1}{2} \operatorname{tr}(\overline{M}_{i+1}(I + D_i D_i^{\top} - W_i^{\mathsf{P}} W_i^{\mathsf{P}\top}))$
Euler-Maru. Noiseless [5]	$\widetilde{V}_{i+1}^{\mu}(X_{i+1}) + L_i^{\mu} + \widetilde{Z}_{i+1}^{\top} D_i$
Euler-Maru. Noisy	$\widetilde{V}_{i+1}^{\mu}(X_{i+1}) + L_i^{\mu} - \widetilde{Z}_{i+1}^{\top} W_i^{\mathsf{P}} + \widetilde{Z}_{i+1}^{\top} D_i$

(respectively lower-) envelope to a smooth sub- (respectively super-) solution (see, e.g., [9] or [24, p. 197-8]).

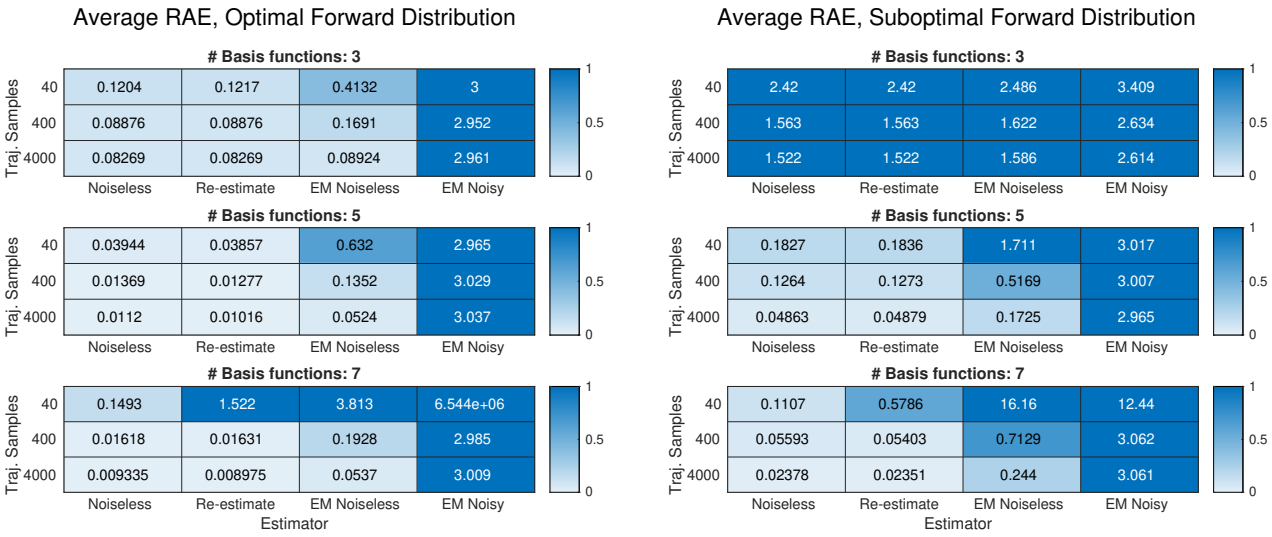
We assume for each example that K_i is selected such that the difference $F_i^{\mu} - K_i$ is bounded by some constant using a construction similar to (8) in Remark 2.1, thus ensuring that the continuous analogs of the examples will satisfy Novikov's condition. Furthermore, for the examples with quadratic cost, we tacitly assume that they are, in fact, only locally quadratic, growing linearly once $\|x\|$ surpasses some (large) constant. This will ensure that in the corresponding continuous SDE formulation the dynamics and cost functions are uniformly Lipschitz, as required by Theorem 2.1.

5.1 Nonlinear 1D Example

Consider the scalar optimal control problem with the dynamics and cost

$$\mathrm{d}X_s = \big(0.1(X_s - 3)^2 + 0.2u_s\big)\mathrm{d}s + 0.8\,\mathrm{d}W_s, \quad x_0 = 7,$$
$$J_t(u_{[t,T]}) = \mathbf{E}\bigg[\int_t^T \big(12\,|X_s - 6| + 0.4\,u_s^2\big)\,\mathrm{d}s + 25\,X_T^2\bigg],$$

over a time interval of length $T = 10$, with $N = 200$ timesteps. We compute a ground-truth optimal value function V_i^* and the optimal policy π^* by directly evaluating the optimal Bellman equation using a finely-gridded state and control space. The values for $\mathbf{E}[V_{i+1}^*(X_{i+1}^{x,u}) | X_i = x, u_i = u]$ are computed by interpolating a convolution which evaluates the expectation over W_i, namely, $V_{i+1}^{*\text{smooth}}(x) = \sum_j p(w_j; \Sigma) V_{i+1}^*(x + w_j)$, where $p(w_j; \Sigma)$ is the probability density of ΣW_i at w_j. The optimal value function is visualized in Fig. 1 (the yellow surface), along with two forward-backward trajectory distributions $\{(X_i, Y_i)\}$ considered for evaluation: (a) the optimal $K_i^{\text{optimal}} = F_i^{\pi^*}$ (the cyan trajectories), and (b) the suboptimal $K_i^{\text{subopt}} = -0.2X_i$ (the orange trajectories). We ran a series of simulations to investigate how



(a) Optimal forward sampling distribution generated with K^{optimal} (On-policy estimators).

(b) Suboptimal forward sampling distribution generated with K^{subopt} (Off-policy estimators).

Fig. 2. Heatmaps of experiments comparing the proposed estimators (Noiseless/Re-estimate) against naïve estimators (EM Noiseless/EM Noisy), with varying numbers of basis functions and numbers of trajectory samples. Each matrix element is the relative absolute error of the value function averaged over both 20 trials and $N = 200$ timesteps.

each estimator performs under different algorithmic conditions, visualized in Fig. 2. Each trial has one forward pass and a single backward pass, corresponding to each estimator. For the purposes of fair comparison we choose the target policy to be the ground-truth optimal policy $\mu = \pi^*$, but the next step value function $\widetilde{V}_{i+1}^{\mu}$ is the approximation produced by that estimator for the previous step in the backward pass. Chebyshev polynomials are used to locally approximate the optimal value function. For evaluation we use the relative absolute error (RAE) metric [23, Chapter 5]

$$\frac{\sum_{x\in\mathcal{C}_i} |\widetilde{V}_i(x) - V_i^*(x)|}{\sum_{x\in\mathcal{C}_i} |\sum_{y\in\mathcal{C}_i} \frac{1}{|\mathcal{C}_i|} V_i^*(y) - V_i^*(x)|}, \tag{32}$$

where $\mathcal{C}_i := \{\overline{x}_i - 3\sigma_i, \ldots, \overline{x}_i + 3\sigma_i\}$ and $\overline{x}_i, \sigma_i$ are the mean and standard deviation [6] of X_i. For each element in Fig. 2 we average the RAE approximations (32) over 20 trials and $N = 200$ time steps.

The results show that in all cases the proposed Taylor-based estimators perform as well as the Euler-Maruyama estimators and for the vast majority perform significantly better. Although the Taylor-based estimators generally perform equally well, there are slight differences in how they perform under different conditions. The Taylor-noiseless estimator seems to outperform the re-estimate estimator when the number of trajectory samples is low, and vice versa when the number is high. Recall that the error analysis suggests that the re-estimate estimator has lower bias but higher variance than the Taylor-noiseless estimator. The simulated results confirm the theoretical results, that is, when the number of trajectory samples is low, high variance makes the re-estimate estimator perform poorly, but when there are enough samples to overcome the variance in the estimator, the low bias properties can result in better accuracy. In typical usage, however, it is likely that the low variance of the Taylor-noiseless estimator is preferable for its simplicity and lower variance.

[6] A small positive constant is used instead if the standard deviation is excessively small.

5.2 L^1 *Inverted Pendulum*

Next, we compared the estimators on a 2-dimensional inverted pendulum problem with dynamics and cost given as follows

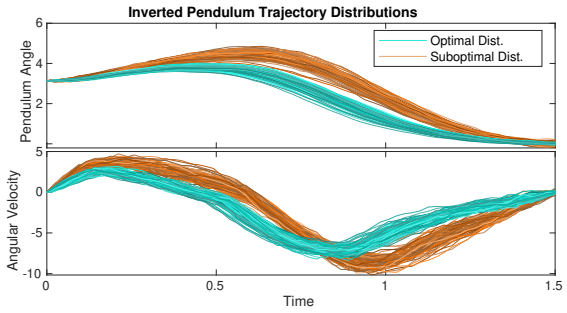
$$\mathrm{d}X_s = \begin{bmatrix} X_s^2 \\ 0.4X_s^2 + 19.62\sin(X_s^1) + 19.62u \end{bmatrix} \mathrm{d}s + \begin{bmatrix} 0.04 & 0 \\ 0 & 0.4 \end{bmatrix} \mathrm{d}W_s, \qquad u \in [-1, 1],$$

$$J_t(u_{[t,T]}) = \mathbf{E}\left[\int_t^T 0.2\,|u_s|\,\mathrm{d}s + 4\,(X_T^1)^2 + 2\,(X_T^2)^2 \right],$$

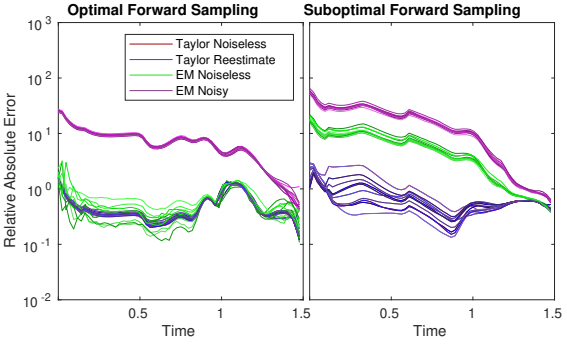
where $x_0 = [0, \pi]^\top$, and the discretization uses $N = 64$ time steps. Note that the cost is different than most approaches to this problem since it has an L^1 penalty in terms of the control, making the optimal policy bang-bang-bang, that is, always contained in the discrete set $\pi^*(x) \in \{-1, 0, 1\}$. We used normalized Chebyshev polynomials of degree 2 and lower for the linear basis functions used in the representation of

$\widetilde{V}^{\mu}$. The suboptimal sampling distribution drift was $K_i^{\text{subopt}} = F_i^* + [k_1\widetilde{W}_i^1, k_2\widetilde{W}_i^2 + k_3(N-i)]^\top$, where F_i^* is the problem dynamics driven by the optimal policy, k_1, k_2, k_3 are constants, and $\widetilde{W}_i^1, \widetilde{W}_i^2$ are normally distributed random variables independent of the problem's noise W_i^1, W_i^2. The trajectory distributions include $M = 2,000$ trajectory samples.

The optimal and suboptimal forward distributions are visualized in Fig. 3(a). A comparison of the RAE, now computed over a 2-dimensional grid of the same width, for each of the four estimators is visualized in Fig. 3(b). The Taylor estimators again outperform the EM estimators by at least an order of magnitude for most of the backward pass on the suboptimal forward sampling condition. Although for the optimal sampling condition the EM Noiseless estimator performs about as well as the Taylor estimators on average, it has higher variance and is thus less reliable. Again, between the Taylor estimators they show nearly equivalent performance.



(a) Trajectory distributions for the two sampling conditions (K_i^{optimal} / K_i^{subopt}).



(b) Accuracy of value function approximation $\widetilde{V}_i^{\mu}$ compared to ground-truth evaluated via relative absolute error (32). The red Taylor Noiseless lines are almost entirely overlapped by the blue Taylor Re-estimate lines.

Fig. 3. Comparison of accuracy of estimators on a 2-dimensional inverted pendulum problem with L^1 running cost.

5.3 LQ 4D Problem

We also tested the proposed estimators on a linearized version of the 4-dimensional finite time cart-pole problem [22] with initial condition $x_0 = [0, 0, \pi/9, 0]^\top$ and $\sigma = \text{diag}(0.01, 0.1, 0.01, 0.1)$. For the suboptimal sampling distribution we selected a time-invariant linear closed-loop feedback policy K_i^{subopt} corresponding to a feedback gain matrix $\begin{bmatrix}0 & 0 & 0.5 & 0.2\end{bmatrix}$. The optimal policy is found through the solution of the associated Riccati equations (distributions visualized in Fig. 4(a)). The value function model for $\widetilde{V}$ again used Chebyshev functions of degree 2 and lower (15 basis functions). The RAE metrics, now computed over a 4-dimensional grid of the same width, (32) are visualized in Fig. 4(b).

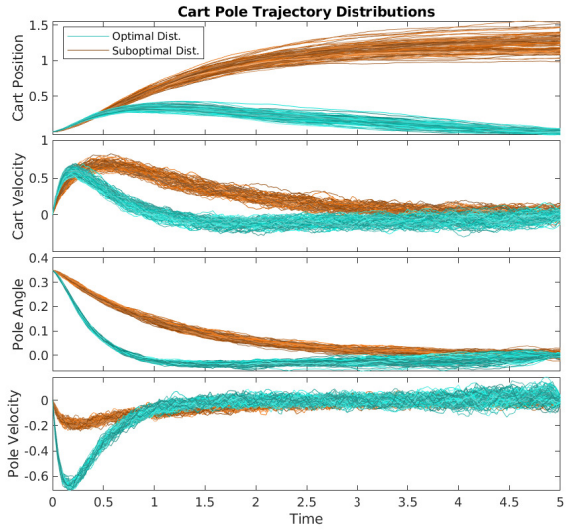
As predicted by the error analysis, since this is an LQ problem and the value function is in the class of quadratic functions, the Taylor expansion-based estimators are able to produce approximations of the value function with accuracy near machine precision for both conditions. For the suboptimal forward sampling the EM estimators diverge quickly during the backward pass. For the optimal forward sampling condition the EM estimators did not perform as well compared to the value function's variance and their error is still several orders of magnitudes higher than the Taylor estimators.

These results confirm that the proposed estimators are able to achieve near perfect performance on the most common problem in stochastic optimal control, namely, linear dynamics with quadratic cost (LQ). Further, they confirm that utilizing second-order derivatives of the value function is crucial for accurate Girsanov-inspired off-policy estimator schemes, contrary to what naïve application of the theory would suggest.

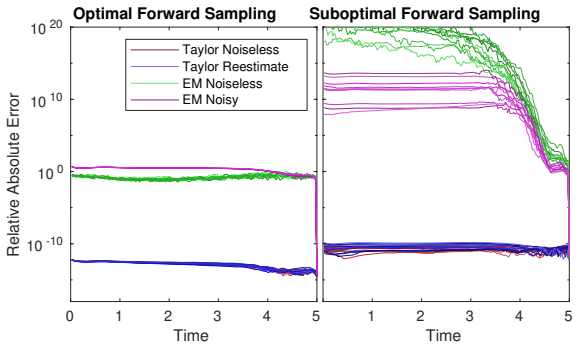
6 Conclusion

We have demonstrated that Taylor-based estimators for numerically solving Feynman-Kac FBSDEs are significantly more accurate than naïve Euler-Maruyama-based estimators through both error analysis and numerical simulation. These estimators are derived by using higher-order Taylor expansions and follow the spirit of the continuous-time Feynman-Kac-Girsanov formulation. Both error analysis and numerical simulations confirm that these estimators have very high accuracy when applied to LQ problems. Further, in simulation, the proposed estimators are orders of magnitude more accurate than the EM estimators in both LQ and nonlinear problems. This paper also proposes a method to use the estimated value function parameters for generating an improved policy in reinforcement learning problems.

Moving forward, the primary challenge with Feynman-Kac FBSDE methods is how to produce robust iterative methods. Although value function approximation can be extremely accurate in the proximity of the initial forward pass, even for off-policy methods, Runge's phenomenon begins dominating outside the sampling distribution. As a consequence, when in some extrapolative region the approximation significantly underestimates the true value function, policy improvement begins to fail and future iterations are constructed based on divergent policies with little room for improvement aside from starting



(a) Trajectory distributions for the two sampling conditions (K_i^{optimal} / K_i^{subopt}).



(b) Accuracy of the value function approximation $\widetilde{V}_i^{\mu}$ compared to ground-truth over time, evaluated via relative absolute error (32).

Fig. 4. Comparison of accuracy of estimators on a 4-dimensional LQ approximation of cart-pole balancing system.

over. To overcome such difficulties, the proposed estimators can be integrated into model-based policy gradient techniques. By alternating between small batches of trajectory samples and small changes to the policy, the trajectory distribution avoids moving significantly off-policy into regions where the current policy and value function estimates are invalid. Although our approach appears similar to [13], our estimators utilize dynamics models without differentiating the drift term or the running cost, instead leveraging only derivatives of the local value function with respect to the state. Further, our estimators are more closely related to off-policy Bellman residual updates as discussed in [21]. Unlike typical off-policy Bellman updates, however, our estimators are nearly free from bias because they directly compensate for taking a step off-policy.

Acknowledgements

This work has been supported by NSF awards CMMI-1662523 and IIS-2008686 and ONR award N00014-18-1-2828. The authors would like to thank Evangelos Theodorou for many helpful discussions as well as the anonymous reviewers for their insightful comments.

References

[1] Samuel N Cohen and Robert James Elliott. *Stochastic calculus and applications*, volume 2. Springer, 2015.

[2] Edwin L Crow and Kunio Shimizu. *Lognormal Distributions*. Marcel Dekker New York, 1987.

[3] Giovanni B Di Masi and Wolfgang J Runggaldier. On measure transformations for combined filtering and parameter estimation in discrete time. *Systems & Control Letters*, 2(1):57–62, 1982.

[4] Nicole El Karoui, Shige Peng, and Marie Claire Quenez. Backward stochastic differential equations in finance. *Mathematical Finance*, 7(1):1–71, 1997.

[5] Ioannis Exarchos and Evangelos A. Theodorou. Stochastic optimal control via forward and backward stochastic differential equations and importance sampling. *Automatica*, 87:159–165, 2018.

[6] Ioannis Exarchos, Evangelos A. Theodorou, and Panagiotis Tsiotras. Stochastic Differential Games: A Sampling Approach via FBSDEs. *Dynamic Games and Applications*, 2018.

[7] Ioannis Exarchos, Evangelos A. Theodorou, and Panagiotis Tsiotras. Stochastic L^1-optimal control via forward and backward sampling. *Systems and Control Letters*, 118:101–108, 2018.

[8] Wendell H. Fleming and Raymond W. Rishel. Deterministic and stochastic optimal control. *Bulletin of the American Mathematical Society*, 82:869–870, 1976.

[9] Wendell H Fleming and Domokos Vermes. Convex Duality Approach to the Optimal Control of Diffusions. *SIAM journal on control and optimization*, 27(5):1136–1155, 1989.

[10] K. Hawkins, A. Pakniyat, E. Theodorou, and P. Tsiotras. Forward-backward rapidly-exploring random trees for stochastic optimal control. *arXiv preprint arXiv:2006.12444*, 2020.

[11] Kelsey P. Hawkins, Ali Pakniyat, Evangelos Theodorou, and Panagiotis Tsiotras. Forward-backward rapidly-exploring random trees for stochastic optimal control. In *2021 60th IEEE Conference on Decision and Control*, Austin, TX, Dec. 13–15, 2021, pages 912–917.

[12] Kelsey P Hawkins, Ali Pakniyat, and Panagiotis Tsiotras. On the Time Discretization of the Feynman-Kac Forward-Backward Stochastic Differential Equations for Value Function Approximation. In *60th IEEE Conference on Decision and Control*, Austin, TX, Dec. 13–15, 2021, pages 892–897.

[13] Nicolas Heess, Greg Wayne, David Silver, Timothy Lillicrap, Yuval Tassa, and Tom Erez. Learning continuous control policies by stochastic value gradients. *arXiv preprint arXiv:1510.09142*, 2015.

[14] David H. Jacobson and David Q. Mayne. *Differential dynamic programming*. North-Holland, New York, NY, 1970.

[15] Peter E Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23. Springer Science and Business Media, 2013.

[16] Francis A. Longstaff and Eduardo S. Schwartz. Valuing American options by simulation: A simple least-squares approach. *Review of Financial Studies*, 14:113–147, 2001.

[17] George Lowther. Girsanov transformations, May 2010.

[18] Jin Ma and Jiongmin Yong. *Forward-Backward Stochastic Differential Equations and their Applications*. Springer, 2007.

[19] E. Pardoux and S. G. Peng. Adapted solution of a backward stochastic differential equation. *Systems and Control Letters*, 14(1):55–61, 1990.

[20] Halil M. Soner and Nizar Touzi. A stochastic representation for the level set equations. *Communications in Partial Differential Equations*, 27(9-10):2031–2053, 2002.

[21] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[22] Russ Tedrake. Underactuated Robotics: Learning, Planning, and Control for Efficient and Agile Machines: Course Notes for MIT 6.832. *Working Draft Edition*, 3, 2009.

[23] Ian H Witten, Eibe Frank, and Mark A Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Inc., 3rd edition, 2011.

[24] Jiongmin Yong and Xun Yu Zhou. *Stochastic Controls: Hamiltonian Systems and HJB Equations*, volume 43. Springer, 1999.

Appendix

A Proof of Theorem 3.2

PROOF. Using (12) and the result (19) of Lemma 3.1 we have $\widehat{Y}_i := \widehat{Y}_{i+1} - \Delta\widehat{Y}_i = \widehat{Y}_{i+1} - \Delta Y_i + (\delta_{i+1}^{\Delta\widehat{Y}} - \mathbf{E}_{\mathsf{Q}}[\delta_{i+1}^{\Delta\widehat{Y}}|X_i, K_i])$, and so the general expression for the bias is $\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}_i|X_i, K_i] = \mathbf{E}_{\mathsf{P}}[Y_{i+1} - \widehat{Y}_{i+1}|X_i, K_i] + \mathbf{E}_{\mathsf{Q}}[\delta_{i+1}^{\Delta\widehat{Y}}|X_i, K_i] - \mathbf{E}_{\mathsf{P}}[\delta_{i+1}^{\Delta\widehat{Y}}|X_i, K_i]$. The variance of the estimator is $\mathrm{Var}_{\mathsf{P}}[\widehat{Y}_i|X_i, K_i] = \mathrm{Var}_{\mathsf{P}}[\widehat{Y}_{i+1} - \Delta Y_i + (\delta_{i+1}^{\Delta\widehat{Y}} - \mathbf{E}_{\mathsf{Q}}[\delta_{i+1}^{\Delta\widehat{Y}}|X_i, K_i])|X_i, K_i] = \mathrm{Var}_{\mathsf{P}}[\delta_{i+1}^{\Delta\widehat{Y}} - (Y_{i+1} - \widehat{Y}_{i+1})|X_i, K_i]$, noting that we can drop the terms Y_i and $\mathbf{E}_{\mathsf{Q}}[\delta_{i+1}^{\Delta\widehat{Y}}|X_i, K_i]$ because they are (X_i, K_i)-measurable.

For the re-estimate estimator we have $Y_{i+1} - \widehat{Y}_{i+1}^{\text{re-est}} = V_{i+1}^{\mu}(X_{i+1}) - \widetilde{V}_{i+1}^{\mu}(X_{i+1})$, and for the noiseless estimator we have $Y_{i+1} - \widehat{Y}_{i+1}^{\text{noiseless}} = V_{i+1}^{\mu}(X_{i+1}) - \widetilde{Y}_{i+1} = V_{i+1}^{\mu}(X_{i+1}) - (\widetilde{V}_{i+1}^{\mu}(X_{i+1}) - \delta_{i+1}^{\text{h.o.t.}\cdot}) = \delta_{i+1}^{\Delta\widehat{Y}}$, due to (16). Plugging these two equalities into the general expressions for the bias and variance yields the result. ∎

B Proof of Proposition 3.3

PROOF. A separate application of Taylor's theorem to $\partial_x\widetilde{V}_{i+1}^{\mu}(X_{i+1})$ can be used to show that $\widetilde{Z}_{i+1} = \overline{Z}_{i+1} + \overline{M}_{i+1}W_i^{\mathsf{P}} + \Sigma_i^{\top}\widetilde{\delta}_{i+1}^{\text{h.o.t.}\cdot}$, where $\widetilde{\delta}_{i+1}^{\text{h.o.t.}\cdot}$ is a new set of residual terms of order three and higher. Substituting $\widetilde{Z}_{i+1}$ and (16)-(17) into the definition of $\widehat{Y}_i^{\text{em-nless}}$, we have $\widehat{Y}_i^{\text{em-nless}} = L_i^{\mu} + \overline{Y}_{i+1} + \overline{Z}_{i+1}^{\top}W_i^{\mathsf{P}} + \frac{1}{2}(W_i^{\mathsf{P}})^{\top}\overline{M}_{i+1}W_i^{\mathsf{P}} + \delta_{i+1}^{\text{h.o.t.}\cdot} + \overline{Z}_{i+1}^{\top}D_i + D_i^{\top}\overline{M}_{i+1}W_i^{\mathsf{P}} + D_i^{\top}\Sigma_i^{\top}\widetilde{\delta}_{i+1}^{\text{h.o.t.}\cdot}$. If we substitute this into $\mathbf{E}_{\mathsf{P}}[\widehat{Y}_i^{\text{em-nless}} - \widehat{Y}_i^{\text{noiseless}}|X_i, K_i]$ and then substitute in (23)-(24), we get $\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}_i^{\text{em-nless}}|\cdot] = \mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}_i^{\text{re-est}}|\cdot] + \frac{1}{2}D_i^{\top}\overline{M}_{i+1}D_i - \mathbf{E}_{\mathsf{P}}[D_i^{\top}\Sigma_i^{\top}\widetilde{\delta}_{i+1}^{\text{h.o.t.}}|\cdot]$.

For the variance result, when taking the conditional variance of $\widehat{Y}_i^{\text{em-nless}}$, the terms $L_i^{\mu}, \overline{Y}_{i+1}, D_i^{\top}\overline{Z}_{i+1}$ drop out because they are (X_i, K_i)-measurable, which results in $\mathrm{Var}_{\mathsf{P}}[\widehat{Y}_i^{\text{em-nless}}|\cdot] = \mathrm{Var}_{\mathsf{P}}[\delta_{i+1}^{\text{h.o.t.}\cdot}|\cdot] + \|\overline{Z}_{i+1} + \overline{M}_{i+1}D_i\|^2 + \mathrm{Var}_{\mathsf{P}}[\frac{1}{2}(W_i^{\mathsf{P}})^{\top}\overline{M}_{i+1}W_i^{\mathsf{P}}|\cdot] + \mathrm{Var}_{\mathsf{P}}[D_i^{\top}\Sigma_i^{\top}\widetilde{\delta}_{i+1}^{\text{h.o.t.}}|\cdot] + \cdots$ where the remainder of the terms are covariances between the terms in $\widehat{Y}_i^{\text{em-nless}}$. Since the second two variance terms are non-negative, we now only need to prove that covariance terms are not significantly large and negative.

Every covariance term but one contains higher order terms and is thus, by our assumptions, relatively small. The only covariance term without higher order terms is $\mathrm{Cov}_{\mathsf{P}}[(\overline{Z}_{i+1} + \overline{M}_{i+1}D_i)^{\top}W_i^{\mathsf{P}}, \frac{1}{2}(W_i^{\mathsf{P}})^{\top}\overline{M}_{i+1}W_i^{\mathsf{P}}|\cdot] = 0$. This can be shown by noting that, for any vector Z and matrix M measurable with respect to the conditional expectation and normally distributed vector W, $\mathrm{Cov}[Z^{\top}W, W^{\top}MW|\cdot] = \sum_{i,j,k}\mathbf{E}[W_iW_jW_k|\cdot]Z_kM_{i,j}$, and since, for distinct i, j, k, $\mathbf{E}[W_iW_jW_k|\cdot] = \mathbf{E}[W_i^2W_j|\cdot] = \mathbf{E}[W_i^3|\cdot] = 0$ by the properties of normal vectors, then $\mathbf{E}[W_iW_jW_k|\cdot] = 0$ for all i, j, k. ∎

C Proof of Proposition 3.4

PROOF. The assumptions of the proposition imply that there exists a constant $0 \leq \alpha \ll 1/7$ such that each of the following terms (conditioned on (X_i, K_i)) $|\mathbf{E}_{\mathsf{Q}}[V_{i+1}^{\mu}(X_{i+1}) - \widetilde{V}_{i+1}^{\mu}(X_{i+1})|\cdot]|$, $|\mathbf{E}_{\mathsf{Q}}[\delta_{i+1}^{\text{h.o.t.}}|\cdot]|$, $|\mathbf{E}_{\mathsf{P}}[\delta_{i+1}^{\text{h.o.t.}}|\cdot]|$, $|\mathbf{E}_{\mathsf{P}}[D_i^{\top}\Sigma_i^{\top}\widetilde{\delta}_{i+1}^{\text{h.o.t.}}|\cdot]| \leq \alpha|\frac{1}{2}D_i^{\top}\overline{M}_{i+1}D_i|$. In light of these assumptions, the triangle inequality immediately yields that $|\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}_i^{\text{re-est}}|\cdot]| \leq 3\alpha|\frac{1}{2}D_i^{\top}\overline{M}_{i+1}D_i|$. A second application yields $|\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}_i^{\text{re-est}}|\cdot] - \mathbf{E}_{\mathsf{P}}[D_i^{\top}\Sigma_i^{\top}\widetilde{\delta}_{i+1}^{\text{h.o.t.}}|\cdot]| \leq 4\alpha|\frac{1}{2}D_i^{\top}\overline{M}_{i+1}D_i|$. Applying the reverse triangle inequality gives the result $|\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}_i^{\text{em-nless}}|\cdot]| = |\frac{1}{2}D_i^{\top}\overline{M}_{i+1}D_i + \mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}_i^{\text{re-est}}|\cdot] - \mathbf{E}_{\mathsf{P}}[D_i^{\top}\Sigma_i^{\top}\widetilde{\delta}_{i+1}^{\text{h.o.t.}}|\cdot]| \geq ||\frac{1}{2}D_i^{\top}\overline{M}_{i+1}D_i| - |\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}_i^{\text{re-est}}|\cdot] - \mathbf{E}_{\mathsf{P}}[D_i^{\top}\Sigma_i^{\top}\widetilde{\delta}_{i+1}^{\text{h.o.t.}}|\cdot]|| \geq (1 - 4\alpha)|\frac{1}{2}D_i^{\top}\overline{M}_{i+1}D_i| \geq (1-4\alpha)|\frac{1}{2}D_i^{\top}\overline{M}_{i+1}D_i| \geq |\mathbf{E}_{\mathsf{P}}[Y_i - \widehat{Y}_i^{\text{re-est}}|\cdot]|$. ∎

D Proof of Theorem 3.5

PROOF. As a result of the change of measure defined in the discrete-time Girsanov theorem [3, Theorem 1], we have $\mathbf{E}_{\mathsf{Q}}[\delta_{i+1}^{\Delta\widehat{Y}}|X_i, K_i] = \mathbf{E}_{\mathsf{P}}[\varphi(D_i, W_i^{\mathsf{P}})\delta_{i+1}^{\Delta\widehat{Y}}|X_i, K_i]$, where $\varphi(d, w) := \exp(-\frac{1}{2}\|d\|^2 + d^{\top}w)$. By the Cauchy-Schwartz inequality, we have that $|\mathbf{E}_{\mathsf{Q}}[\delta_{i+1}^{\Delta\widehat{Y}}|X_i, K_i]| \leq \mathbf{E}_{\mathsf{P}}[\varphi(D_i, W_i^{\mathsf{P}})^2|X_i, K_i]^{1/2}\mathbf{E}_{\mathsf{P}}[(\delta_{i+1}^{\Delta\widehat{Y}})^2|X_i, K_i]^{1/2}$. Using properties of log-normal distributions [2] we have $\mathbf{E}_{\mathsf{P}}[\varphi(D_i, W_i^{\mathsf{P}})^2|X_i, K_i] = \mathbf{E}_{\mathsf{P}}[\exp(\|D_i\|^2)|X_i, K_i] = \exp(\|D_i\|^2)$, which, upon substitution, yields the desired result. ∎