

# A deep solver for backward stochastic Volterra integral equations

Alessandro Gnoatto\*    Camilo Andrés García Trillos†    Kristoffer Andersson‡

May 23, 2025

## Abstract

We present the first deep-learning solver for backward stochastic Volterra integral equations (BSVIEs) and their fully-coupled forward-backward variants. The method trains a neural network to approximate the two solution fields in a single stage, avoiding the use of nested time-stepping cycles that limit classical algorithms. For the decoupled case we prove a non-asymptotic error bound composed of an a posteriori residual plus the familiar square root dependence on the time step. Numerical experiments confirm this rate and reveal two key properties: *scalability*, in the sense that accuracy remains stable from low dimension up to 500 spatial variables while GPU batching keeps wall-clock time nearly constant; and *generality*, since the same method handles coupled systems whose forward dynamics depend on the backward solution. These results open practical access to a family of high-dimensional, path-dependent problems in stochastic control and quantitative finance.

## 1 Introduction

Backward stochastic Volterra integral equations (BSVIEs) represent a natural extension of backward stochastic differential equations (BSDEs) by allowing for memory effects and more general dependence structures. This makes them well-suited for problems in finance, control theory, and other applications where past states influence future evolution.

To motivate the need for BSVIEs, first recall a standard situation. Assume the state process satisfies the forward stochastic differential equation (FSDE)

$$X_t = x_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s.$$

For functions  $g$  and  $f$  that depend on arguments taken at a single time, the quantity

$$Y_t = \mathbb{E} \left[ g(X_T) + \int_t^T f(s, X_s) ds \mid \mathcal{F}_t \right]$$

admits the well known BSDE representation

$$Y_t = g(X_T) + \int_t^T f(s, X_s) ds - \int_t^T Z_s dW_s.$$

We now expand the setting by allowing  $g$  and  $f$  to also depend on the evaluation date  $t$ . Set

$$Y_t = \mathbb{E} \left[ g(t, X_T) + \int_t^T f(t, s, X_s) ds \mid \mathcal{F}_t \right]. \quad (1)$$

---

\*Department of Economics, University of Verona, via Cantarane, 24 - 37129 Verona, Italy.  
Email: alessandro.gnoatto@univr.it

†Department of Mathematics, University College London, Gower Street, London. Email: camilo.garcia@ucl.ac.uk

‡Department of Economics, University of Verona, via Cantarane, 24 - 37129 Verona, Italy.  
Email: kristofferherbet.andersson@univr.it

When this extra time argument can be factorized in a multiplicative way, a change of variables restores the ordinary BSDE framework. For illustration, let us consider an example that can be interpreted as a pricing model in finance. Take a possibly stochastic short rate  $r$  and define

$$g(t, x) = \exp\left(-\int_t^T r_s ds\right) \tilde{g}(x), \quad f(t, s, x) = \exp\left(-\int_t^s r_u du\right) \tilde{f}(s, x).$$

Introduce the discount factor  $M_t = \exp(-\int_0^t r_u du)$ . Since  $e^{-\int_t^s r_u du} = M_s/M_t$  we obtain

$$Y_t M_t = \mathbb{E}\left[M_T \tilde{g}(X_T) + \int_t^T M_s \tilde{f}(s, X_s) ds \mid \mathcal{F}_t\right].$$

The term inside the expectation no longer depends on the evaluation date. Hence  $Y_t M_t$  satisfies a classical BSDE and dividing by  $M_t$  recovers  $Y_t$ .

The need for a wider Volterra formulation appears once the second time argument cannot be removed by such a factorization. In this scenario, the backward representation of (1) is instead a BSVIE of the form

$$Y_t = g(t, X_T) + \int_t^T f(t, s, X_s) ds - \int_t^T Z_{t,s} dW_s.$$

Contrary to the BSDE case the control process  $Z$  now carries two time indices, where the additional dependence on the evaluation date  $t$  reflects the memory that the Volterra structure introduces and is discussed in detail in the following sections. Similar to the classical BSDE framework, the functions  $f$  and  $g$  may also depend on the solution pair  $(Y, Z)$  itself, so that the conditional expectation in (1) becomes implicit. In addition one can consider fully coupled systems in which the coefficients  $b$  and  $\sigma$  of the forward equation depend on the BSVIE solution.

Below we present two practical situations in which the conditional expectation naturally involves both the observation time  $t$  and the integration time  $s$ . In both cases, the two-clock dependence cannot be disentangled. Consequently, the standard BSDE machinery breaks down, making a Volterra formulation essential.

**Example 1.1** (Social discounting, see [7]). *A sovereign wealth fund must value uncertain long-dated cash-flows  $C_s$  over the horizon  $[0, T]$ . To reflect inter-generational welfare it adopts the declining hyperbolic kernel*

$$D(t, s) = \frac{1}{(1 + \alpha(s - t))^\beta}, \quad s \geq t, \quad \alpha, \beta > 0,$$

*whose slower decay ensures that payments decades ahead are not virtually ignored. The present social cost at time  $t$  is*

$$Y_t = \mathbb{E}\left[\int_t^T D(t, s) C_s ds \mid \mathcal{F}_t\right].$$

*Because  $D$  depends on both the evaluation time  $t$  and the integration time  $s$ , no driver of the form  $f(s, \cdot)$  exists, so a BSDE representation fails. The appropriate formulation is the BSVIE*

$$Y_t = \int_t^T D(t, s) C_s ds - \int_t^T Z_{t,s} dW_s,$$

*whose control field  $Z_{t,s}$  carries both time indices and captures the memory effect of the moving discount kernel.*

**Example 1.2.** *Suppose a bank enters into an OTC derivative position with terminal payoff  $g$  against a counterparty subject to default risk. The default risk of the counterparty is described by the deterministic positive hazard rate function  $\lambda$ . We assume that, in case of default, the reference value of the transaction is the whole value of the position including the valuation adjustment due to counterparty risk. This gives rise to the valuation formula (see e.g. [6])*

$$Y_t = \mathbb{E}\left[e^{-\int_t^T r_s ds} g(X_T) - LGD \int_t^T e^{-\int_t^s (r_u + \lambda_u) du} (Y_s)^+ ds \mid \mathcal{F}_t\right] \quad (2)$$

where  $r$  is the deterministic function describing the overnight rate and  $LGD \in (0, 1]$  is a constant loss given default. While the discount factor can be factorized as above, the presence of the hazard rate turns this valuation problem into a BSVIE

$$Y_t = e^{-\int_t^T r_s ds} g(X_T) - LGD \int_t^T e^{-\int_t^s (r_u + \lambda_u) du} (Y_s)^+ ds - \int_t^T Z_{t,s} dW_s \quad (3)$$

Beyond serving as backward representations of conditional expectations, BSVIEs provide a powerful framework for time inconsistent stochastic optimal control problems. When an objective functional changes with the initial date, whether through non exponential discounting, dynamic risk measures, or any other mechanism, the dynamic programming principle collapses and the classical BSDE adjoint is no longer adequate. Because a BSVIE driver may depend simultaneously on the evaluation time and the integration time, it naturally captures the evolving preferences and path dependence that come with time inconsistency, yielding equilibrium conditions that can be derived either through a Pontryagin maximum principle or, in an HJB-like formulation, by treating the BSVIE variable  $Y$  as the value function. In practice the BSVIE is coupled with an FSDE or a forward stochastic Volterra integral equation (FSVIE), producing a well-posed FSDE-BSVIE or forward-backward stochastic Volterra integral equation (FBSVIE) system whose solution characterizes admissible equilibrium controls even when the underlying value process is non-Markovian. See, for instance, [41, 33, 35] for time-inconsistent stochastic optimal control problems formulated as FBSVIEs. In [34] the authors study optimal control of FBSDEs, where the state itself follows a controlled FBSDE. This setting naturally leads to an FBSDE-BSVIE system and even accommodates a conditional mean variance portfolio optimization problem.

Finally, we mention dynamic risk measures, which is another arena where the two-time structure of a BSVIE is essential. When the exposure is a whole cash-flow stream and the risk weights or discount factors shift with the observation date, a classical BSDE cannot retain the resulting memory and horizon-dependence, whereas a BSVIE captures both through its simultaneous  $(t, s)$  driver. This representation underlies modern insurance reserving, capital allocation rules for banking groups, and portfolio policies constrained by pathwise VaR or CVaR limits; see, for example, [39, 10, 32].

Since the seminal work of [26] launched the field, the theory of BSVIEs has expanded rapidly, see, for instance, the introduction of type-II BSVIEs (when the driver takes both  $Z_{t,s}$  and  $Z_{s,t}$  as inputs) in [38], the well-posedness and regularity of M-solutions for type-II BSVIEs in [40] and control-theoretic applications surveyed in [37]. By contrast, the numerical side has hardly kept pace.

On the numerical side, the literature is remarkably thin: to the best of our knowledge there are only three published schemes for BSVIEs: (i) the finite-difference analysis in the PhD thesis of [29]; (ii) an implicit backward-Euler scheme for type-I BSVIEs proposed by [36]; and (iii) the recent explicit backward-Euler method for type-II equations of [15]. The discretization schemes put forward in [29, 36, 15] assume that the conditional expectations can be computed exactly; consequently, they remain semi-discrete and cannot be executed without an additional (unspecified) approximation layer. To the best of our knowledge, the algorithm developed in the following is the first fully implementable end-to-end solver for BSVIEs.

By contrast, BSDEs already have several neural-network solvers, pioneered by the seminal deep-BSDE method of Han, Jentzen, and E [17]. Following this work, several deep learning-based strategies have emerged, notably [3, 5, 30, 23, 2], with convergence analyses provided in, *e.g.*, [18, 20, 14, 22, 1, 31, 13]. Concurrently, a separate branch known as backward-type methods, closer in spirit to classical dynamic programming algorithms, has developed, see *e.g.*, [19, 8, 11, 27, 24, 12]. For a comprehensive survey of numerical methods for approximating BSDEs and PDEs, see [9], and for neural-network-based approaches specifically, we refer to [4]. Yet no neural-network-based method has been put forward for BSVIEs. This imbalance between abundant theory and scarce algorithms motivates the present work.

In this paper, we contribute to the literature on BSVIEs in the following ways:

1. We propose a neural network-based approximation method for BSVIEs. The method directly approximates the functional form of the BSVIE solution processes and does not rely on reformulating the problem as a family of BSDEs.

2. We prove that the overall (mean-squared) simulation error of our method can, up to a multiplicative constant, be bounded by the size of the temporal discretization and the mean-squared error of the free-term condition (the BSVIE analog of the terminal condition in BSDEs). In this sense, our analysis extends the results of [18] to the BSVIE framework.

Taken together, these two results deliver the first machine-learning solver for BSVIEs that comes with a non-asymptotic error guarantee, thereby closing the algorithmic gap identified above and paving the way for high-dimensional, path-dependent applications that were previously out of reach.

In Section 2 we state the class of equations studied and the assumptions that guarantee existence and uniqueness of their solutions. Section 3 reformulates the problem in variational form and introduces its Euler–Maruyama time discretization. The error analysis is carried out in Section 4. Section 5 describes the complete algorithm and the main neural-network details. Finally, Section 6 reports numerical results.

## 2 Preliminaries

This section sets out the notation and presents the two problem settings we study. First, we introduce the decoupled FBSVIE and list the assumptions that ensure a unique adapted solution. This is the setting used for the error analysis. We then show how the same algorithm applies to a coupled FSDE–BSVIE and briefly recall the known existence and uniqueness results for this case. The algorithm can be implemented for both settings, but the error analysis in this paper is carried out only for the first one.

Throughout this paper, we let  $T \in (0, \infty)$ ,  $d, \ell \in \mathbb{N}$ ,  $x_0 \in \mathbb{R}^d$ ,  $(W_t)_{t \in [0, T]}$  be an  $\ell$ –dimensional standard Brownian motion on a filtered probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [0, T]}, \mathbb{P})$  and define  $\Delta[0, T]^2 := \{(t, s) \in [0, T]^2 \mid t \leq s\}$ .

### 2.1 A decoupled FBSVIE

In this subsection, the problem coefficients are given by  $\varphi: [0, T] \rightarrow \mathbb{R}^d$ ,  $b: \Delta[0, T]^2 \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ ,  $\sigma: \Delta[0, T]^2 \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times \ell}$ ,  $g: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$  and  $f: \Delta[0, T]^2 \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^\ell \rightarrow \mathbb{R}$ . We consider a decoupled FBSVIE of the form

$$\begin{cases} X_t = \varphi(t) + \int_0^t b(t, s, X_s) ds + \int_0^t \sigma(t, s, X_s) dW_s, \\ Y_t = g(t, X_T) + \int_t^T f(t, s, X_s, Y_s, Z_{t,s}) ds - \int_t^T Z_{t,s}^\top dW_s. \end{cases} \quad (4)$$

We let the following assumptions, which are equivalent to Assumptions A1–A4 in [36] hold true.

**Assumption 1.** For  $(t_1, s_1), (t_2, s_2) \in \Delta[0, T]^2$ ,  $x \in \mathbb{R}^d$ ,  $y \in \mathbb{R}$  and  $z \in \mathbb{R}^\ell$ , there exists a constant  $K_1$  such that

$$\begin{aligned} |f(t_1, s_1, x, y, z) - f(t_2, s_2, x, y, z)| + |g(t_1, x) - g(t_2, x)| &\leq K_1(|t_1 - t_2|^{1/2} + |s_1 - s_2|^{1/2}), \\ |f(\cdot, \cdot, 0, 0, 0)| + |g(\cdot, 0)| &\leq K_1. \end{aligned}$$

Moreover,  $f$  and  $g$  have continuous and uniformly bounded first and second partial derivatives with respect to  $x$ ,  $y$  and  $z$ , and  $x$  (with boundary  $K_1$ ).

**Assumption 2.** For  $(t_1, s_1), (t_2, s_2) \in \Delta[0, T]^2$ , and  $x \in \mathbb{R}^d$ , there exists a constant  $K_2$  such that

$$\begin{aligned} |b(t_1, s_1, x) - b(t_2, s_2, x)| + |\sigma(t_1, s_1, x) - \sigma(t_2, s_2, x)| &\leq K_2(|t_1 - t_2|^{1/2} + |s_1 - s_2|^{1/2}) \\ |b(\cdot, \cdot, 0)| + |\sigma(\cdot, \cdot, 0)| &\leq K_2. \end{aligned}$$

Moreover,  $b$  and  $\sigma$  have continuous and uniformly bounded first and second partial derivatives with respect to  $x$  (with boundary  $K_2$ ).

**Assumption 3.**  $\varphi$  is an  $\mathbb{F}$ -adapted continuous process with  $\varphi(t) \in \mathbb{D}^{1,2}(\mathbb{R})$  for all  $t \in [0, T]$  and for  $t, t_1, t_2 \in [0, T]$  and  $(\theta_1, \theta_2) \in \Delta[0, T]^2$ , there exist constants  $p_0 > 2$  and  $K_3$  such that

$$\begin{aligned} \mathbb{E}|\varphi(t_1) - \varphi(t_2)|^2 &\leq K_3|t_1 - t_2|, \\ \mathbb{E}|D_{\theta_1}\varphi(t) - D_{\theta_2}\varphi(t)|^2 &\leq K_3|\theta_1 - \theta_2|, \\ \sup_{0 \leq \theta_1, \theta_2 \leq t \leq T} \mathbb{E}[|\varphi(t)|^{2p_0} + |D_{\theta_1}\varphi(t)|^{2p_0} + |D_{\theta_1}D_{\theta_2}\varphi(t)|^{p_0}] &\leq K_3^{2p_0}; \end{aligned}$$

where  $D_{\theta}\varphi(t)$  denotes the Malliavin derivative of the random variable  $\varphi(t)$  at time  $\theta$ .

The above assumptions are sufficient to guarantee the existence of a unique adapted solution to the FBSVIE (4), which is stated in the following theorem.

**Remark 2.1.** Assumption 3 is satisfied whenever  $\varphi$  is deterministic and Lipschitz continuous, or if  $\varphi(t) = \tilde{\phi}(\tilde{X}_t)$  for  $\tilde{\phi} \in C_b^4(\mathbb{R}^d; \mathbb{R})$  provided that  $\tilde{X}_t$  is an SDE with bounded and Lipschitz coefficients. The latter is a consequence of the chain rule and Lipschitz and Malliavin regularity results for SDEs. See, e.g., Theorem 2.2.1 in [28].

**Theorem 2.1.** Under assumptions 1-3, it holds that:

1. The FSVIE (4) admits a unique adapted solution  $X = (X_t)_{t \in [0, T]}$ ,
2. The BSVIE admits a unique solution  $(Y, Z) = (Y_s, Z_{t,s})_{(t,s) \in \Delta[0, T]^2}$  where  $Y$  is adapted and, for each fixed  $t$ , the map  $s \mapsto Z_{t,s}$  is  $\mathcal{F}_s$ -adapted on  $[t, T]$ .

*Proof.* For 1, we refer to classical results on well-posedness for FSVIE in [21] and 2 is exactly [36, Theorem 2.3].  $\square$

## 2.2 A coupled FSDE-BSVIE

In this subsection, the problem coefficients are given by  $x_0 \in \mathbb{R}^d$ ,  $b: [0, T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^\ell \rightarrow \mathbb{R}^d$ ,  $\sigma: [0, T] \times \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^{d \times \ell}$ ,  $g: [0, T] \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  and  $f: \Delta[0, T]^2 \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^\ell \times \mathbb{R}^\ell \rightarrow \mathbb{R}$ . We consider a coupled FSDE-BSVIE of the form

$$\begin{cases} X_t = x_0 + \int_0^t b(s, X_s, Y_s, Z_{s,s})ds + \int_0^t \sigma(s, X_s, Y_s)dW_s, \\ Y_t = g(t, X_t, X_T) + \int_t^T f(t, s, X_s, Y_s, Z_{t,s}, Z_{s,s})ds - \int_t^T Z_{t,s}^\top dW_s. \end{cases} \quad (5)$$

The coupled FSDE-BSVIE was first presented in [33] as a model for a time-inconsistent stochastic optimal control problem. In that work the authors showed that (5) can be rewritten as an HJB equation. They proved that, under suitable conditions, the HJB has a classical solution, which in turn guarantees that the FSDE-BSVIE has a unique adapted solution. Because we do not study this example in our error analysis, we omit the detailed conditions and refer the reader to [33] for full descriptions.

## 3 Variational formulations and temporal discretization

We first rewrite the problem as a variational problem that is continuous in time. This form is similar to the one used in the deep BSDE method [16], but is adjusted to BSVIEs. Next, we discretize the time interval to obtain a semi-discrete problem, which then serves as the starting point of our numerical method.

### 3.1 A time continuous variational formulation

We begin with the familiar variational formulation for a standard FBSDE. This example gives the reader a clear reference point before we extend the same ideas to the broader FBSVIE and FSDE-BSVIE cases that follow. Consider the following variational problem:

$$\begin{cases} \underset{y_0, \mathcal{Z}}{\text{minimize}} \mathbb{E}|Y_T^{y_0, \mathcal{Z}} - g(X_T^{y_0, \mathcal{Z}})|^2, & \text{where for } t \in [0, T], \\ X_t^{y_0, \mathcal{Z}} = x_0 + \int_0^t b(s, X_s^{y_0, \mathcal{Z}}, Y_s^{y_0, \mathcal{Z}}, \mathcal{Z}_s) ds + \int_0^t \sigma(s, X_s^{y_0, \mathcal{Z}}, Y_s^{y_0, \mathcal{Z}}, \mathcal{Z}_s) dW_s, \\ Y_t^{y_0, \mathcal{Z}} = y_0 - \int_0^t f(s, X_s^{y_0, \mathcal{Z}}, Y_s^{y_0, \mathcal{Z}}, \mathcal{Z}_s) ds + \int_0^t (Z_s^{y_0, \mathcal{Z}})^\top dW_s. \end{cases} \quad (6)$$

A solution to the FBSDE

$$\begin{cases} X_t = x_0 + \int_0^t b(s, X_s, Y_s, Z_s) ds + \int_0^t \sigma(s, X_s, Y_s, Z_s) dW_s, \\ Y_t = g(X_T) + \int_t^T f(s, X_s, Y_s, Z_s) ds - \int_t^T Z_s^\top dW_s, \end{cases} \quad (7)$$

clearly solves (6) and standard well-posedness conditions for the FBSDE guarantee that this solution is unique. Moreover, under suitable conditions, one further obtains the feedback forms  $y_0 = y_0(x_0)$  and  $\mathcal{Z}_t = \mathcal{Z}(t, X_t^{y_0, \mathcal{Z}})$ . This variational formulation is exactly what inspires the deep-BSDE method [17].

Just as an FBSDE can be written in a variational form, an FBSVIE or FSDE-BSVIE can as well. We seek processes  $Y$  and  $Z$  that satisfy the free-term dynamics. Choose coefficients  $\varphi, b, \sigma, g$ , and  $f$  such that the system is either the FBSVIE in (4) or the FSDE-BSVIE in (5). We assume that the chosen system has a unique adapted solution. With this assumption the equation is equivalent to the following variational problem:

$$\begin{cases} \underset{\mathcal{Y}, \mathcal{Z}}{\text{minimize}} \int_0^T \mathbb{E}|\mathcal{Y}_T^{\mathcal{Y}, \mathcal{Z}}(t) - g(t, X_t^{\mathcal{Y}, \mathcal{Z}}, X_T^{\mathcal{Y}, \mathcal{Z}})|^2 dt, & \text{where for } t \in [0, T], \\ X_t^{\mathcal{Y}, \mathcal{Z}} = \varphi(t) + \int_0^t b(t, s, X_s^{\mathcal{Y}, \mathcal{Z}}, \mathcal{Y}_s, \mathcal{Z}_{s,s}) ds + \int_0^t \sigma(t, s, X_s^{\mathcal{Y}, \mathcal{Z}}, \mathcal{Y}_s) dW_s, \\ \mathcal{Y}_T^{\mathcal{Y}, \mathcal{Z}}(t) = \mathcal{Y}_t - \int_t^T f(t, s, X_s^{\mathcal{Y}, \mathcal{Z}}, \mathcal{Y}_s, \mathcal{Z}_{t,s}, \mathcal{Z}_{s,s}) ds + \int_t^T \mathcal{Z}_{t,s}^\top dW_s. \end{cases} \quad (8)$$

Here  $\mathcal{Y}_T^{\mathcal{Y}, \mathcal{Z}}(t)$  is  $\mathcal{F}_T$ -measurable and can be viewed as the target from the perspective of the evaluation time  $t$ . Concretely, at time  $t$  one wants to find processes  $(\mathcal{Y}_s)_{s \in [t, T]}$  and  $(\mathcal{Z}_{t,s})_{s \in [t, T]}$  on the interval  $[t, T]$  such that

$$\mathcal{Y}_T^{\mathcal{Y}, \mathcal{Z}}(t) = g(t, X_t^{\mathcal{Y}, \mathcal{Z}}, X_T^{\mathcal{Y}, \mathcal{Z}}).$$

Then, at a later time  $t + u \leq T$ , one looks again for processes  $(\mathcal{Y}_s)_{s \in [t+u, T]}$  and  $(\mathcal{Z}_{t+u,s})_{s \in [t+u, T]}$  so that

$$\mathcal{Y}_T^{\mathcal{Y}, \mathcal{Z}}(t + u) = g(t + u, X_{t+u}^{\mathcal{Y}, \mathcal{Z}}, X_T^{\mathcal{Y}, \mathcal{Z}}).$$

Importantly, this should not be interpreted to mean that  $\mathcal{Y}_s$  and  $\mathcal{Z}_{t,s}$  for  $s \in [t, T]$  must be  $\mathcal{F}_t$ -measurable. Instead, it simply reflects the idea that, at each time  $t$ , one considers the relevant processes over the time interval  $[t, T]$  in order to satisfy the above relationship for  $\mathcal{Y}_T^{\mathcal{Y}, \mathcal{Z}}(t)$ .

In this setting, under suitable conditions, the processes  $Y$  and  $Z$  can be expressed in the feedback forms  $\mathcal{Y}_t = \mathcal{Y}(t, X_t^{\mathcal{Y}, \mathcal{Z}})$ , and  $\mathcal{Z}_{t,s} = \mathcal{Z}(t, s, X_t^{\mathcal{Y}, \mathcal{Z}}, X_s^{\mathcal{Y}, \mathcal{Z}})$ .

A solution to (5) clearly solves (8) since the objective function becomes identically zero. Furthermore, under appropriate conditions ensuring well-posedness of (5), this solution is unique.

### 3.2 A time discrete variational formulation

For some  $N \in \mathbb{N}$ , let  $\pi := \{0 = t_0 < t_1 < \dots < t_{N-1} < t_N = T\}$  be an equidistant grid with  $h = T/N$  and define  $\pi(t) := \inf\{k \mid t \in [t_k, t_{k+1})\}$  and  $\Delta^\pi[0, T]^2 := \{(\pi(t), \pi(s)) \mid (t, s) \in \Delta[0, T]^2\}$ .

For  $n \in \{0, 1, \dots, N-1\}$ , let  $\Delta W_n$  be a vector of  $\ell$  i.i.d. normally distributed random variables, each with mean 0 and variance  $h$ .

Below, we present the semi-discrete version of the variational problem (8) and, again, we assume that  $\varphi$ ,  $b$ ,  $\sigma$ ,  $f$  and  $g$  are chosen so that the system is either the FBSVIE in (4) or the FSDE-BSVIE in (5). This formulation employs the feedback forms for  $Y$  and  $Z$  introduced above, and reads

$$\left\{ \begin{array}{l} \text{minimize}_{\mathcal{Y}, \mathcal{Z}} \sum_{n=0}^{N-1} \mathbb{E} |\mathcal{Y}_N^{\pi, \mathcal{Y}, \mathcal{Z}}(n) - g(t_n, X_n^{\pi, \mathcal{Y}, \mathcal{Z}}, X_N^{\pi, \mathcal{Y}})|^2 h, \quad \text{where for } t \in [0, T], \\ X_n^{\mathcal{Y}, \mathcal{Z}} = \varphi(t) + \sum_{k=0}^{n-1} b(t_n, t_k, X_k^{\pi, \mathcal{Y}, \mathcal{Z}}, Y_k^{\pi, \mathcal{Y}, \mathcal{Z}}, Z_{k,k}^{\pi, \mathcal{Y}, \mathcal{Z}}) h \\ \quad + \sum_{k=0}^{n-1} \sigma(t_n, t_k, X_k^{\pi, \mathcal{Y}, \mathcal{Z}}, Y_k^{\pi, \mathcal{Y}, \mathcal{Z}}) \Delta W_k, \\ \mathcal{Y}_N^{\pi, \mathcal{Y}, \mathcal{Z}}(n) = Y_n^{\pi, \mathcal{Y}, \mathcal{Z}} - \sum_{k=n}^{N-1} f(t_n, t_k, X_k^{\pi, \mathcal{Y}, \mathcal{Z}}, Y_k^{\pi, \mathcal{Y}, \mathcal{Z}}, Z_{n,k}^{\pi, \mathcal{Y}, \mathcal{Z}}, Z_{k,k}^{\pi, \mathcal{Y}, \mathcal{Z}}) h + \sum_{k=n}^{N-1} (Z_{n,k}^{\pi, \mathcal{Y}, \mathcal{Z}})^\top \Delta W_k, \\ Y_n^{\pi, \mathcal{Y}, \mathcal{Z}} = \mathcal{Y}(t_n, X_n^{\pi, \mathcal{Y}, \mathcal{Z}}), \quad Z_{n,k}^{\pi, \mathcal{Y}, \mathcal{Z}} = \mathcal{Z}(t_n, t_k, X_n^{\pi, \mathcal{Y}, \mathcal{Z}}, X_k^{\pi, \mathcal{Y}, \mathcal{Z}}). \end{array} \right. \quad (9)$$

Note that when we consider a decoupled FBSVIE, we have assumed that  $g(t, x_t, x) = g(t, x)$ , and hence the feedback form for  $\mathcal{Z}$  reduces to  $\mathcal{Z}(t_n, t_k, X_k^{\pi, \mathcal{Y}, \mathcal{Z}})$ .

At this point the scheme is almost ready to implement. The final aspects to be cleared are

1. an approximation of the expectation,
2. explicit prescriptions for the functions  $\mathcal{Y}$  and  $\mathcal{Z}$ , and
3. a practical optimization routine.

**Expectation.** We assume that the expectation can be approximated to any desired accuracy. In practice we use Monte-Carlo simulation, whose mean-square error decreases at the classical rate  $O(M^{-1/2})$  when the sample size  $M$  grows.

**Functions  $\mathcal{Y}$  and  $\mathcal{Z}$ .** In a subsequent section we represent  $\mathcal{Y}$  and  $\mathcal{Z}$  with neural networks. For the error analysis in the next section, however, we keep the choice of function class completely open.

**Optimization.** The numerical experiments later in the paper employ mini-batch stochastic gradient descent with the Adam algorithm. The specific optimizer does not influence the error bounds developed in the following section, so its details are omitted for now.

## 4 Error analysis

In this section, we assume the setting of Section 2.1, *i.e.*, we let the randomness of the BSVIE stem from an FSVIE, which does not take the solution of a backward equation as inputs

$$X_t = \varphi(t) + \int_0^t b(t, s, X_s) ds + \int_0^t \sigma(t, s, X_s) dW_s. \quad (10)$$

The aim is to prove an a posteriori error bound for the BSVIE

$$Y_t = g(t, X_T) + \int_t^T f(t, s, X_s, Y_s, Z_{t,s}) ds - \int_t^T Z_{t,s}^\top dW_s. \quad (11)$$

We work under Assumptions 1-3; by Theorem 2.1 these conditions guarantee a unique adapted solution.

Below we introduce the Euler–Maruyama scheme for the FSVIE in (4) which is used throughout this section

$$X_n^\pi = \varphi(t_n) + \sum_{k=0}^{n-1} b(t_n, t_k, X_k^\pi)h + \sum_{k=0}^{n-1} \sigma(t_n, t_k, X_k^\pi)\Delta W_k, \quad n \in \{0, 1, \dots, N\}, \quad (12)$$

where the empty sum should be interpreted as zero. The following theorem states convergence and square integrability of (12).

**Theorem 4.1.** *Let Assumption 3 hold true. Then there exist constants  $C_x$  and  $K_x$  such that*

$$\sup_{t \in [0, T]} \mathbb{E}|X_t - X_{\pi(t)}^\pi|^2 \leq C_x h, \quad \max_{n \in \{0, 1, \dots, N\}} \mathbb{E}|X_n^\pi|^2 \leq K_x.$$

*Proof.* The proof of the first statement directly follows from the convergence at discretization points stated in [36, Theorem 3.1] and the path regularity of  $X$  stated in [36, Lemma 2.2]. The second statement follows from the square integrability of  $X_t$  stated in [36, Lemma 3.1] combined with the convergence of the Euler–Maruyama scheme.  $\square$

Note that the above Theorem holds true also in the special case when  $X$  is described by a standard FSDE.

Below we present a generic discretization scheme for a BSVIE of the form (11)

$$\begin{cases} \mathcal{Y}_N^\pi(n) = Y_n^\pi - \sum_{k=n}^{N-1} f(t_n, t_k, X_k^\pi, Y_k^\pi, Z_{n,k}^\pi)h + \sum_{k=n}^{N-1} (Z_{n,k}^\pi)^\top \Delta W_k, \\ Y_n^\pi = \mathcal{Y}(t_n, X_n^\pi), \quad Z_{n,k}^\pi = \mathcal{Z}(t_n, t_k, X_k^\pi), \quad (n, k) \in \Delta^\pi[0, T]^2. \end{cases} \quad (13)$$

The scheme above is of feedback form for the approximations of  $Y$  and  $Z$ . Moreover, it uses an Euler–Maruyama scheme to compute  $\mathcal{Y}_N^\pi(n)$ .

**Assumption 4.** *The functions  $\mathcal{Y}: [0, T] \times \mathbb{R}^d$  and  $\mathcal{Z}: \Delta[0, T]^2 \times \mathbb{R}^d \rightarrow \mathbb{R}^\ell$  satisfy a linear growth condition, i.e., for  $t_0 \in [0, T]$ ,  $(t, s) \in \Delta[0, T]^2$ ,  $x \in \mathbb{R}^d$  there exist constants  $K_Y$  and  $K_Z$  such that*

$$|\mathcal{Y}(t_0, x)|^2 \leq K_Y(1 + |x|^2), \quad |\mathcal{Z}(t, s, x)|^2 \leq K_Z(1 + |x|^2).$$

In the following, we present the main result of this section, which is an a posteriori error estimate for the above scheme.

**Theorem 4.2.** *Let Assumptions 1–4 hold true and suppose that  $f = f(t, s, x, y)$  or  $f = f(t, s, x, z)$ . Then for sufficiently small  $h$ , there exists a constant  $C$ , depending on  $T$  and  $K_1$ , such that*

$$\int_0^T \mathbb{E}|Y_t - Y_{\pi(t)}^\pi|^2 dt + \int_0^T \int_t^T \mathbb{E}|Z_{t,s} - Z_{\pi(t), \pi(s)}^\pi|^2 ds dt \leq C(h + \sum_{n=0}^{N-1} \mathbb{E}|\mathcal{Y}_N^\pi(n) - g(t_n, X_N^\pi)|^2 h).$$

To prove Theorem 4.2, we first present several intermediate results which are used in the final argument. The overall proof strategy is as follows:

1. **Approximation by coupled BSDEs.** We introduce a family of  $N$  coupled BSDEs that approximates the BSVIE (11) in the limit as  $h \rightarrow 0$  (equivalently  $N \rightarrow \infty$ ).
2. **Discretization of the BSDE family.** We present an explicit backward Euler–Maruyama scheme for these BSDEs which converges to the solution of the BSDE family.
3. **Stability estimate.** We derive a stability result that compares (i) the BSVIE scheme (13) and (ii) the discretized BSDE family (15). This gives a precise bound on the difference between the two schemes.
4. **Conclusion of the proof.** Finally, we combine the convergence of the BSDE scheme and the stability estimate to conclude that our BSVIE scheme converges, thereby establishing the error bounds in Theorem 4.2.

We note that Steps 1–2 in our outline are established directly via the main results of [36]. Meanwhile, Steps 3–4, are in spirit similar to the a posteriori error bounds for coupled FBSDE presented in [18].

In the remainder of this section, we introduce the coupled BSDEs in (14), recall key results from [36] regarding their numerical discretization, and prove the necessary stability estimates. At the end, we assemble all of these ingredients in a final proof of Theorem 4.2.

To approximate the BSVIE (11), we introduce a family of  $N$  coupled BSDEs, one for each  $k = 0, \dots, N-1$ . In the limit as  $N \rightarrow \infty$  (equivalently  $h \rightarrow 0$ ), these BSDEs collectively recover the BSVIE solution. Concretely, for  $0 \leq k \leq N-1$  and  $t \in [t_k, T]$ , we set

$$Y_t^k = g(t_k, X_T) + \int_t^T f(t_k, s, X_s, Y_s^{\pi(s)}, Z_s^k) ds - \int_t^T Z_s^k dW_s. \quad (14)$$

Note that this defines a system of  $N$  coupled BSDEs, where  $Y^k$  is defined for  $t \in [t_k, T]$ . For each  $n \in \{k+1, \dots, N-1\}$ , the driver of the  $k$ :th BSDE takes  $Y^n$  and  $Z^k$  as inputs in  $f$  on the interval  $[t_n, t_{n+1})$ . The following Theorem states that for each  $k$  (14) admits a unique adapted solution, and converges to (11) as the size of the temporal steps goes to zero.

**Theorem 4.3.** *Under assumptions 1-3, it holds that:*

1. *The family of BSDEs (14) admits a unique adapted solution  $(Y_s^{\pi(t)}, Z_s^{\pi(t)})_{(t,s) \in \Delta[0,T]^2}$ .*
2. *For each  $k \in \{0, 1, \dots, N-1\}$ , there exists a constant  $C_1$ , depending on  $T$  and  $K_1$ , such that*

$$\int_0^T \mathbb{E}|Y_t - Y_t^{\pi(t)}|^2 dt + \int_0^T \int_t^T \mathbb{E}|Z_{t,s} - Z_s^{\pi(t)}|^2 ds dt \leq C_1 h.$$

*Proof.* This Theorem states the results of Theorem 2.3 and Lemma 4.12 in [36]. □

The scheme below, proposed in [36], is an explicit backward type Euler–Maruyama scheme for the family of BSDEs (14).

$$\begin{cases} Y_n^{k,\pi} = \mathbb{E}[Y_{n+1}^{k,\pi} | \mathcal{F}_{t_n}] + f(t_k, t_n, X_n^\pi, Y_n^{n,\pi}, Z_n^{k,\pi})h, \\ Z_n^{k,\pi} = \frac{1}{h} \mathbb{E}[\Delta W_n Y_{n+1}^{k,\pi} | \mathcal{F}_{t_n}], \\ Y_N^{k,\pi} = g(t_k, X_N^\pi), \quad (n, k) \in \Delta^\pi[0, T]^2. \end{cases} \quad (15)$$

The following theorem, which combines results from [36], states that the scheme (15) converges to the solution of the BSDE family (14).

**Theorem 4.4.** *Let Assumption 1-3 hold true and suppose that  $f = f(t, s, x, y)$  or  $f = f(t, s, x, z)$ . Then, for each  $k \in \{0, 1, \dots, N-1\}$ , there exists a constant  $C_2$ , depending on  $T$  and  $K_1$ , such that*

$$\int_{t_k}^{t_{k+1}} \mathbb{E}|Y_t^k - Y_{\pi(t)}^{k,\pi}|^2 dt + h \int_{t_k}^T \mathbb{E}|Z_t^k - Z_{\pi(t)}^{k,\pi}|^2 dt \leq C_2 h^2.$$

*Proof.* This is a direct application of [36, Lemma 4.5 and Lemma 4.12]. □

**Remark 4.1.** *The scheme in [36] uses  $Y_{n+1}^{n,\pi}$  rather than  $Y_n^{n,\pi}$  in the driver. This makes the scheme implicit in the equation for  $\bar{Z}_n^{k,\pi}$  since  $f(t_k, t_n, X_n^\pi, Y_{n+1}^{n,\pi}, Z_n^{k,\pi})$  is  $\mathcal{F}_{t_{n+1}}$  measurable (rather than  $\mathcal{F}_{t_n}$ -measurable) which yields  $Z_n^{k,\pi} = \frac{1}{h} \mathbb{E}[\Delta W_n Y_{n+1}^{k,\pi} + f(t_k, t_n, X_n^\pi, Y_{n+1}^{n,\pi}, Z_n^{k,\pi}) \Delta W_n | \mathcal{F}_{t_n}]$ . Nevertheless, the same techniques apply, and thus one can prove the same error bound for the explicit scheme considered here. Alternatively, one can use the more general scheme for type-II BSVIEs proposed in [15], which when applied to type-I BSVIEs coincides with (15).*

To bound the difference between our BSVIE scheme (13) and the scheme for the BSDE family (15), we introduce the following general scheme

$$\widehat{Y}_{n+1}^{k,\pi} = \widehat{Y}_n^{k,\pi} - f(t_k, t_n, X_n^\pi, \widehat{Y}_n^{n,\pi}, \widehat{Z}_n^{k,\pi})h + (\widehat{Z}_n^{k,\pi})^\top \Delta W_k, \quad (k, n) \in \Delta^\pi[0, T]^2. \quad (16)$$

Because no initial or terminal condition is imposed, the general scheme above admits infinitely many solutions. Furthermore, we observe that  $\widehat{Y}_n^{k,\pi} = \mathbb{E}[\widehat{Y}_{n+1}^{k,\pi} | \mathcal{F}_{t_n}] + f(t_k, t_n, X_n^\pi, \widehat{Y}_n^{n,\pi}, \widehat{Z}_n^{k,\pi})h$ , and  $\widehat{Z}_n^{k,\pi} = \frac{1}{h} \mathbb{E}[\Delta W_n \widehat{Y}_{n+1}^{k,\pi} | \mathcal{F}_{t_n}]$ , implying that (15) is a special case of (16). It is, in fact, also possible to express our BSVIE scheme via the generic scheme (16). To illustrate this, we introduce the following notation

$$\begin{cases} \mathcal{Y}_k^\pi(n+1) = \mathcal{Y}_k^\pi(n) - f(t_k, t_n, X_n^\pi, \mathcal{Y}_n^\pi(n), \mathcal{Z}_n^\pi(k))h + (\mathcal{Z}_k^\pi(n))^\top \Delta W_n, \\ \mathcal{Y}_n^\pi(n) = \mathcal{Y}(t_n, X_n^\pi), \quad \mathcal{Z}_k^\pi(n) = \mathcal{Z}(t_k, t_n, X_n^\pi), \quad (n, k) \in \Delta^\pi[0, T]^2. \end{cases} \quad (17)$$

Note that the above is equivalent to (13), with  $\mathcal{Y}_k^\pi(k) = Y_k^\pi$ . We emphasize that the scheme presented above is included purely for illustration and is not intended for practical use in this form.

The following lemma provides a stability estimate between two solutions to (16).

**Lemma 4.1.** *Let Assumptions 1-3 hold true and assume that  $h$  is small enough so that  $(2K_1 + \frac{1}{2})h < 1$ . For  $j \in \{1, 2\}$ , suppose  $\{\widehat{Y}_n^{k,\pi,j}, \widehat{Z}_n^{k,\pi,j}\}_{(k,n) \in \Delta^\pi[0, T]^2}$  are two square integrable solutions to (16). Define the differences*

$$\delta Y_n^k = \widehat{Y}_n^{k,\pi,1} - \widehat{Y}_n^{k,\pi,2}, \quad \delta Z_n^k = \widehat{Z}_n^{k,\pi,1} - \widehat{Z}_n^{k,\pi,2}.$$

Then there exist constants  $C_Y$  and  $C_Z$ , depending only on  $T$  and  $K_1$ , such that for every  $(k, n) \in \Delta^\pi[0, T]^2$ , the following estimates hold:

$$\begin{aligned} \mathbb{E}|\delta Y_n^k|^2 &\leq C_Y \left( \mathbb{E}|\delta Y_N^k|^2 + \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h \right), \\ \mathbb{E}|\delta Z_n^k|^2 h &\leq C_Z \left( \mathbb{E}|\delta Y_N^k|^2 + \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h \right). \end{aligned}$$

*Proof.* We first derive a discrete Grönwall inequality for the error terms  $\mathbb{E}|\delta Y_n^k|^2$  and  $\mathbb{E}|\delta Z_n^k|^2$ . Inspecting the recursion shows that, once the time grid is fine enough, i.e., for  $N \geq (2K_1 + \frac{1}{2})T$ , where  $K_1$  is the Lipschitz-growth constant in Assumption 1 and  $T$  is the time horizon, the associated constants are monotone decreasing in  $N$  and therefore uniformly bounded. Because the  $Z$ -process carries two time indices, we treat the cases  $n > k$  and  $n = k$  separately before combining the estimates. Full algebraic details are given in Appendix A.  $\square$

We are now ready to prove Theorem 4.2.

*Proof of Theorem 4.2.* We decompose the overall error into three main contributions:

- (i) the approximation error arising from replacing the original BSVIE with a family of BSDEs,
- (ii) the discretization error incurred when approximating the BSDE family with a backward Euler–Maruyama scheme,
- (iii) the error due to the difference between the backward discretization scheme for the BSDE family and our BSVIE discretization scheme.

Denote these contributions by  $\text{Err}_1(h)$ ,  $\text{Err}_2(h)$ , and  $\text{Err}_3(h)$ , respectively. Then we have

$$\int_0^T \mathbb{E}|Y_t - Y_{\pi(t)}^\pi|^2 dt + \int_0^T \int_t^T \mathbb{E}|Z_{t,s} - Z_{\pi(t),\pi(s)}^\pi|^2 ds dt \leq \text{Err}_1(h) + \text{Err}_2(h) + \text{Err}_3(h),$$

where

$$\begin{aligned}\text{Err}_1(h) &= \int_0^T \mathbb{E}|Y_t - Y_t^{\pi(t)}|^2 dt + \int_0^T \int_t^T \mathbb{E}|Z_{t,s} - Z_s^{\pi(t)}|^2 ds dt, \\ \text{Err}_2(h) &= \int_0^T \mathbb{E}|Y_t^{\pi(t)} - Y_{\pi(t)}^{\pi(t),\pi}|^2 dt + \int_0^T \int_t^T \mathbb{E}|Z_s^{\pi(t)} - Z_{\pi(t),\pi(s)}^\pi|^2 ds dt, \\ \text{Err}_3(h) &= \int_0^T \mathbb{E}|Y_{\pi(t)}^{\pi(t),\pi} - Y_{\pi(t)}^\pi|^2 dt + \int_0^T \int_t^T \mathbb{E}|Z_{\pi(t)}^{\pi(s),\pi} - Z_{\pi(t),\pi(s)}^\pi|^2 ds dt.\end{aligned}$$

From Theorem 4.3, we have  $\text{Err}_1(h) \leq C_1 h$ . A slight re-write of  $\text{Err}_2(h)$  and applying Theorem 4.4 yield

$$\text{Err}_2(h) = \sum_{k=0}^{N-1} \left( \int_{t_k}^{t_{k+1}} \mathbb{E}|Y_t^{\pi(t)} - Y_{\pi(t)}^{\pi(t),\pi}|^2 dt + h \int_{t_k}^T \mathbb{E}|Z_s^{\pi(t)} - Z_{\pi(t),\pi(s)}^\pi|^2 ds \right) \leq C_2 h.$$

For  $\text{Err}_3(h)$ , we note that

$$\text{Err}_3(h) = \sum_{n=0}^{N-1} \left( \mathbb{E}|Y_n^{n,\pi} - Y_n^\pi|^2 h + h \sum_{k=n}^{N-1} \mathbb{E}|Z_n^{k,\pi} - Z_{k,n}^\pi|^2 ds \right),$$

to which we want to apply Lemma 4.1. Define the two discrete processes as follows. First, set  $\{Y_n^{n,\pi,1}, Z_n^{k,\pi,1}\}_{(k,n) \in \Delta^\pi[0,T]^2} = \{Y_n^{n,\pi}, Z_n^{k,\pi}\}_{(k,n) \in \Delta^\pi[0,T]^2}$ , where  $\{Y_n^{n,\pi}, Z_n^{k,\pi}\}$  is produced by the backward Euler–Maruyama scheme (15). Next, set  $\{Y_n^{n,\pi,2}, Z_n^{k,\pi,2}\}_{(k,n) \in \Delta^\pi[0,T]^2} = \{\mathcal{Y}_n^\pi(n), \mathcal{Z}_n^\pi(k)\}_{(k,n) \in \Delta^\pi[0,T]^2}$ , where  $\{\mathcal{Y}_n^\pi(n), \mathcal{Z}_n^\pi(k)\}$  is obtained from the discretization scheme for the BSVIE (13), using the notation in (17). Then

$$\begin{aligned}\text{Err}_3(h) &\leq C_Y \sum_{n=0}^{N-1} \left( \mathbb{E}|\delta Y_N^n|^2 + \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h \right) + C_Z \sum_{n=0}^{N-1} h \sum_{k=n}^{N-1} \left( \mathbb{E}|\delta Y_N^k|^2 + \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h \right) \\ &\leq (C_Y + TC_Z)(1+T) \sum_{n=0}^{N-1} \mathbb{E}|\delta Y_N^n|^2 \\ &= (C_Y + TC_Z)(1+T) \sum_{n=0}^{N-1} \mathbb{E}|\mathcal{Y}_N^\pi(n) - g(t_n, X_N^\pi)|^2.\end{aligned}$$

Combining the estimates for  $\text{Err}_1(h)$ ,  $\text{Err}_2(h)$ , and  $\text{Err}_3(h)$ , we obtain the overall error bound, which completes the proof.  $\square$

## 5 Fully implementable scheme and neural network details

In this section, we present a fully discretized problem formulation and introduce neural networks as function approximators.

### 5.1 Fully implementable algorithms

Without further specifications, (9) assumes exact optimization over an unspecified set of functions  $\mathcal{Y}$  and  $\mathcal{Z}$  and the exact computation of expectations. To define a fully implementable scheme, we introduce the parametric functions  $\mathcal{Y}^\theta: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$  and  $\mathcal{Z}^\theta: \Delta[0, T]^2 \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^\ell$ . Here  $\mathcal{Y}^{\theta_y}$  and  $\mathcal{Z}^{\theta_z}$  are neural networks and  $\theta = (\theta_y, \theta_z)$  represents all trainable parameters. We assume that  $\theta$  takes values in some parameter space  $\Theta$ . Moreover, expectations are approximated with batch Monte-Carlo simulation. Let  $K_{\text{epochs}} \geq 1, K_{\text{batch}} \in \mathbb{N}$  be the number of epochs and the number of batches per epoch, respectively. Let further  $M_{\text{train}}, M_{\text{batch}} \in \mathbb{N}$  be the size of the training data set and batch, respectively. We assume that  $M_{\text{train}}/M_{\text{batch}} = K_{\text{batch}} \in \mathbb{N}$ . Training data are  $M_{\text{train}}$  independent realizations of the Wiener increments  $\Delta W_0, \dots, \Delta W_{N-1} \sim \mathcal{N}(0, h)$  and the training data are reused

in  $K_{\text{epoch}}$  epochs. The training is initialized by random sampling of  $\theta^0 \in \Theta$ . For each update step in an epoch of the training algorithm, we take  $M_{\text{batch}}$  Wiener increments  $\Delta W_0(m), \dots, \Delta W_{N-1}(m)$ ,  $m = 1, 2, \dots, M_{\text{batch}}$  from the training data set that were not previously used during the epoch and update  $\theta$  by approximate optimization of the following problem:

$$\left\{ \begin{array}{l} \text{minimize}_{\mathcal{Y}, \mathcal{Z}} \sum_{n=0}^{N-1} \frac{1}{M_{\text{batch}}} \sum_{m=1}^{M_{\text{batch}}} |\mathcal{Y}_N^{\pi, \theta}(n)(m) - g(t_n, X_n^{\pi, \theta}(m), X_N^{\pi, \theta}(m))|^2 h, \\ \text{For } m = 1, \dots, M_{\text{batch}} : \\ X_n^{\pi, \theta}(m) = \varphi(t) + \sum_{k=0}^{n-1} b(t_n, t_k, X_k^{\pi, \theta}(m), Y_k^{\pi, \theta}(m), Z_{k,k}^{\pi, \theta}(m))h, \\ \quad + \sum_{k=0}^{n-1} \sigma(t_n, t_k, X_k^{\pi, \theta}(m), Y_k^{\pi, \theta}(m)) \Delta W_k(m), \\ \mathcal{Y}_N^{\pi, \theta}(n)(m) = Y_n^{\pi, \theta}(m) - \sum_{k=n}^{N-1} f(t_n, t_k, X_k^{\pi, \theta}(m), Y_k^{\pi, \theta}(m), Z_{n,k}^{\pi, \theta}(m), Z_{k,k}^{\pi, \theta}(m))h \\ \quad + \sum_{k=n}^{N-1} (Z_{n,k}^{\pi, \theta}(m))^\top \Delta W_k(m), \\ Y_n^{\pi, \theta}(m) = \mathcal{Y}^\theta(t_n, X_n^{\pi, \theta}(m)), \quad Z_{n,k}^{\pi, \theta}(m) = \mathcal{Z}(t_n, t_k, X_n^{\pi, \theta}(m), X_k^{\pi, \theta}(m)), \end{array} \right. \quad (18)$$

When all training data has been used, a new epoch starts. When the  $K_{\text{epoch}}$ :th epoch is finished, the algorithm terminates. The neural network parameters at termination are  $\theta^*$ . It is an approximation of the parameters  $\theta^{**}$  that optimize (18) in the limit  $M_{\text{batch}} \rightarrow \infty$ . Hence,  $Y^{\pi^{**}, \theta}$  and  $Z^{\pi^{**}, \theta}$  and the solution to the discrete variational problem (9) are controlled by the representation error. In turn, if the class of neural networks is dense in the function space we optimize over in (9), then the representation error vanishes.

## 5.2 Specification of the neural networks

Here, we introduce the neural networks that we use in our implementations in Section 5.1. The generality is kept to a minimum and more general architectures are of course possible. For  $\mathcal{Y}^\theta: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$  and  $\mathcal{Z}^\theta: \Delta[0, T]^2 \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^\ell$ , we employ fully-connected, feed-forward networks with three hidden layers; because the input dimension of  $\mathcal{Z}^\theta$  is larger than that of  $\mathcal{Y}^\theta$ , we use 100 neurons in each hidden layer for  $\mathcal{Z}^\theta$  and 50 neurons in each hidden layer for  $\mathcal{Y}^\theta$ . In both architectures, each affine transformation in the hidden layers is followed by the element-wise ReLU activation function,  $\mathfrak{R}(x) = \max(0, x)$ , while the output layer remains unactivated. More precisely, for  $x, x_t \in \mathbb{R}^d$  and  $(t, s) \in \Delta[0, T]^2$  denote by  $\mathbf{x}_y = \text{Concat}(t, x) \in \mathbb{R}^{d+1}$  and  $\mathbf{x}_z = \text{Concat}(t, s, x_t, x) \in \mathbb{R}^{2d+2}$

$$\begin{aligned} \mathcal{Y}^{\theta_y}(\mathbf{x}_y) &= \mathcal{W}_y^4 \mathfrak{R}(\mathcal{W}_y^3 \mathfrak{R}(\mathcal{W}_y^2 \mathfrak{R}(\mathcal{W}_y^1 \mathbf{x}_y + b_y^1) + b_y^2) + b_y^3) + b_y^4, \\ \mathcal{Z}^{\theta_z}(\mathbf{x}_z) &= \mathcal{W}_z^4 \mathfrak{R}(\mathcal{W}_z^3 \mathfrak{R}(\mathcal{W}_z^2 \mathfrak{R}(\mathcal{W}_z^1 \mathbf{x}_z + b_z^1) + b_z^2) + b_z^3) + b_z^4, \end{aligned}$$

with weight matrices  $\mathcal{W}_y^1 \in \mathbb{R}^{50 \times d+1}$ ,  $\mathcal{W}_y^2, \mathcal{W}_y^3 \in \mathbb{R}^{50 \times 50}$ ,  $\mathcal{W}_y^4 \in \mathbb{R}^{1 \times 50}$ , and  $\mathcal{W}_z^1 \in \mathbb{R}^{100 \times 2d+2}$ ,  $\mathcal{W}_z^2, \mathcal{W}_z^3 \in \mathbb{R}^{100 \times 100}$ ,  $\mathcal{W}_z^4 \in \mathbb{R}^{\ell \times 100}$  and bias vectors  $b_y^1, b_y^2, b_y^3 \in \mathbb{R}^{50}$ ,  $b_y^4 \in \mathbb{R}$ , and  $b_z^1, b_z^2, b_z^3 \in \mathbb{R}^{100}$ ,  $b_z^4 \in \mathbb{R}^\ell$ . Finally, denote by  $\theta_y = (\mathcal{W}_y^1, \mathcal{W}_y^2, \mathcal{W}_y^3, \mathcal{W}_y^4, b_y^1, b_y^2, b_y^3, b_y^4)$ ,  $\theta_z = (\mathcal{W}_z^1, \mathcal{W}_z^2, \mathcal{W}_z^3, \mathcal{W}_z^4, b_z^1, b_z^2, b_z^3, b_z^4)$ , and  $\theta = (\theta_y, \theta_z)$  where the matrices are considered vectorized before concatenation.

## 6 Numerical experiments

This section is divided into two parts. Section 6.1 focuses on test problems that satisfy the assumptions of Section 4, whereas Section 6.2 relaxes those assumptions and investigates coupled FBSVIEs.

Throughout all experiments we adopt the same hyper-parameter configuration. The mini-batch size is fixed at  $M_{\text{batch}} = 2^{11}$ , and the total number of training paths at  $M_{\text{train}} = 2^{18}$ . Each path is therefore processed ten times, giving  $K_{\text{epoch}} = 10$  training epochs with random shuffling between epochs. Both the  $Y$ - and  $Z$ -networks contain three hidden layers (input and output layers excluded); the  $Y$ -network has 50 neurons per layer, while the  $Z$ -network has 100, due to one extra temporal dimension for the  $Z$ -network. The learning rate is initialized at 0.005 and follows an exponential decay schedule, being multiplied by  $e^{-0.2}$  after every epoch. Optimization is carried out with the Adam algorithm [25]. Further implementation details can be found at <https://github.com/AlessandroGnoatto/DeepBSVIE>.

Although we report results with a single feed-forward architecture for all problems, during development we experimented with a wide range of hyper-parameters: between 1 and 6 hidden layers, 10–300 neurons per layer, ReLU versus tanh activations, and both Adam and SGD optimizers. Once the network had sufficient capacity, the solver’s accuracy and convergence rate were essentially insensitive to the exact depth, width, or activation choice. We therefore settled on a network with three hidden layers and 100 neurons per layer, oversized for the low-dimensional examples yet still tractable in up to 500 dimensions, which performed on par with both slimmer and deeper alternatives. This configuration is thus reported as the smallest architecture that remained robust across all test cases.

## 6.1 Examples where the error analysis apply

In this section, we consider two coupled FSDE–BSVIE systems in which both the driver  $f$  and the free term  $g$  depend explicitly on the time variable  $t$ . That is, we study systems of the form

$$\begin{cases} X_t = x_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s, \\ Y_t = g(t, X_T) + \int_t^T f(t, s, X_s, Y_s, Z_{t,s}) ds - \int_t^T Z_{t,s}^\top dW_s, \quad (t, s) \in \Delta[0, T]^2. \end{cases} \quad (19)$$

Assuming that the coefficients satisfy the required regularity conditions, our numerical analysis is applicable to this class of systems.

We consider two examples: one in which the FSDE is driven by additive noise, and another in which it is driven by multiplicative noise.

### 6.1.1 Example 1A: Additive noise

Let  $d \in \mathbb{N}$ ,  $k \in \mathbb{R}$ ,  $\mu, x_0 \in \mathbb{R}^d$  and  $\sigma \in \mathbb{R}^{d \times d}$  be constants with  $\sigma$  invertible. We consider the following system

$$\begin{cases} X_t = x_0 + \mu t + \sigma W_t, \quad \bar{X}_t = \frac{1}{d} \sum_{i=1}^d X_t^i, \\ Y_t = t \sin(k \bar{X}_T) + \int_t^T \left( \frac{tk^2}{2d^2} \sin(k \bar{X}_s) \|\sigma\|^2 - \mu^\top \sigma^{-1} Z_{t,s} \right) ds - \int_t^T Z_{t,s}^\top dW_s, \quad (t, s) \in \Delta[0, T]^2. \end{cases} \quad (20)$$

A direct calculation shows that the unique solution  $\{(Y_t, Z_{t,s})\}_{t \leq s \leq T}$  is given by the closed-form expressions

$$Y_t = t \sin(k \bar{X}_t) \quad \text{and} \quad Z_{t,s} = \frac{tk}{d} \cos(k \bar{X}_s) \sigma \mathbf{1}_d, \quad (t, s) \in \Delta[0, T]^2. \quad (21)$$

Here,  $\mathbf{1}_d$  is a  $d$ -dimensional vector consisting of ones.

Let  $k = 5$ ,  $d = 5$ ,  $T = 1$ ,  $x_0 = (1, 1, 1, 1, 1)^\top$ ,  $\mu = \text{diag}(0.25, 0.25, 0.25, 0.25, 0.25)$ , and  $\sigma = \text{diag}(0.8, 0.9, 1.0, 1.1, 1.2)$ .

Figure 1 displays the approximate  $Y$ -process alongside the analytical reference solution. In the left frame, three representative sample paths are shown, while the right frame presents the sample mean together with the 25th and 75th sample percentiles.

Figure 2 displays the first component of the  $Z_{t,\cdot}$ -process for different values of  $t$  compared with the analytical reference solutions for Example 1A. In the left frame, one representative sample path is shown, while the right frame presents the sample mean.

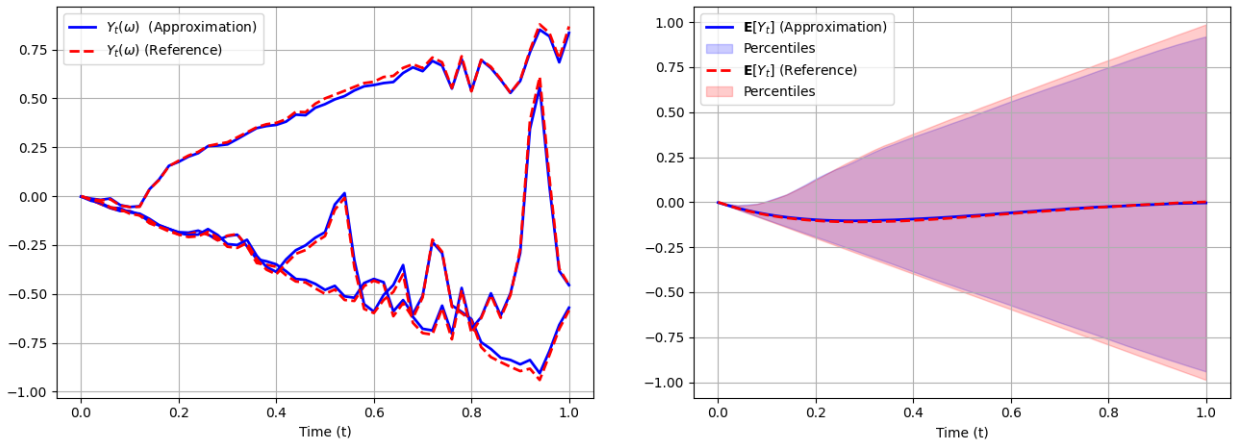


Figure 1: Comparison of the approximated  $Y$  with the reference solutions for Example 1A. **Left:** Three representative sample paths. **Right:** The sample mean and the 25th and 75th percentiles.

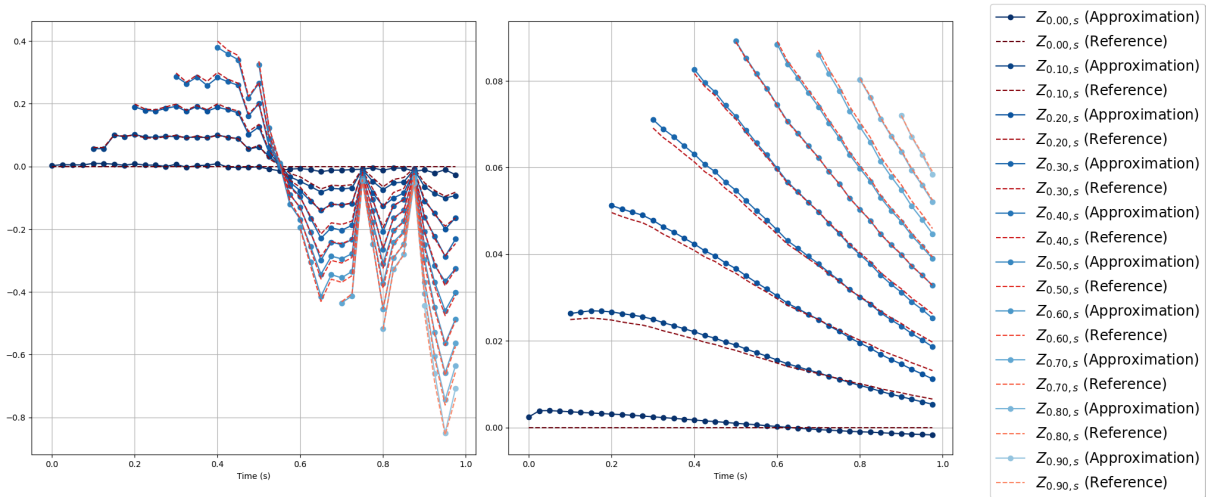


Figure 2: Comparison of the approximated  $Z_{t,s}$  with the reference solution for different values of  $t$  for Example 1A. **Left:** One representative sample path of the first (of 5) component of  $Z_{t,s}$ . **Right:** A sample mean for the first component of  $Z_{t,s}$ .

### 6.1.2 Example 1B: Multiplicative noise

Let  $d \in \mathbb{N}$ ,  $k \in \mathbb{R}$ ,  $\mu, x_0 \in \mathbb{R}^d$  and  $\sigma \in \mathbb{R}^{d \times d}$  be constants with  $\sigma$  invertible. We consider the following system

$$\begin{cases} X_t = x_0 + \int_0^t \text{diag}(\mu) X_s ds + \int_0^t \text{diag}(\sigma X_s) dW_s, & \bar{X}_t = \frac{1}{d} \sum_{i=1}^d X_t^i, \\ Y_t = t \sin(k \bar{X}_T) + \int_t^T \left( \frac{tk^2}{2d^2} \sin(k \bar{X}_s) \|\sigma X_s\|^2 - \mu^\top \sigma^{-1} Z_{t,s} \right) ds - \int_t^T Z_{t,s}^\top dW_s, & (t, s) \in \Delta[0, T]^2. \end{cases} \quad (22)$$

A direct calculation shows that the unique solution  $\{(Y_t, Z_{t,s})\}_{(t,s) \in \Delta[0,T]^2}$  is given by the closed-form expressions

$$Y_t = t \sin(k \bar{X}_t) \quad \text{and} \quad Z_{t,s} = \frac{tk}{d} \cos(k \bar{X}_s) \sigma X_s, \quad (t, s) \in \Delta[0, T]^2. \quad (23)$$

Let  $d = 5$ ,  $T = 1$ ,  $x_0 = (1, 1, 1, 1, 1)^\top$ ,  $\mu = \text{diag}(0.05, 0.05, 0.05, 0.05, 0.05)$ , and  $\sigma = \text{diag}(0.2, 0.25, 0.3, 0.35, 0.45)$ . Figure 3 displays the approximate  $Y$ -process alongside the analytical reference solution. In the left frame, three representative sample paths are shown, while the right frame presents the sample mean together with the 5th and 95th sample percentiles. Due to the difference in variance between these examples, different percentile ranges were chosen. In the first example, where the variance is higher, the 25th and 75th percentiles provide a clear view of the central tendency. In the second example, with lower variance, it is more illustrative to use the 5th and 95th percentiles to capture a broader perspective of the distribution.

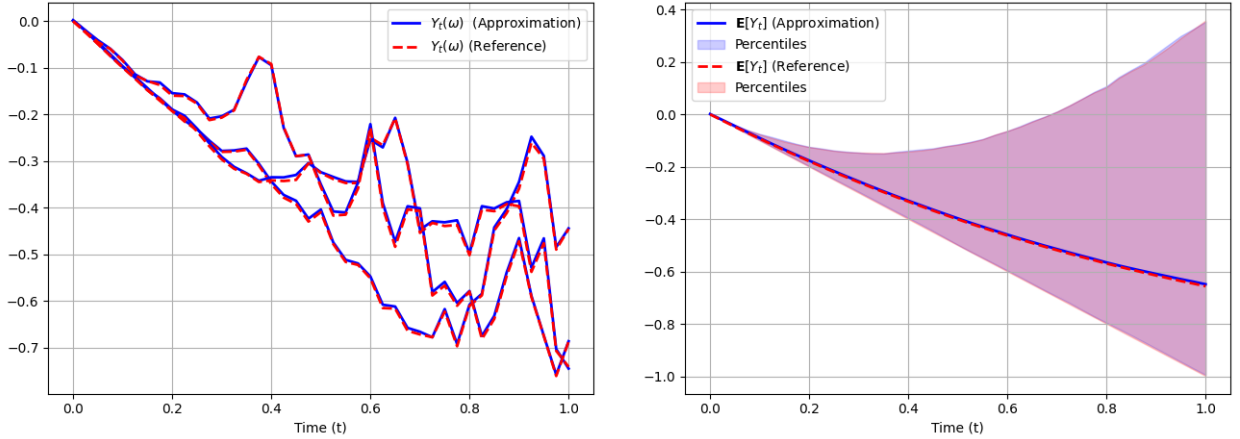


Figure 3: Comparison of the approximated  $Y$  with the reference solutions for Example 1B. **Left:** Three representative sample paths. **Right:** The sample mean and the 5th and 95th percentiles.

Figure 4 displays the first component of the  $Z_{t,\cdot}$ -process for different values of  $t$  compared with the analytical reference solutions. In the left frame, one representative sample path is shown, while the right frame presents the sample mean.

### 6.1.3 Example 1: Empirical error analysis

Figure 5 displays the empirical convergence of our approximation in terms of the stepsize  $h$  (or the number of discretization steps  $N$ ). We choose  $N \in \{10, 20, 30, 40, 50\}$  for Example 1A and  $N \in \{10, 20, 30, 40\}$  for Example 1B, where the variance of the solution is lower. Note that in Example 1A, where the FSDE is an arithmetic Brownian motion, the Euler–Maruyama scheme coincides with the exact solution. This implies that the entire discretization error is attributable to the discretization of the BSVIE. For Example 1B, the FSDE is a geometric Brownian motion, for which the Euler–Maruyama scheme has a strong discretization error of order 0.5. Moreover, we have access to a closed-form solution for the geometric Brownian motion, which is used when the reference solution is computed (while the Euler–Maruyama scheme is employed for the FSDE in our approximate solution).

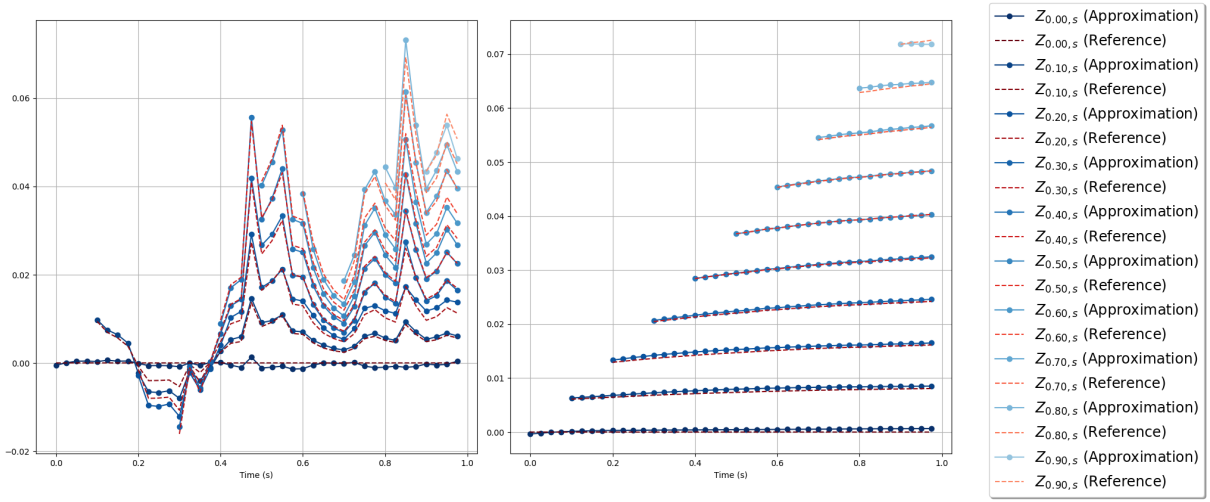


Figure 4: Comparison of the approximated  $Z_{t,s}$  with the reference solution for different values of  $t$  for Example 1B. **Left:** One representative sample path of the first (of 5) component of  $Z_{t,s}$ . **Right:** A sample mean the first component of  $Z_{t,s}$ .

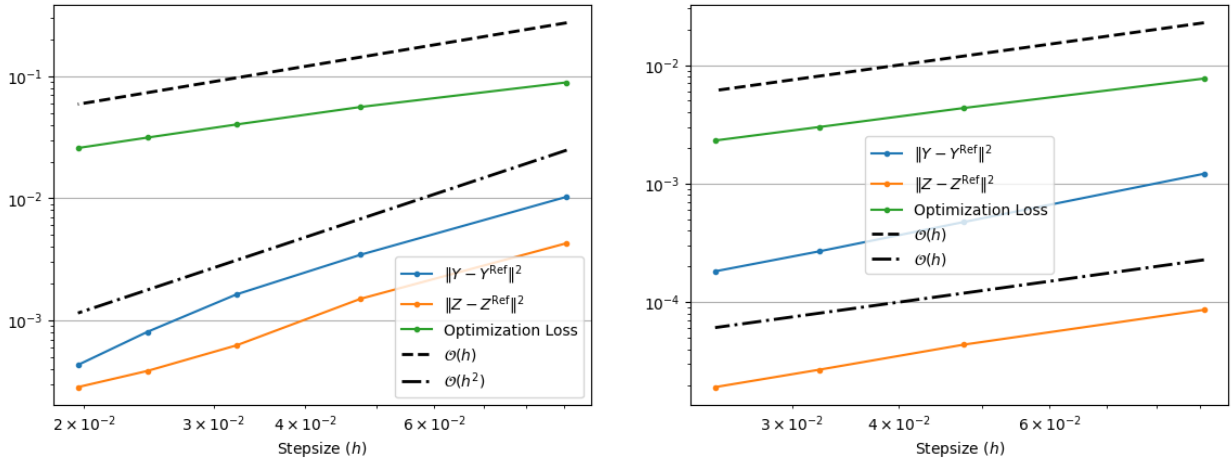


Figure 5: Empirical convergence plot for our approximate  $Y$ ,  $Z$  and the loss function. **Left:** Example 1A. **Right:** Example 1B.

In both Example 1A and Example 1B, we observe an empirical convergence order of 1 for the optimization loss. However, for  $Y$  and  $Z$ , the convergence order is 1 in Example 1A and 0.5 in Example 1B. This suggests that employing a higher-order approximation for the FSDE can improve the overall convergence order achieved for the BSVIE.

#### 6.1.4 Example 1: CPU time and scalability in the spatial dimension

In this subsection, we study how the method scales with the spatial dimension  $d$  in terms of both accuracy and wall-clock runtime. As a test case, we consider the BSVIE (22). For several values of  $d$  we report the wall-clock runtime and the simulation errors defined as

$$\begin{aligned}\text{Err}(Y) &:= \frac{1}{M} \sum_{m=1}^M \sum_{n=0}^{N-1} \left| Y_{t_n}(m) - Y_n^{\pi, \theta^*}(m) \right|^2 h, \\ \text{Err}(Z) &:= \frac{1}{M} \sum_{m=1}^M \sum_{k=0}^{N-1} \sum_{n=k}^{N-1} \left| Z_{t_k, t_n}(m) - Z_{k,n}^{\pi, \theta^*}(m) \right|^2 h^2.\end{aligned}$$

Table 1: Numerical results for various dimensions  $d$ .

d	Err( $Y$ )	Err( $Z$ )	Runtime (s)
1	$7.7 \times 10^{-5}$	$8.2 \times 10^{-5}$	430
5	$9.6 \times 10^{-5}$	$2.9 \times 10^{-5}$	480
20	$1.8 \times 10^{-4}$	$7.9 \times 10^{-6}$	480
100	$1.3 \times 10^{-3}$	$4.5 \times 10^{-6}$	480
500	$8.0 \times 10^{-3}$	$2.9 \times 10^{-6}$	480

The algorithm runs on a Google Colab instance equipped with an NVIDIA A100 GPU. Each training batch launches hundreds of small GPU kernels and transfers pseudo-random numbers from the CPU to the device; this fixed “administrative” latency dwarfs the arithmetic itself, so processing a 500-dimensional state vector takes almost the same wall-clock runtime as a 1-dimensional one. The method therefore scales extremely well in the state dimension  $d$ . Finally, most of the time is spent in these per-step launches and transfers, not in mathematics, so rewriting the loop in a more vectorised, GPU-friendly way should yield a much faster algorithm that would retain the same favorable scaling in  $d$ .

## 6.2 Examples of general FSDE-BSVIE systems

In this section, we treat more general forms of FSDE-BSVIEs, where the numerical analysis, in the form given in this paper, no longer applies.

### 6.2.1 Example 2: An FSDE-BSVIE system with a quadratic solution

Let  $\mu, x_0 \in \mathbb{R}^d$  and  $\sigma \in \mathbb{R}^{d \times d}$  be constant, with  $\sigma$  invertible. We consider the following system

$$\begin{cases} X_t = x_0 + \int_0^t \mu X_s ds + \int_0^t \text{diag}(\sigma X_s) dW_s, \\ Y_t = \langle t + X_t, X_T \rangle - \int_t^T \mu^\top \text{diag}(t + X_t) \sigma^{-1} (\text{diag}(t + X_t))^{-1} Z_{t,s} ds - \int_t^T Z_{t,s}^\top dW_s, \end{cases} \quad (t, s) \in \Delta[0, T]^2. \quad (24)$$

A direct calculation shows that the unique solution  $\{(Y_t, Z_{t,s})\}_{(t,s) \in \Delta[0, T]^2}$  is given by the closed-form expressions

$$Y_t = \langle t + X_t, X_t \rangle \quad \text{and} \quad Z_{t,s} = \text{diag}(t + X_t) \sigma X_s, \quad (t, s) \in \Delta[0, T]^2. \quad (25)$$

Let  $d = 20$ ,  $T = 1$ ,  $x_0 = (1, \dots, 1)^\top$ ,  $\mu = (-0.05, \dots, -0.05)^\top$ ,  $\sigma = \text{diag}(0.3, 0.375, 0.45, 0.375, 0.3, 0.375, 0.45, 0.375, 0.3, 0.375, 0.45, 0.375, 0.3, 0.375, 0.45, 0.375)$ ,  $N = 40$ .

Figures 6 and 7 display the approximate  $Y$ - and  $Z$ -processes alongside their respective analytical reference solutions. For the  $Y$ -process, we present three representative sample paths, a sample mean, as well as the 5th and 95th sample percentiles. For the  $Z$ -process, one representative sample path is shown, together with a sample mean of its first component.

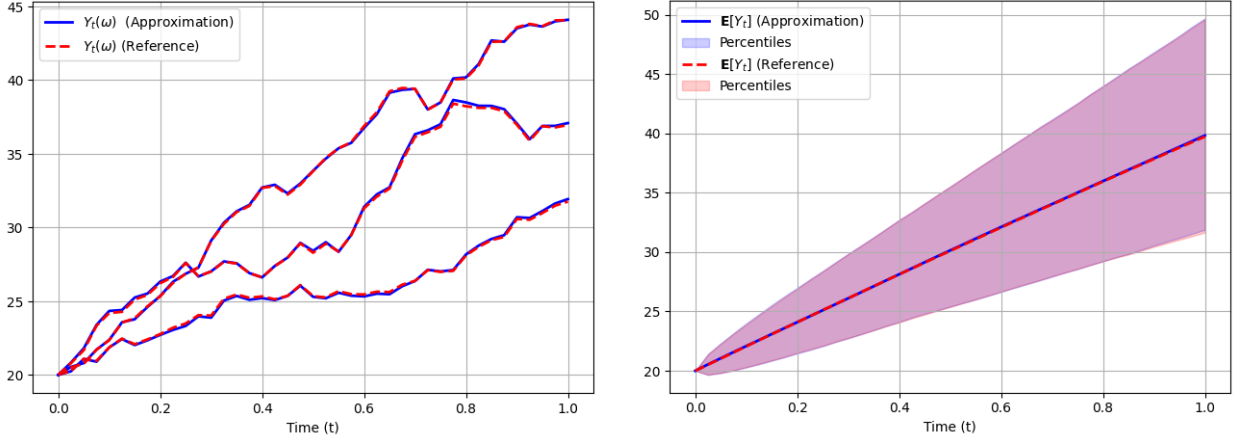


Figure 6: Comparison of the approximated  $Y$  with the reference solutions for Example 2. **Left:** Three representative sample paths. **Right:** The sample mean and the 5th and 95th percentiles.

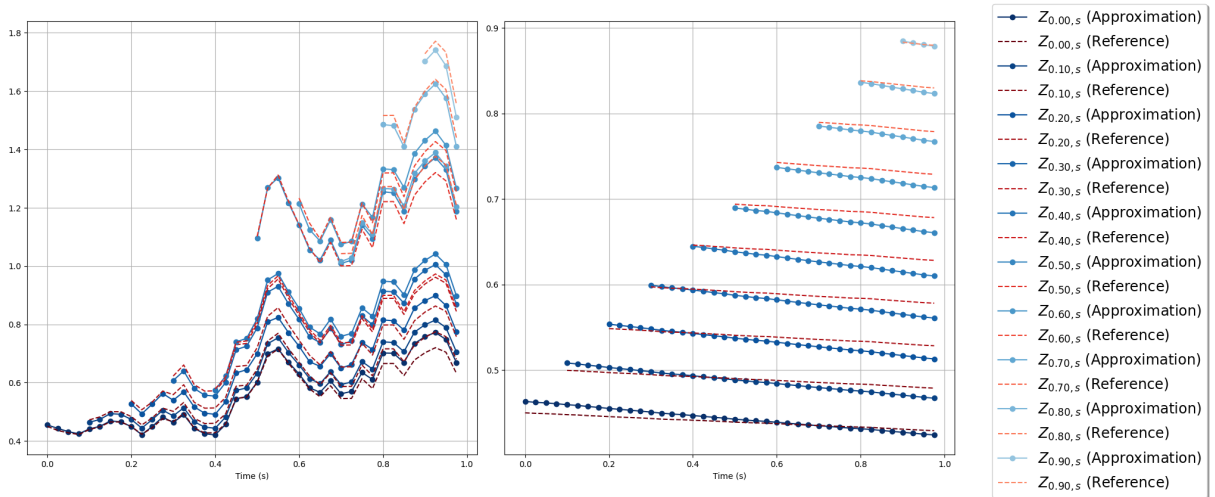


Figure 7: Comparison of the approximated  $Z_{t,s}$  with the reference solution for different values of  $t$  for Example 2. **Left:** One representative sample path of the first (of 20) component of  $Z_{t,s}$ . **Right:** A sample mean the first component of  $Z_{t,s}$ .

Because we do not carry out an empirical error analysis in this section, these figures do not offer insight into the accuracy of our approximation for the remaining 19 components of the  $Z$ -process. Consequently, Figure 8 illustrates a representative sample path for our approximations of  $Z_{0,s}$  and  $Z_{0.5,s}$ , compared to their corresponding analytical reference solutions.

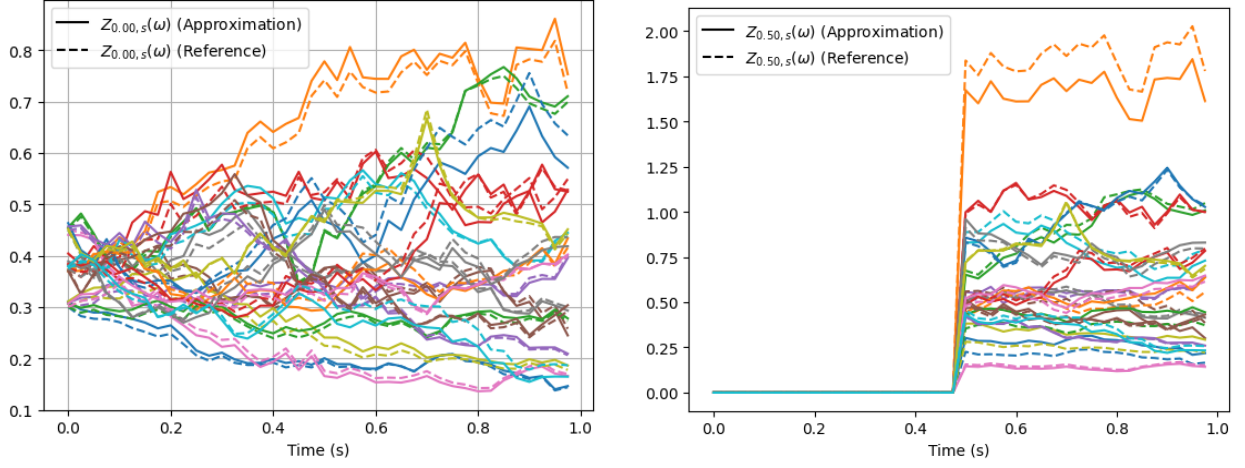


Figure 8: Comparison of one representative sample path of the approximated  $Z_{t,s}$  with its reference solution for Example 2. **Left:**  $t = 0$ . **Right:**  $t = 0.5$ . Since  $Z_{t,s}$  is only defined for  $s \geq t$ , we set  $Z_{t,s} = \mathbf{0}$  for  $s < t$  purely for illustrative purposes.

### 6.2.2 Example 3: An FSDE-BSVIE system coupled in $Y$ and $Z$

For  $d \in \mathbb{N}$ ,  $a, x_0 \in \mathbb{R}^d$ ,  $b, c, k \in \mathbb{R}$ , and  $\sigma \in \mathbb{R}^{d \times d}$ . We consider the following coupled FSDE-BSVIE system

$$\begin{cases} X_t = x_0 + \int_0^t (a + bZ_{s,s}) ds + \int_0^t (c + Y_s) \sigma W_t, & \bar{X}_t = \frac{1}{d} \sum_{i=0}^d X_t^i, \\ Y_t = t \sin(k\bar{X}_T) + \int_t^T \left( \frac{tk^2}{2d^2} \sin(k\bar{X}_s) (c + Y_s)^2 \|\mathbf{1}^\top \sigma\|^2 - \cos(k\bar{X}_s) (t\mathbf{1}^\top a + sb\mathbf{1}^\top Z_{t,s}) \right) ds \\ - \int_t^T Z_{t,s}^\top dW_s, & (t, s) \in \Delta[0, T]^2. \end{cases} \quad (26)$$

A direct calculation shows that the unique solution  $\{(Y_t, Z_{t,s})\}_{(t,s) \in \Delta[0,T]^2}$  is given by the closed-form expressions

$$Y_t = t \sin(k\bar{X}_t) \quad \text{and} \quad Z_{t,s} = t \cos(k\bar{X}_s) (c + s \sin(k\bar{X}_s)) \mathbf{1}^\top \sigma, \quad (t, s) \in \Delta[0, T]^2. \quad (27)$$

Let  $d = k = 5$ ,  $T = 1$ ,  $x_0 = 1$ ,  $a = (0.15, 0.075, 0.0, -0.075, -0.15)^\top$ ,  $\sigma = \text{diag}(0.4, 0.5, 0.6, 0.7, 0.9)$ ,  $c = 1.001$  ( $c > 1$  to guarantee enough ellipticity), and  $N = 40$ .

Figure 9-10 shows our approximate  $X$ -,  $Y$ - and  $Z$ -processes, compared with the semi-analytic (we have to approximate the FSDE with an Euler–Maruyama scheme) reference solutions. In particular, the reference solution for the FSDE is obtained by substituting  $Y_t = t \sin(k\bar{X}_t)$ , and  $Z_{t,t} = \frac{tk}{d} \cos(k\bar{X}_t) (c + t \sin(k\bar{X}_t)) \mathbf{1}^\top \sigma$  into the drift coefficient and approximating the decoupled FSDE via the Euler–Maruyama scheme.

### 6.3 Comparison with existing methods

We do not benchmark our method against existing BSVIE solvers because, strictly speaking, no complete solver is yet available: the methods of Wang [36], Hamaguchi & Taguchi [15] and Pokalyuk [29] provide only the time-grid recursion while leaving the crucial conditional-expectation step unspecified. Implementing those schemes therefore requires choosing and tuning an additional regression, quantization or cubature layer—decisions that are outside the scope of their papers and would introduce a subjective bias into any comparison. Consequently, our tests focus on accuracy versus analytical solutions and on convergence-rate diagnostics, which are the only fair yardsticks at present.

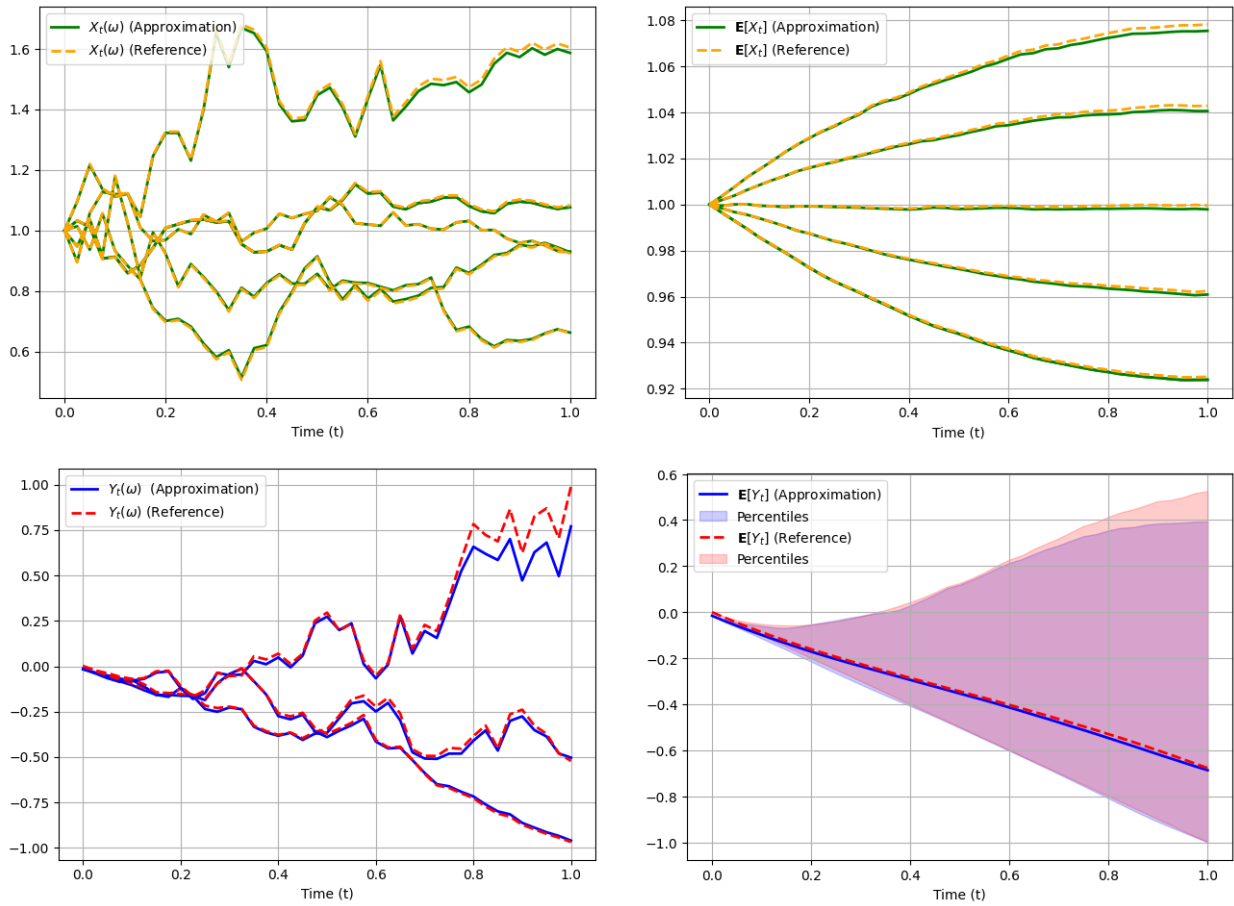


Figure 9: Comparison of the approximated  $X$  and  $Y$  with the reference solutions for Example 3. **Left:** Three representative sample paths. **Right:** The sample mean and, for  $Y$ , the 5th and 95th percentiles.

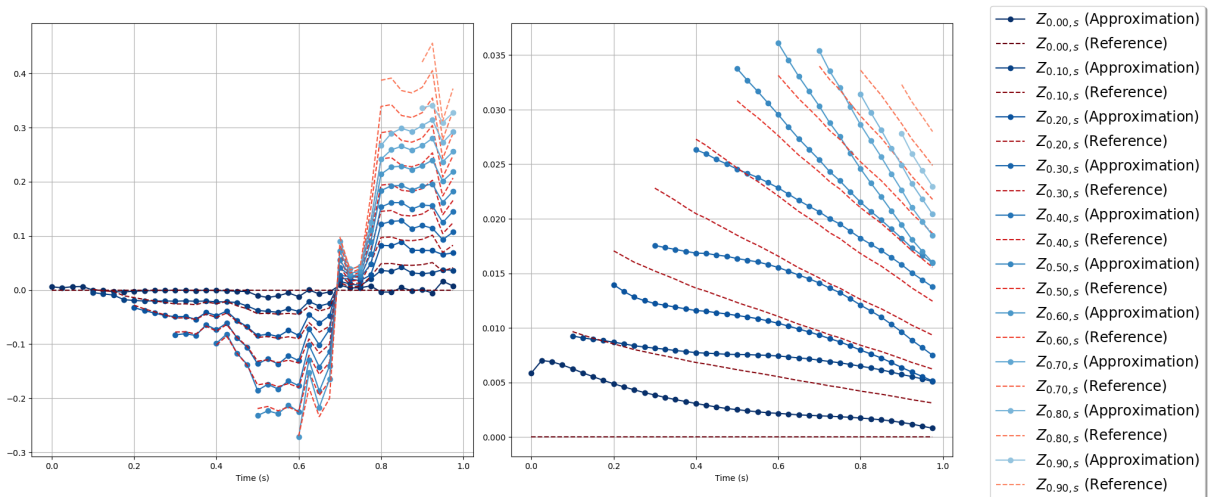


Figure 10: Comparison of component 1 (of 5) of the approximated  $Z_{t,s}$  with the reference solution for different values of  $t$  for Example 3. **Left:** One representative sample path of  $Z_{t,s}$ . **Right:** A sample mean of  $Z_{t,s}$ .

## Acknowledgments

The research of Kristoffer Andersson was funded by the RIBA2022 grant.

## References

- [1] Kristoffer Andersson, Adam Andersson, and Cornelis W Oosterlee. Convergence of a robust deep FBSDE method for stochastic control. *SIAM Journal on Scientific Computing*, 45(1):A226–A255, 2023.
- [2] Kristoffer Andersson, Alessandro Gnoatto, Marco Patacca, and Athena Picarelli. A deep solver for BSDEs with jumps. *SIAM Journal on Financial Mathematics - accepted*, 2025.
- [3] Christian Beck, Sebastian Becker, Philipp Grohs, Nor Jaafari, and Arnulf Jentzen. Solving the Kolmogorov *PDE* by means of deep learning. *Journal of Scientific Computing*, 88(3):1–28, 2021.
- [4] Christian Beck, Martin Hutzenthaler, Arnulf Jentzen, and Benno Kuckuck. An overview on deep learning-based approximation methods for partial differential equations. *Discrete and Continuous Dynamical Systems - B*, 2023, 28(6): 3697–3746. doi: 10.3934/dcdsb.2022238, 2023.
- [5] Christian Beck, Arnulf Jentzen, et al. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *Journal of Nonlinear Science*, 29(4):1563–1619, 2019.
- [6] Francesca Biagini, Alessandro Gnoatto, and Immacolata Oliva. A unified approach to xVA with CSA discounting and initial margin. *SIAM Journal on Financial Mathematics*, 12(3):1013–1053, 2021.
- [7] Dorje C Brody and Lane P Hughston. Social discounting and the long rate of interest. *Mathematical Finance*, 28(1):306–334, 2018.
- [8] Quentin Chan-Wai-Nam, Joseph Mikael, and Xavier Warin. Machine learning for semi linear *PDEs*. *Journal of Scientific Computing*, 79(3):1667–1712, 2019.
- [9] Jared Chessari, Reiichiro Kawai, Yuji Shinozaki, and Toshihiro Yamada. Numerical methods for backward stochastic differential equations: A survey. *Probability Surveys*, 20:486–567, 2023.
- [10] Giulia Di Nunno and Emanuela Rosazza Gianin. Fully dynamic risk measures: Horizon risk, time-consistency, and relations with BSDEs and BSVIEs. *SIAM Journal on Financial Mathematics*, 15(2):399–435, 2024.
- [11] Fang Fang and Cornelis W Oosterlee. A novel pricing method for European options based on Fourier-cosine series expansions. *SIAM Journal on Scientific Computing*, 31(2):826–848, 2009.
- [12] Maximilien Germain, Huyen Pham, and Xavier Warin. Approximation error analysis of some deep backward schemes for nonlinear *PDEs*. *SIAM Journal on Scientific Computing*, 44(1):A28–A56, 2022.
- [13] Alessandro Gnoatto, Katharina Oberpriller, and Athena Picarelli. Convergence of a deep BSDE solver with jumps. *arXiv preprint arXiv:2501.09727*, 2025.
- [14] Philipp Grohs, Fabian Hornung, Arnulf Jentzen, and Philippe von Wurstemberger. *A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations*, volume 284 (1410) of *Memoirs of the American Mathematical Society*. American Mathematical Society, Providence, RI, 2023.
- [15] Yushi Hamaguchi and Dai Taguchi. Approximations for adapted M-solutions of type-ii backward stochastic Volterra integral equations. *ESAIM: Probability and Statistics*, 27:19–79, 2023.

- [16] Jiequn Han. Deep learning approximation for stochastic control problems. *Deep Reinforcement Learning Workshop, NIPS*, 2016.
- [17] Jiequn Han, Arnulf Jentzen, and E Weinan. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.
- [18] Jiequn Han and Jihao Long. Convergence of the deep *BSDE* method for coupled *FBSDEs*. *Probability, Uncertainty and Quantitative Risk*, 5(1):1–33, 2020.
- [19] Côme Huré, Huyên Pham, and Xavier Warin. Deep backward schemes for high-dimensional nonlinear *PDEs*. *Mathematics of Computation*, 89(324):1547–1579, 2020.
- [20] Martin Hutzenthaler, Arnulf Jentzen, Thomas Kruse, and Tuan Anh Nguyen. A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *SN partial differential equations and applications*, 1(2):1–34, 2020.
- [21] Ichiro Ito. On the existence and uniqueness of solutions of stochastic integral equations of the Volterra type. *Kodai Mathematical Journal*, 2(2):158–170, 1979.
- [22] Arnulf Jentzen, Diyora Salimova, and Timo Welti. A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. *Communications in Mathematical Sciences*, 19(5):1167–1205, 2021.
- [23] Shaolin Ji, Shige Peng, Ying Peng, and Xichuan Zhang. Three algorithms for solving high-dimensional fully coupled *FBSDEs* through deep learning. *IEEE Intelligent Systems*, 35(3):71–84, 2020.
- [24] Lorenc Kapllani and Long Teng. A backward differential deep learning-based algorithm for solving high-dimensional nonlinear backward stochastic differential equations. *IMA Journal of Numerical Analysis*, page draf022, 2025.
- [25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Preprint arXiv:1412.6980*, 2014.
- [26] Jianzhong Lin. Adapted solution of a backward stochastic nonlinear Volterra integral equation. *Stochastic Analysis and Applications*, 20(1):165–183, 2002.
- [27] Balint Negyesi, Kristoffer Andersson, and Cornelis W Oosterlee. The one step Malliavin scheme: new discretization of BSDEs implemented with deep learning regressions. *IMA Journal of Numerical Analysis*, 44(6):3595–3647, 2024.
- [28] David Nualart. *The Malliavin Calculus and Related Topics*, volume 1995. Springer, 2006.
- [29] Stanislav Pokalyuk. *Discretization of backward stochastic Volterra integral equations*. PhD thesis, Saarland University, Saarbrücken, Germany, 2012.
- [30] Maziar Raissi. Forward–backward stochastic neural networks: deep learning of high-dimensional partial differential equations. In *Peter Carr Gedenkschrift: Research Advances in Mathematical Finance*, pages 637–655. World Scientific, 2024.
- [31] Christoph Reisinger, Wolfgang Stockinger, and Yufei Zhang. A posteriori error estimates for fully coupled mckean–vlasov forward-backward sdes. *IMA Journal of Numerical Analysis*, 44(4):2323–2369, 2024.
- [32] Hanxiao Wang, Jingrui Sun, and Jiongmin Yong. Recursive utility processes, dynamic risk measures and quadratic backward stochastic volterra integral equations. *Applied Mathematics & Optimization*, 84(1):145–190, 2021.

- [33] Hanxiao Wang and Jiongmin Yong. Time-inconsistent stochastic optimal control problems and backward stochastic Volterra integral equations. *ESAIM: Control, Optimisation and Calculus of Variations*, 27:22, 2021.
- [34] Hanxiao Wang, Jiongmin Yong, and Chao Zhou. Optimal controls for forward-backward stochastic differential equations: Time-inconsistency and time-consistent solutions. *Journal de Mathématiques Pures et Appliquées*, 190:103603, 2024.
- [35] Tianxiao Wang and Harry Zheng. Closed-loop equilibrium strategies for general time-inconsistent optimal control problems. *SIAM Journal on Control and Optimization*, 59(5):3152–3178, 2021.
- [36] Yanqing Wang. A numerical scheme for BSVIEs. *arXiv preprint arXiv:1605.04865*, 2016.
- [37] Jiong-min Yong. Backward stochastic Volterra integral equations—a brief survey. *Applied Mathematics-A Journal of Chinese Universities*, 28(4):383–394, 2013.
- [38] Jiongmin Yong. Backward stochastic Volterra integral equations and some related problems. *Stochastic Processes and their Applications*, 116(5):779–795, 2006.
- [39] Jiongmin Yong. Continuous-time dynamic risk measures by backward stochastic volterra integral equations. *Applicable Analysis*, 86(11):1429–1442, 2007.
- [40] Jiongmin Yong. Well-posedness and regularity of backward stochastic Volterra integral equations. *Probability Theory and Related Fields*, 142(1):21–77, 2008.
- [41] Jiongmin Yong. Time-inconsistent optimal control problems and the equilibrium HJB equation. *Mathematical Control and Related Fields*, 2(3): 271–329. doi: 10.3934/mcrf.2012.2.271, 2012.

## A Proof of Lemma 4.1

In this appendix we provide the full proof of Lemma 4.1, restated here for convenience.

**Lemma A.1.** *Assume that  $h$  is small enough so that  $(2K_1 + \frac{1}{2})h < 1$ . For  $j \in \{1, 2\}$ , suppose  $\{\hat{Y}_n^{k,\pi,j}, \hat{Z}_n^{k,\pi,j}\}_{(k,n) \in \Delta^\pi[0,T]^2}$  are two square integrable solutions to (16). Define the differences*

$$\delta Y_n^k = \hat{Y}_n^{k,\pi,1} - \hat{Y}_n^{k,\pi,2}, \quad \delta Z_n^k = \hat{Z}_n^{k,\pi,1} - \hat{Z}_n^{k,\pi,2}.$$

*Then there exist constants  $C_Y$  and  $C_Z$ , depending only on  $T$  and  $K_1$ , such that for every  $(k, n) \in \Delta^\pi[0, T]^2$ , the following estimates hold:*

$$\begin{aligned} \mathbb{E}|\delta Y_n^k|^2 &\leq C_Y \left( \mathbb{E}|\delta Y_N^k|^2 + \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h \right), \\ \mathbb{E}|\delta Z_n^k|^2 h &\leq C_Z \left( \mathbb{E}|\delta Y_N^k|^2 + \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h \right). \end{aligned}$$

*Proof.* Let  $\delta f_n^k = f(t_k, t_n, X_n^\pi, Y_n^{n,\pi,1}, Z_n^{k,\pi,1}) - f(t_k, t_n, X_n^\pi, Y_n^{n,\pi,2}, Z_n^{k,\pi,2})$ , then

$$\delta Y_{n+1}^k = \mathbb{E}[\delta Y_{n+1}^k | \mathcal{F}_{t_n}] + \delta f_n^k h, \quad \delta Z_n^k = \frac{1}{h} \mathbb{E}[\delta Y_{n+1}^k \Delta W_n | \mathcal{F}_{t_n}].$$

Then, by the martingale representation theorem, there exists an adapted, square integrable process  $(\delta Z_t^k)_{t \in [t_n, n+1]}$ , such that

$$\delta Y_{n+1}^k = \mathbb{E}[\delta Y_{n+1}^k | \mathcal{F}_{t_n}] + \int_{t_n}^{t_{n+1}} \delta Z_n^k dW_t = \delta Y_n^k - \delta f_n^k h + \int_{t_n}^{t_{n+1}} \delta Z_n^k dW_t.$$

Since  $\delta f_n^k$  and  $\delta Y_n^k$  are  $\mathcal{F}_{t_n}$ -measurable, it holds that  $\mathbb{E}[\delta Y_n^k \int_{t_n}^{t_{n+1}} \delta Z_n^k dW_t] = \mathbb{E}[\delta f_n^k \int_{t_n}^{t_{n+1}} \delta Z_n^k dW_t] = 0$  and by Itô-Isometry, we have

$$\mathbb{E}|\delta Y_{n+1}^k|^2 \geq \mathbb{E}|\delta Y_n^k - \delta f_n^k h|^2 + \int_{t_n}^{t_{n+1}} \mathbb{E}|\delta Z_t^k|^2 ds.$$

The first term,  $\mathbb{E}|\delta Y_n^k - \delta f_n^k h|^2$  is treated in two separate cases.

**Case I** ( $n > k$ ):

For the first term on the right-hand side, we use Lipschitz continuity of  $f$ , and then apply the root-mean-square and geometric-mean inequalities with  $\lambda > 0$  to obtain

$$\mathbb{E}|\delta Y_n^k - \delta f_n^k|^2 \geq \mathbb{E}|\delta Y_n^k|^2 - 2h\mathbb{E}[\delta Y_n^k \delta f_n^k] \geq \mathbb{E}|\delta Y_n^k|^2 - \lambda\mathbb{E}|\delta Y_n^k|^2 h - K_1\lambda^{-1}(\mathbb{E}|\delta Y_n^k|^2 + \mathbb{E}|\delta Z_n^k|^2)h.$$

We observe that  $\mathbb{E}[\int_{t_n}^{t_{n+1}} \delta Z_t^k dt | \mathcal{F}_{t_n}] = h\delta Z_n^k$ , see *e.g.*, [18, Lemma 1], which, by the Cauchy-Schwartz inequality, implies that  $\int_{t_n}^{t_{n+1}} \mathbb{E}|\delta Z_t^k|^2 dt \geq \mathbb{E}|\delta Z_n^k|^2 h$ . Collecting terms then yields

$$\mathbb{E}|\delta Y_{n+1}^k|^2 \geq (1 - \lambda h)\mathbb{E}|\delta Y_n^k|^2 - \lambda^{-1}K_1\mathbb{E}|\delta Y_n^k|^2 h + (1 - K_1\lambda^{-1})\mathbb{E}|\delta Z_n^k|^2 h. \quad (28)$$

setting  $\lambda = 2K_1$  (assuming  $K_1 > 0$ ) and rearranging yields

$$(1 - 2K_1 h)\mathbb{E}|\delta Y_n^k|^2 + \frac{1}{2}\mathbb{E}|\delta Z_n^k|^2 h \leq 2\mathbb{E}|\delta Y_{n+1}^k|^2 + \mathbb{E}|\delta Y_n^k|^2 h.$$

Then, for sufficiently small  $h$ , such that  $2K_1 h < 1$ , we have

$$\mathbb{E}|\delta Y_n^k|^2 \leq (1 - 2K_1 h)^{-1} \left( \mathbb{E}|\delta Y_{n+1}^k|^2 + \frac{1}{2}\mathbb{E}|\delta Y_n^k|^2 h \right).$$

Setting  $A_1 = (1 - 2K_1 h)^{-1}$  and iteratively applying the bound for  $\mathbb{E}|\delta Y_n^k|^2$ , yield

$$\mathbb{E}|\delta Y_n^k|^2 \leq A_1^{N-n}\mathbb{E}|\delta Y_N^k|^2 + \frac{1}{2} \sum_{\ell=n}^{N-1} A_1^{\ell+1-n}\mathbb{E}|\delta Y_\ell^k|^2 h. \quad (29)$$

**Case II** ( $n = k$ ):

We, again use the root-mean-square and geometric-mean inequalities with  $\lambda = 2K_1$  and Lipschitz continuity of  $f$  to obtain

$$\begin{aligned} \mathbb{E}|\delta Y_{n+1}^n|^2 &\geq \mathbb{E}|\delta Y_n^n|^2 - 2h\mathbb{E}[\delta Y_n^n \delta f_n^n] + \mathbb{E}|\delta Z_n^n|^2 h \\ &\geq \left(1 - (2K_1 + \frac{1}{2})h\right)\mathbb{E}|\delta Y_n^n|^2 + \frac{1}{2}\mathbb{E}|\delta Z_n^n|^2 h. \end{aligned}$$

The above implies that we have the following bounds

$$\mathbb{E}|\delta Y_n^n|^2 \leq \left(1 - (2K_1 + \frac{1}{2})h\right)^{-1} \mathbb{E}|\delta Y_{n+1}^n|^2, \quad \mathbb{E}|\delta Z_n^n|^2 h \leq 2\mathbb{E}|\delta Y_{n+1}^n|^2. \quad (30)$$

Setting  $A_2 = \left(1 - (2K_1 + \frac{1}{2})h\right)^{-1} A_1$  and combining (29) and (30) iteratively, we obtain

$$\begin{aligned} \mathbb{E}|\delta Y_n^k|^2 &\leq A_1^{N-n}\mathbb{E}|\delta Y_N^k|^2 + A_2 \sum_{\ell=n}^{N-1} A_1^{N-\ell-1} (1 + A_2 h)^{n-\ell} \mathbb{E}|\delta Y_N^\ell|^2 h, \\ \mathbb{E}|\delta Y_n^k|^2 &\leq A_1^N \mathbb{E}|\delta Y_N^k|^2 + A_1^N A_2 (1 + A_2 h)^N \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h, \end{aligned}$$

where in the last step we use the fact that  $A_1, A_2 > 1$  by construction. To bound  $\mathbb{E}|\delta Z_n^k|^2$  we note that, (28) yields

$$\begin{aligned}\mathbb{E}|\delta Z_n^k|^2 h &\leq 2\mathbb{E}|\delta Y_{n+1}^k|^2 + \mathbb{E}|\delta Y_n^n|^2 h \\ &\leq A_1^N \mathbb{E}|\delta Y_N^k|^2 + (1+h)A_1^N A_2^N \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h.\end{aligned}$$

The remaining task is to analyze  $A_1^N$ ,  $A_2^N$ , and  $(1+A_2h)^N$ . Noting that  $h = \frac{T}{N}$ , we begin with the first term

$$A_1^N = \left( \frac{1}{1-2K_1h} \right)^N = \left( 1 + \frac{2K_1T}{N-2K_1T} \right)^N.$$

Hence, for  $N > 2K_1T$ ,  $A_1^N$  is a decreasing sequence in  $N$ , converging to  $e^{2K_1T}$ . Similarly, for  $A_2^N$ , we have that for  $N > (2K_1 + \frac{1}{2})T$

$$A_2^N = \left( 1 + \frac{(2K_1 + \frac{1}{2})T}{N - (2K_1 + \frac{1}{2})T} \right)^N \left( 1 + \frac{2K_1T}{N-2K_1T} \right)^N$$

is a decreasing sequence in  $N$ , converging to  $e^{(3K_1 + \frac{1}{2})T}$ . For the last term,  $(1+A_2h)^N = \left(1 + \frac{A_2T}{N}\right)^N \leq e^{A_2T}$ , which is also decreasing in  $N$ . Finally,  $A_1^N$ ,  $A_2^N$ , and  $(1+A_2h)^N$  are bounded and decreasing sequences in  $N$ , and hence, there exist bounded constants  $C_Y$  and  $C_Z$  such that

$$\begin{aligned}\mathbb{E}|\delta Y_n^k|^2 &\leq C_Y \left( \mathbb{E}|\delta Y_N^k|^2 + \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h \right), \\ \mathbb{E}|\delta Z_n^k|^2 h &\leq C_Z \left( \mathbb{E}|\delta Y_N^k|^2 + \sum_{\ell=n}^{N-1} \mathbb{E}|\delta Y_N^\ell|^2 h \right).\end{aligned}$$

□