RESEARCH   **Open Access**

Convergence of the deep BSDE method for coupled FBSDEs

Jiequn Han · Jihao Long

Abstract The recently proposed numerical algorithm, deep BSDE method, has shown remarkable performance in solving high-dimensional forward-backward stochastic differential equations (FBSDEs) and parabolic partial differential equations (PDEs). This article lays a theoretical foundation for the deep BSDE method in the general case of coupled FBSDEs. In particular, a posteriori error estimation of the solution is provided and it is proved that the error converges to zero given the universal approximation capability of neural networks. Numerical results are presented to demonstrate the accuracy of the analyzed algorithm in solving high-dimensional coupled FBSDEs.

Keywords Forward-backward SDE · Weakly coupled condition · Deep learning · High dimension · Numerics

Abbreviations

PDE: Partial differential equation
SDE: Stochastic differential equation
BSDE: Backward stochastic differential equation
FBSDE: Forward-backward stochastic differential equation

J. Han (✉)
Department of Mathematics, Princeton University, Princeton 08544, NJ, USA
e-mail: jiequnh@princeton.edu

J. Long
School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China

1 Introduction

Forward-backward stochastic differential equations (FBSDEs) and partial differential equations (PDEs) of parabolic type have found numerous applications in stochastic control, finance, physics, etc., as a ubiquitous modeling tool. In most situations encountered in practice the equations cannot be solved analytically but require certain numerical algorithms to provide approximate solutions. On the one hand, the dominant choices of numerical algorithms for PDEs are mesh-based methods, such as finite differences, finite elements, etc. On the other hand, FBSDEs can be tackled directly through probabilistic means, with appropriate methods for the approximation of conditional expectation. Since these two kinds of equations are intimately connected through the nonlinear Feynman–Kac formula (Pardoux and Peng 1992), the algorithms designed for one kind of equation can often be used to solve another one.

However, the aforementioned numerical algorithms become more and more difficult, if not impossible, when the dimension increases. They are doomed to run into the so-called "curse of dimensionality" (Bellman 1957) when the dimension is high, namely, the computational complexity grows exponentially as the dimension grows. The classical mesh-based algorithms for PDEs require a mesh of size $O(N^d)$. The simulation of FBSDEs faces a similar difficulty in the general nonlinear cases, due to the need to compute conditional expectation in high dimension. The conventional methods, including the least squares regression (Bender and Steiner 2012), Malliavin approach (Bouchard et al. 2004), and kernel regression (Bouchard and Touzi 2004), are all of exponential complexity. There are a limited number of cases where practical high-dimensional algorithms are available. For example, in the linear case, Feynman–Kac formula and Monte Carlo simulation together provide an efficient approach to solving PDEs and associated BSDEs numerically. In addition, methods based on the branching diffusion process (Henry-Labordere 2012; Henry-Labordere et al. 2019) and multilevel Picard iteration (Hutzenthaler et al. 2016, 2018; E et al. 2019) overcome the curse of dimensionality in their considered settings. We refer (E et al. 2019) for the detailed discussion on the complexity of the algorithms mentioned above. Overall there is no numerical algorithm in literature so far proved to overcome the curse of dimensionality for general quasilinear parabolic PDEs and the corresponding FBSDEs.

A recently developed algorithm, called the deep BSDE method (Han et al. 2018; E et al. 2017), has shown astonishing power in solving general high-dimensional FBSDEs and parabolic PDEs (Beck et al. 2017; Han and Hu 2019; Han et al. 2020). In contrast to conventional methods, the deep BSDE method employs neural networks to approximate unknown gradients and reformulates the original equation-solving problem into a stochastic optimization problem. Thanks to the universal approximation capability and parsimonious parameterization of neural networks, in practice the objective function can be effectively optimized in high-dimensional cases, and the function values of interests are obtained quite accurately.

The deep BSDE method was initially proposed for decoupled FBSDEs. In this paper, we extend the method to deal with coupled FBSDEs and a broader class of quasilinear parabolic PDEs. Furthermore, we present an error analysis of the proposed scheme, including decoupled FBSDEs as a special case. Our theoretical

result consists of two theorems. Theorem 1 provides a posteriori error estimation of the deep BSDE method. As long as the objective function is optimized to be close to zero under fine time discretization, the approximate solution is close to the true solution. In other words, in practice, the accuracy of the numerical solution is effectively indicated by the value of the objective function. Theorem 2 shows that such a situation is attainable, by relating the infimum of the objective function to the expression ability of neural networks. As an implication of the universal approximation property (in the L^2 sense), there exist neural networks with suitable parameters such that the obtained numerical solution is approximately accurate. To the best of our knowledge, this is the first theoretical result of the deep BSDE method for solving FBSDEs and parabolic PDEs. Although our numerical algorithm is based on neural networks, the theoretical result provided here is equally applicable to the algorithms based on other forms of function approximations.

The article is organized as follows. In Section 2, we precisely state our numerical scheme for coupled FBSDEs and quasilinear parabolic PDEs and give the main theoretical results of the proposed numerical scheme. In Section 3, the basic assumptions and some useful results from the literature are given for later use. The proofs of the two main theorems are provided in Sections 4 and 5, respectively. Some numerical experiments with the proposed scheme are presented in Section 6.

2 A numerical scheme for coupled FBSDEs and main results

Let $T \in (0, +\infty)$ be the terminal time, $\left(\Omega, \mathbb{F}, \{\mathcal{F}_t\}_{0 \leq t \leq T}, \mathbb{P}\right)$ be a filtered probability space equipped with a d-dimensional standard Brownian motion $\{W_t\}_{0 \leq t \leq T}$ starting from 0. ξ is a square-integrable random variable independent of $\{W_t\}_{0 \leq t \leq T}$. We use the same notation $\left(\Omega, \mathbb{F}, \{\mathcal{F}_t\}_{0 \leq t \leq T}, \mathbb{P}\right)$ to denote the filtered probability space generated by $\{W_t + \xi\}_{0 \leq t \leq T}$. The notation $|x|$ denotes the Euclidean norm of a vector x and $\|A\| = \sqrt{\text{trace}\left(A^{\mathrm{T}} A\right)}$ denotes the Frobenius norm of a matrix A.

Consider the following coupled FBSDEs

$$
\begin{cases}
X_t = \xi + \displaystyle\int_0^t b(s, X_s, Y_s)\, \mathrm{d}s + \int_0^t \sigma(s, X_s, Y_s)\, \mathrm{d}W_s, & (2.1) \\[4mm]
Y_t = g(X_T) + \displaystyle\int_t^T f(s, X_s, Y_s, Z_s)\, \mathrm{d}s - \int_t^T (Z_s)^{\mathrm{T}}\, \mathrm{d}W_s, & (2.2)
\end{cases}
$$

in which X_t takes values in $\mathbb{R}^m$, Y_t takes values in $\mathbb{R}$, and Z_t takes values in $\mathbb{R}^d$. Here we assume Y_t to be one-dimensional to simplify the presentation. The result can be extended without any difficulty to the case where Y_t is multi-dimensional. We say (X_t, Y_t, Z_t) is a solution of the above FBSDEs, if all its components are $\mathcal{F}_t$-adapted and square-integrable, together satisfying Eqs. (2.1) and (2.2).

Solving coupled FBSDEs numerically is more difficult than solving decoupled FBSDEs. Except the Picard iteration method developed in Bender and Zhang (2008), most methods exploit the relation to quasilinear parabolic PDEs via the

four-time-step-scheme in Ma et al. (1994). This type of methods suffers from high dimensionality due to spatial discretization of PDEs. In contrast, our strategy, starting from simulating the coupled FBSDEs directly, is a new purely probabilistic scheme. To state the numerical algorithm precisely, we consider a partition of the time interval $[0, T]$, $\pi : 0 = t_0 < t_1 < \cdots < t_N = T$ with $h = T/N$ and $t_i = ih$. Let $\Delta W_i := W_{t_{i+1}} - W_{t_i}$ for $i = 0, 1, \ldots, N - 1$. Inspired by the nonlinear Feynman–Kac formula that will be introduced below, we view Y_0 as a function of X_0 and view Z_t as a function of X_t and Y_t. Equipped with this viewpoint, our goal becomes finding appropriate functions $\mu_0^\pi : \mathbb{R}^m \to \mathbb{R}$ and $\phi_i^\pi : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^d$ for $i = 0, 1, \ldots, N - 1$ such that $\mu_0^\pi(\xi)$ and $\phi_i^\pi \left(X_{t_i}^\pi, Y_{t_i}^\pi \right)$ can serve as good surrogates of Y_0 and Z_{t_i}, respectively. To this end, we consider the classical Euler scheme

$$
\begin{cases}
X_0^\pi = \xi, \quad Y_0^\pi = \mu_0^\pi(\xi), \\
X_{t_{i+1}}^\pi = X_{t_i}^\pi + b\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi \right) h + \sigma \left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi \right) \Delta W_i, \\
Z_{t_i}^\pi = \phi_i^\pi \left(X_{t_i}^\pi, Y_{t_i}^\pi \right), \\
Y_{t_{i+1}}^\pi = Y_{t_i}^\pi - f\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi, Z_{t_i}^\pi \right) h + \left(Z_{t_i}^\pi \right)^{\mathrm{T}} \Delta W_i.
\end{cases}
\tag{2.3}
$$

Without loss of clarity, here we use the notation X_0^π as $X_{t_0}^\pi$, X_T^π as $X_{t_N}^\pi$, etc.

Following the spirit of the deep BSDE method, we employ a stochastic optimizer to solve the following stochastic optimization problem

$$
\inf_{\mu_0^\pi \in \mathcal{N}_0', \phi_i^\pi \in \mathcal{N}_i} F\left(\mu_0^\pi, \phi_0^\pi, \ldots, \phi_{N-1}^\pi \right) := E|g\left(X_T^\pi \right) - Y_T^\pi|^2,
\tag{2.4}
$$

where $\mathcal{N}_0'$ and $\mathcal{N}_i$ ($0 \leq i \leq N - 1$) are parametric function spaces generated by neural networks. To see intuitively where the objective function (2.4) comes from, we consider the following variational problem:

$$
\inf_{Y_0, \{Z_t\}_{0 \leq t \leq T}} E|g(X_T) - Y_T|^2,
$$

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

$$
Y_t = Y_0 - \int_0^t f(s, X_s, Y_s, Z_s)\,ds + \int_0^t (Z_s)^{\mathrm{T}}\,dW_s,
$$

where Y_0 is $\mathcal{F}_0$-measurable and square-integrable, and Z_t is a $\mathcal{F}_t$-adapted square-integrable process. The solution of the FBSDEs (2.1) and (2.2) is a minimizer of the above problem since the loss function attains zero when it is evaluated at the solution. In addition, the wellposedness of the FBSDEs (under some regularity conditions) ensures the existence and uniqueness of the minimizer. Therefore, we expect (2.4), as a discretized counterpart of (2.5), defines a benign optimization problem and the associated near-optimal solution provides us a good approximate solution of the original FBSDEs. The reason we do not represent Z_{t_i} as a function of X_{t_i} only is that the process $\{X_{t_i}^\pi\}_{0 \leq i \leq N}$ is not Markovian, while the process $\left\{ X_{t_i}^\pi, Y_{t_i}^\pi \right\}_{0 \leq i \leq N}$ is Markovian, which facilitates our analysis considerably. If b and σ are both independent of Y, then the FBSDEs (2.1) and (2.2) are decoupled, we can take ϕ_i^π as a function of $X_{t_i}^\pi$ only, as the numerical scheme introduced in Han et al. (2018) and E et al. (2017).

Our two main theorems regarding the deep BSDE method are the following, mainly on the justification and property of the objective function (2.4) in the general coupled case, regardless of the specific choice of parametric function spaces. An important assumption for the two theorems is the so-called weak coupling or monotonicity condition, which will be explained in detail in Section 3. The precise statement of the theorems can be found in Theorem 1' (Section 4) and Theorem 2' (Section 5), respectively.

Theorem 1 *Under some assumptions, there exists a constant C, independent of h, d, and m, such that for sufficiently small h,*

$$
\sup_{t\in[0,T]} \left(E|X_t - \hat{X}_t^\pi|^2 + E|Y_t - \hat{Y}_t^\pi|^2 \right) + \int_0^T E|Z_t - \hat{Z}_t^\pi|^2 \, \mathrm{d}t \le C \left[h + E|g\left(X_T^\pi\right) - Y_T^\pi|^2 \right],
$$

(2.5)

where $\hat{X}_t^\pi = X_{t_i}^\pi$, $\hat{Y}_t^\pi = Y_{t_i}^\pi$, $\hat{Z}_t^\pi = Z_{t_i}^\pi$ for $t \in [t_i, t_{i+1})$.

Theorem 2 *Under some assumptions, there exists a constant C, independent of h, d and m, such that for sufficiently small h,*

$$
\inf_{\mu_0^\pi \in \mathcal{N}_0', \phi_i^\pi \in \mathcal{N}_i} E|g\left(X_T^\pi\right) - Y_T^\pi|^2
$$

$$
\le C \left\{ h + \inf_{\mu_0^\pi \in \mathcal{N}_0', \phi_i^\pi \in \mathcal{N}_i} \left[E|Y_0 - \mu_0^\pi(\xi)|^2 \right.\right.
$$

$$
\left.\left. + \sum_{i=0}^{N-1} E|E\left[\tilde{Z}_{t_i}|X_{t_i}^\pi, Y_{t_i}^\pi\right] - \phi_i^\pi\left(X_{t_i}^\pi, Y_{t_i}^\pi\right)|^2 h \right] \right\},
$$

where $\tilde{Z}_{t_i} = h^{-1} E\left[\int_{t_i}^{t_{i+1}} Z_t \, \mathrm{d}t | \mathcal{F}_{t_i}\right]$. If b and σ are independent of Y, the term $E[\tilde{Z}_{t_i}|X_{t_i}^\pi, Y_{t_i}^\pi]$ can be replaced with $E[\tilde{Z}_{t_i}|X_{t_i}^\pi]$.

Briefly speaking, Theorem 1 states that the simulation error (left side of Eq. (2.5)) can be bounded through the value of the objective function (2.4). To the best of our knowledge, this is the first result for the error estimation of the coupled FBSDEs, concerning both time discretization error and terminal distance. Theorem 2 states that the optimal value of the objective function can be small if the approximation capability of the parametric function spaces ($\mathcal{N}_0'$ and $\mathcal{N}_i$ above) is high. Neural networks are a promising candidate for such a requirement, especially in high-dimensional problems. There are numerous results, dating back to the 90s (see, e.g., Cybenko (1989), Funahashi (1989), Hornik et al. (1989), Barron (1993), Arora et al. (2018), Eldan and Shamir (2016), Cohen et al. (2016), Mhaskar and Poggio (2016), Bölcskei et al. (2017), and Liang and Srikant (2017)), in regard to the universal approximation and complexity of neural networks. There are also some recent analysis (Grohs et al. 2018; Jentzen et al. 2018; Berner et al. 2018; Hutzenthaler et al. 2020) on approximating the solutions of certain parabolic partial differential equations with neural networks. However, the problem is still far from resolved. Theorem 2 implies that if the involved conditional expectations can be approximated by neural networks whose

numbers of parameters growing at most polynomially both in the dimension and the reciprocal of the required accuracy, then the solutions of the considered FBSDEs can be represented in practice without the curse of dimensionality. Under what conditions this assumption is true is beyond the scope of this work and remains for further investigation.

The above-mentioned scheme in (2.3) and (2.4) is for solving FBSDEs. The so-called nonlinear Feynman–Kac formula, connecting FBSDEs with the quasilinear parabolic PDEs, provides an approach to numerically solve quasilinear parabolic PDEs (2.6) below through the same scheme. We recall a concrete version of the nonlinear Feynman–Kac formula in Theorem 3 below and refer interested readers to e.g., (Ma and Yong 2007) for more details. According to this formula, the term $E|Y_0 - Y_0^\pi|^2$ can be interpreted as $E|u(0, \xi) - \mu_0^\pi(\xi)|^2$. Therefore, we can choose the random variable ξ with a delta distribution, a uniform distribution in a bounded region, or any other distribution we are interested in. After solving the optimization problem, we obtain $\mu_0^\pi(\xi)$ as an approximation of $u(0, \xi)$. See (Han et al. 2018; E et al. 2017) for more details.

Theorem 3 *Assume*

1. $m = d$ and $b(t, x, y)$, $\sigma(t, x, y)$, $f(t, x, y, z)$ are smooth functions with bounded first-order derivatives with respect to x, y, z.
2. There exist a positive continuous function v and a constant μ, satisfying that

$$v(|y|)\mathbf{I} \le \sigma\sigma^{\mathrm{T}}(t, x, y) \le \mu\mathbf{I},$$
$$|b(t, x, 0)| + |f(t, x, 0, z)| \le \mu.$$

3. There exists a constant $\alpha \in (0, 1)$ such that g is bounded in the Hölder space $C^{2,\alpha}(\mathbb{R}^m)$.

Then the following quasilinear PDE has a unique classical solution $u(t, x)$ that is bounded with bounded u_t, $\nabla_x u$, and $\nabla_x^2 u$,

$$\begin{cases} u_t + \dfrac{1}{2}trace\left(\sigma\sigma^{\mathrm{T}}(t, x, u)\nabla_x^2 u\right) \\ \quad + b^{\mathrm{T}}(t, x, u)\nabla_x u + f\left(t, x, u, \sigma^{\mathrm{T}}(t, x, u)\nabla_x u\right) = 0, \\ u(T, x) = g(x). \end{cases} \tag{2.6}$$

The associated FBSDEs (2.1) and (2.2) have a unique solution (X_t, Y_t, Z_t) with $Y_t = u(t, X_t)$, $Z_t = \sigma^{\mathrm{T}}(t, X_t, u(t, X_t))\nabla_x u(t, X_t)$, and X_t is the solution of the following SDE

$$X_t = \xi + \int_0^t b(s, X_s, u(s, X_s))\,\mathrm{d}s + \int_0^t \sigma(s, X_s, u(s, X_s))\,\mathrm{d}W_s.$$

Remark 1 *The statement regarding FBSDEs (2.1) and (2.2) in Theorem 3 is developed through a PDE-based argument, which requires $m = d$, uniform ellipticity of σ, and high-order smoothness of b, σ, f, and g. An analogous result through probabilistic argument is given below in Theorem 4 (point 4). In that case, we only need the Lipschitz condition for all of the involved functions, in addition to some weak*

coupling or monotonicity conditions demonstrated in Assumption 3. Note that the Lipschitz condition alone does not guarantee the existence of a solution to the coupled FBSDEs, even in the situation when b, f, σ are linear (see Bender and Zhang (2008) and Ma and Yong (2007) for a concrete counterexample).

Remark 2 *Theorem 3 also implies that the assumption that the drift function b only depends on x, y is general. If b depends on z as well, one can move the associated term in (2.6) into the nonlinearity f and apply the nonlinear Feynman–Kac formula back to obtain an equivalent system of coupled FBSDEs, in which the new drift function is independent of z.*

3 Preliminaries

In this section, we introduce our assumptions and two useful results in Bender and Zhang (2008). We use the notation $\Delta x = x_1 - x_2$, $\Delta y = y_1 - y_2$, $\Delta z = z_1 - z_2$.

Assumption 1 1. *There exist (possibly negative) constants k_b, k_f such that*

$$[b(t, x_1, y) - b(t, x_2, y)]^{\mathrm{T}} \Delta x \le k_b |\Delta x|^2,$$
$$[f(t, x, y_1, z) - f(t, x, y_2, z)] \Delta y \le k_f |\Delta y|^2.$$

2. *b, σ, f, g are uniformly Lipschitz continuous with respect to (x,y,z). In particular, there are non-negative constants K, b_y, σ_x, σ_y, f_x, f_z, and g_x such that*

$$|b(t, x_1, y_1) - b(t, x_2, y_2)|^2 \le K|\Delta x|^2 + b_y|\Delta y|^2,$$
$$\|\sigma(t, x_1, y_1) - \sigma(t, x_2, y_2)\|^2 \le \sigma_x|\Delta x|^2 + \sigma_y|\Delta y|^2,$$
$$|f(t, x_1, y_1, z_1) - f(t, x_2, y_2, z_2)|^2 \le f_x|\Delta x|^2 + K|\Delta y|^2 + f_z|\Delta z|^2,$$
$$|g(x_1) - g(x_2)|^2 \le g_x|\Delta x|^2.$$

3. *$b(t, 0, 0)$, $f(t, 0, 0, 0)$, and $\sigma(t, 0, 0)$ are bounded. In particular, there are constants b_0, σ_0, f_0, and g_0 such that*

$$|b(t, x, y)|^2 \le b_0 + K|x|^2 + b_y|y|^2,$$
$$\|\sigma(t, x, y)\|^2 \le \sigma_0 + \sigma_x|x|^2 + \sigma_y|y|^2,$$
$$|f(t, x, y, z)|^2 \le f_0 + f_x|x|^2 + K|y|^2 + f_z|z|^2,$$
$$|g(x)|^2 \le g_0 + g_x|x|^2.$$

We note here b_y et al. are all constants, not partial derivatives. For convenience, we use $\mathscr{L}$ to denote the set of all the constants mentioned above and assume K is the upper bound of $\mathscr{L}$.

Assumption 2 *b, σ, f are uniformly Hölder-$\frac{1}{2}$ continuous with respect to t. We assume the same constant K to be the upper bound of the square of the Hölder constants as well.*

Assumption 3 *One of the following five cases holds:*

1. *Small time duration, that is, T is small.*
2. *Weak coupling of Y into the forward SDE* (2.1), *that is, b_y and σ_y are small. In particular, if $b_y = \sigma_y = 0$, then the forward equation does not depend on the backward one and, thus, Eqs.* (2.1) *and* (2.2) *are decoupled.*
3. *Weak coupling of X into the backward SDE* (2.2), *that is, f_x and g_x are small. In particular, if $f_x = g_x = 0$, then the backward equation does not depend on the forward one and, thus, Eqs.* (2.1) *and* (2.2) *are also decoupled. In fact, in this case, Z = 0 and* (2.2) *reduces to an ODE.*
4. *f is strongly decreasing in y, that is, k_f is very negative.*
5. *b is strongly decreasing in x, that is, k_b is very negative.*

The assumptions stated above are usually called weak coupling and monotonicity conditions in literature (Bender and Zhang 2008; Antonelli 1993; Pardoux and Tang 1999). To make it more precise, we define

$$L_0 = [b_y + \sigma_y][g_x + f_x T]T e^{[b_y + \sigma_y][g_x + f_x T]T + [2k_b + 2k_f + 2 + \sigma_x + f_z]T},$$

$$L_1 = [g_x + f_x T][e^{[b_y + \sigma_y][g_x + f_x T]T + [2k_b + 2k_f + 2 + \sigma_x + f_z]T + 1} \vee 1],$$

$$\Gamma_0(x) = \frac{e^x - 1}{x}, \quad (x > 0),$$

$$\Gamma_1(x, y) = \sup_{0 < \theta < 1} \theta e^{\theta x} \Gamma_0(y),$$

$$\begin{aligned}
c = \inf_{\lambda_1 > 0} \Big\{ &\Big[e^{[2k_b + 1 + \sigma_x + [b_y + \sigma_y]L_1]T} \vee 1 \Big] \left(1 + \lambda_1^{-1}\right) [b_y + \sigma_y]T \\
&\times \Big[g_x \Gamma_1([2k_f + 1 + f_z]T, [2k_b + 1 + \sigma_x + (1 + \lambda_1)[b_y + \sigma_y]L_1]T) \\
&+ f_x T \Gamma_0([2k_f + 1 + f_z]T) \\
&\times \Gamma_0(2k_b + 1 + \sigma_x + (1 + \lambda_1)[b_y + \sigma_y]L_1]T) \Big] \Big\}.
\end{aligned}$$

Then, a specific quantitative form of the above five conditions can be summarized as:

$$L_0 < e^{-1} \text{ and } c < 1. \tag{3.1}$$

In other words, if any of the five conditions of the weak coupling and monotonicity conditions holds to certain extent, the two inequalities in (3.1) hold. Below, we refer to (3.1) as Assumption 3 and the five general qualitative conditions described above as the weak coupling and monotonicity conditions.

The above three assumptions are basic assumptions in Bender and Zhang (2008), which we need in order to use the results from (Bender and Zhang 2008), as stated in Theorems 4 and 5 below. Theorem 4 gives the connections between coupled FBS-DEs and quasilinear parabolic PDEs under weaker conditions. Theorem 5 provides the convergence of the implicit scheme for coupled FBSDEs. Our work primarily uses the same set of assumptions except that we assume some further quantitative restrictions related to the weak coupling and monotonicity conditions, which will be transparent through the extra constants we define in proofs. Our aim is to provide

explicit conditions on which our results hold and more clearly present the relationship between these constants and the error estimates. As will be seen in the proof, roughly speaking, the weaker the coupling (resp., the stronger the monotonicity, the smaller the time horizon) is, the easier the condition is satisfied, and the smaller the constant C related with error estimates are.

Theorem 4 *Under Assumptions* 1, 2, *and* 3, *there exists a function* $u: \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ *that satisfies the following statements.*

1. $|u(t, x_1) - u(t, x_2)|^2 \leq L_1 |x_1 - x_2|^2.$
2. $|u(s, x) - u(t, x)|^2 \leq C \left(1 + |x|^2\right) |s - t|$ *with some constant C depending on* $\mathscr{L}$ *and T.*
3. *u is a viscosity solution of the PDE* (2.6).
4. *The FBSDEs* (2.1) *and* (2.2) *have a unique solution* (X_t, Y_t, Z_t) *and* $Y_t = u(t, X_t)$. *Thus,* (X_t, Y_t, Z_t) *satisfies decoupled FBSDEs*

$$
\begin{cases}
X_t = \xi + \int_0^t b(s, X_s, u(s, X_s)) \, ds + \int_0^t \sigma(s, X_s, u(s, X_s)) \, dW_s, \\
Y_t = g(X_T) + \int_t^T f(s, X_s, Y_s, Z_s) \, ds - \int_t^T (Z_s)^{\mathrm{T}} \, dW_s.
\end{cases}
$$

Furthermore, the solution of the FBSDEs satisfies the path regularity with some constant C depending on $\mathscr{L}$ *and T*

$$
\sup_{t \in [0,T]} \left(E|X_t - \tilde{X}_t|^2 + E|Y_t - \tilde{Y}_t|^2 \right) + \int_0^T E|Z_t - \tilde{Z}_t|^2 \, dt \leq C \left(1 + E|\xi|^2\right) h,
\tag{3.2}
$$

in which $\tilde{X}_t = X_{t_i}$, $\tilde{Y}_t = Y_{t_i}$, $\tilde{Z}_t = h^{-1} E \left[\int_{t_i}^{t_{i+1}} Z_t \, dt | \mathcal{F}_{t_i} \right]$ *for* $t \in [t_i, t_{i+1})$. *If* Z_t *is càdlàg, we can replace* $h^{-1} E \left[\int_{t_i}^{t_{i+1}} Z_t \, dt | \mathcal{F}_{t_i} \right]$ *with* Z_{t_i}.

Remark 3 *Several conditions can guarantee* Z_t *admits a càdlàg version, such as* $m = d$ *and* $\sigma \sigma^{\mathrm{T}} \geq \delta I$ *with some* $\delta > 0$, *see e.g.,* (Zhang 2004).

Theorem 5 *Under Assumptions* 1, 2, *and* 3, *for sufficiently small h, the following discrete-time equation* $(0 \leq i \leq N - 1)$

$$
\begin{cases}
\overline{X}_0^\pi = \xi, \\
\overline{X}_{t_{i+1}}^\pi = \overline{X}_{t_i}^\pi + b \left(t_i, \overline{X}_{t_i}^\pi, \overline{Y}_{t_i}^\pi\right) h + \sigma \left(t_i, \overline{X}_{t_i}^\pi, \overline{Y}_{t_i}^\pi\right) \Delta W_i, \\
\overline{Y}_T^\pi = g \left(\overline{X}_T^\pi\right), \\
\overline{Z}_{t_i}^\pi = \frac{1}{h} E \left[\overline{Y}_{t_{i+1}}^\pi \Delta W_i | \mathcal{F}_{t_i}\right], \\
\overline{Y}_{t_i}^\pi = E \left[\overline{Y}_{t_{i+1}}^\pi + f \left(t_i, \overline{X}_{t_i}^\pi, \overline{Y}_{t_i}^\pi, \overline{Z}_{t_i}^\pi\right) h | \mathcal{F}_{t_i}\right],
\end{cases}
\tag{3.3}
$$

has a solution $\left(\overline{X}_{t_i}^{\pi}, \overline{Y}_{t_i}^{\pi}, \overline{Z}_{t_i}^{\pi}\right)$ such that $\overline{X}_{t_i}^{\pi} \in L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P})$ and

$$\sup_{t \in [0,T]} \left(E|X_t - \overline{X}_t^{\pi}|^2 + E|Y_t - \overline{Y}_t^{\pi}|^2\right) + \int_0^T E|Z_t - \overline{Z}_t^{\pi}|^2 \, dt \leq C \left(1 + E|\xi|^2\right) h,$$

(3.4)

where $\overline{X}_t^{\pi} = \overline{X}_{t_i}^{\pi}$, $\overline{Y}_t^{\pi} = \overline{Y}_{t_i}^{\pi}$, $\overline{Z}_t^{\pi} = \overline{Z}_{t_i}^{\pi}$ for $t \in [t_i, t_{i+1})$, and C is a constant depending on $\mathcal{L}$ and T.

Remark 4 *In Bender and Zhang (2008), the above result (existence and convergence) is proved for the explicit scheme, which is formulated as replacing $f\left(t_i, \overline{X}_{t_i}^{\pi}, \overline{Y}_{t_i}^{\pi}, \overline{Z}_{t_i}^{\pi}\right)$ with $f\left(t_i, \overline{X}_{t_i}^{\pi}, \overline{Y}_{t_{i+1}}^{\pi}, \overline{Z}_{t_i}^{\pi}\right)$ in the last equation of (3.3). The same techniques can be used to prove the implicit scheme, as we state in Theorem 5.*

Finally, to make sure the system in (2.3) is well-defined, we restrict our parametric function spaces $\mathcal{N}_0'$ and $\mathcal{N}_i$ as in Assumption 4 below. Note that neural networks with common activation functions, including ReLU and sigmoid function, satisfy this assumption. Under Assumption 1 and 4, one can easily prove by induction that $\left\{X_{t_i}^{\pi}\right\}_{0 \leq i \leq N}$, $\left\{Y_{t_i}^{\pi}\right\}_{0 \leq i \leq N}$ and $\left\{Z_{t_i}^{\pi}\right\}_{0 \leq i \leq N-1}$ defined in (2.3) are all measurable and square-integrable random variables.

Assumption 4 $\mathcal{N}_0'$ and $\mathcal{N}_i (0 \leq i \leq N-1)$ are subsets of measurable functions from $\mathbb{R}^m$ to $\mathbb{R}$ and $\mathbb{R}^m \times \mathbb{R}$ to $\mathbb{R}^d$ with linear growth, namely, μ_0^{π} and $\left\{\phi_i^{\pi}\right\}_{0 \leq i \leq N-1}$ in (2.3) satisfy $|\mu_0^{\pi}(x)|^2 \leq A_0' + B_0'|x|^2$ and $|\phi_i^{\pi}(x, y)|^2 \leq A_i + B_i|x|^2 + C_i|y|^2$ for $0 \leq i \leq N-1$.

4 A posteriori estimation of the simulation error

We prove Theorem 1 in this section. Comparing the statements of Theorems 1 and 5, we wish to bound the differences between $\left(X_{t_i}^{\pi}, Y_{t_i}^{\pi}, Z_{t_i}^{\pi}\right)$ and $\left(\overline{X}_{t_i}^{\pi}, \overline{Y}_{t_i}^{\pi}, \overline{Z}_{t_i}^{\pi}\right)$ with the objective function $E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2$. Recalling the definition of the system of Eq. (2.3), we have

$$\begin{cases} X_{t_{i+1}}^{\pi} = X_{t_i}^{\pi} + b\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right) h + \sigma\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right) \Delta W_i, & (4.1) \\ Y_{t_{i+1}}^{\pi} = Y_{t_i}^{\pi} - f\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}, Z_{t_i}^{\pi}\right) h + \left(Z_{t_i}^{\pi}\right)^{\mathrm{T}} \Delta W_i. & (4.2) \end{cases}$$

Taking the expectation $E[\cdot|\mathcal{F}_{t_i}]$ on both sides of (4.2), we obtain

$$Y_{t_i}^{\pi} = E\left[Y_{t_{i+1}}^{\pi} + f\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}, Z_{t_i}^{\pi}\right) h|\mathcal{F}_{t_i}\right].$$

Right multiplying $(\Delta W_i)^{\mathrm{T}}$ on both sides of (4.2) and taking the expectation $E[\cdot|\mathcal{F}_{t_i}]$ again, we obtain

$$Z_{t_i}^{\pi} = \frac{1}{h}\left[Y_{t_{i+1}}^{\pi} \Delta W_i|\mathcal{F}_{t_i}\right].$$

The above observation motivates us to consider the following system of equations

$$
\begin{cases}
X_0^\pi = \xi, \\
X_{t_{i+1}}^\pi = X_{t_i}^\pi + b\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi\right) h + \sigma\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi\right) \Delta W_i, \\
Z_{t_i}^\pi = \dfrac{1}{h} E\left[Y_{t_{i+1}}^\pi \Delta W_i | \mathcal{F}_{t_i}\right], \\
Y_{t_i}^\pi = E\left[Y_{t_{i+1}}^\pi + f\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi, Z_{t_i}^\pi\right) h | \mathcal{F}_{t_i}\right].
\end{cases}
\tag{4.3}
$$

Note that (4.3) is defined just like the FBSDEs (2.1) and (2.2), where the X component is defined forwardly and the Y, Z components are defined backwardly. However, since we do not specify the terminal condition of Y_T^π, the system of Eq. (4.3) has infinitely many solutions. The following lemma gives an estimate of the difference between two such solutions.

Lemma 1 *For $j = 1, 2$, suppose $\left(\left\{X_{t_i}^{\pi,j}\right\}_{0 \leq i \leq N}, \left\{Y_{t_i}^{\pi,j}\right\}_{0 \leq i \leq N}, \left\{Z_{t_i}^{\pi,j}\right\}_{0 \leq i \leq N-1}\right)$ are two solutions of (4.3), with $X_{t_i}^{\pi,j}, Y_{t_i}^{\pi,j} \in L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P})$, $0 \leq i \leq N$. For any $\lambda_1 > 0, \lambda_2 \geq f_z$, and sufficiently small h, denote*

$$
\begin{aligned}
A_1 &:= 2k_b + \lambda_1 + \sigma_x + Kh, \\
A_2 &:= \left(\lambda_1^{-1} + h\right) b_y + \sigma_y, \\
A_3 &:= -\frac{\ln[1 - (2k_f + \lambda_2)h]}{h}, \\
A_4 &:= \frac{f_x}{[1 - (2k_f + \lambda_2)h]\lambda_2}.
\end{aligned}
\tag{4.4}
$$

Let $\delta X_i = X_{t_i}^{\pi,1} - X_{t_i}^{\pi,2}, \delta Y_i = Y_{t_i}^{\pi,1} - Y_{t_i}^{\pi,2}$, then we have, for $0 \leq n \leq N$,

$$
E|\delta X_n|^2 \leq A_2 \sum_{i=0}^{n-1} e^{A_1(n-i-1)h} E|\delta Y_i|^2 h,
$$

$$
E|\delta Y_n|^2 \leq e^{A_3(N-n)h} E|\delta Y_N|^2 + A_4 \sum_{i=n}^{N-1} e^{A_3(i-n)h} E|\delta X_i|^2 h.
$$

To prove Lemma 1, we need the following lemma to handle the Z component.

Lemma 2 *Let $0 \leq s_1 < s_2$, given $Q \in L^2(\Omega, \mathcal{F}_{s_2}, \mathbb{P})$, by the martingale representation theorem, there exists an $\mathcal{F}_t$-adapted process $\{H_s\}_{s_1 \leq s \leq s_2}$ such that $\int_{s_1}^{s_2} E|H_s|^2 ds < \infty$ and $Q = E[Q|\mathcal{F}_{s_1}] + \int_{s_1}^{s_2} H_s dW_s$. Then we have $E[Q(W_{s_2} - W_{s_1})|\mathcal{F}_{s_1}] = E[\int_{s_1}^{s_2} H_s ds|\mathcal{F}_{s_1}]$.*

Proof Consider the auxiliary stochastic process $Q_s = (E[Q|\mathcal{F}_{s_1}] + \int_{s_1}^{s} H_t \, dW_t)(W_s - W_{s_1})$ for $s \in [s_1, s_2]$. By Itô's formula,

$$dQ_s = (W_s - W_{s_1})H_s \, dW_s + \left(E[Q|\mathcal{F}_{s_1}] + \int_{s_1}^{s} H_t \, dW_t \right) dW_s + H_s \, ds.$$

Noting that $Q_{s_1} = 0$, we have

$$E[Q(W_{s_2} - W_{s_1})|\mathcal{F}_{s_1}] = E[Q_{s_2}|\mathcal{F}_{s_1}] = E\left[\int_{s_1}^{s_2} H_s \, ds | \mathcal{F}_{s_1} \right].$$

$\square$

Proof of Lemma 1 Let

$$\delta Z_i = Z_{t_i}^{\pi,1} - Z_{t_i}^{\pi,2},$$
$$\delta b_i = b\left(t_i, X_{t_i}^{\pi,1}, Y_{t_i}^{\pi,1}\right) - b\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,2}\right),$$
$$\delta \sigma_i = \sigma\left(t_i, X_{t_i}^{\pi,1}, Y_{t_i}^{\pi,1}\right) - \sigma\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,2}\right),$$
$$\delta f_i = f\left(t_i, X_{t_i}^{\pi,1}, Y_{t_i}^{\pi,1}, Z_{t_i}^{\pi,1}\right) - f\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,2}, Z_{t_i}^{\pi,2}\right).$$

Then we have

$$\delta X_{i+1} = \delta X_i + \delta b_i h + \delta \sigma_i \Delta W_i, \tag{4.5}$$

$$\delta Z_i = \frac{1}{h} E[\delta Y_{i+1} \Delta W_i | \mathcal{F}_{t_i}], \tag{4.6}$$

$$\delta Y_i = E[\delta Y_{i+1} + \delta f_i h | \mathcal{F}_{t_i}]. \tag{4.7}$$

By the martingale representation theorem, there exists an $\mathcal{F}_t$-adapted square-integrable process $\{\delta Z_t\}_{t_i \leq t \leq t_{i+1}}$ such that

$$\delta Y_{i+1} = E[\delta Y_{i+1}|\mathcal{F}_{t_i}] + \int_{t_i}^{t_{i+1}} (\delta Z_t)^{\mathrm{T}} \, dW_t,$$

or, equivalently,

$$\delta Y_{i+1} = \delta Y_i - \delta f_i h + \int_{t_i}^{t_{i+1}} (\delta Z_t)^{\mathrm{T}} \, dW_t. \tag{4.8}$$

From Eqs. (4.5) and (4.8), noting that δX_i, δY_i, δb_i, $\delta \sigma_i$, and δf_i are all $\mathcal{F}_{t_i}$-measurable, and $E[\Delta W_i|\mathcal{F}_{t_i}] = 0$, $E\left[\int_{t_i}^{t_{i+1}} (\delta Z_t)^{\mathrm{T}} \, dW_t | \mathcal{F}_{t_i}\right] = 0$, we have

$$E|\delta X_{i+1}|^2 = E|\delta X_i + \delta b_i h|^2 + E\left[(\Delta W_i)^{\mathrm{T}} (\delta \sigma_i)^{\mathrm{T}} \delta \sigma_i \Delta W_i\right]$$
$$= E|\delta X_i + \delta b_i h|^2 + h E\|\delta \sigma_i\|^2, \tag{4.9}$$

$$E|\delta Y_{i+1}|^2 = E|\delta Y_i - \delta f_i h|^2 + \int_{t_i}^{t_{i+1}} E|\delta Z_t|^2 \, dt. \tag{4.10}$$

From Eq. (4.9), by Assumptions 1 and 2 and the root-mean square and geometric mean inequality (RMS-GM inequality), for any $\lambda_1 > 0$, we have

$$
\begin{aligned}
& E|\delta X_{i+1}|^2 \\
= {}& E|\delta X_i|^2 + E|\delta b_i|^2 h^2 + h E \|\delta \sigma_i\|^2 \\
& + 2h E\left[\left(b\left(t_i, X_{t_i}^{\pi,1}, Y_{t_i}^{\pi,1}\right) - b\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,1}\right)\right)^{\mathrm{T}} \delta X_i\right] \\
& + 2h E\left[\left(b\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,1}\right) - b\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,2}\right)\right)^{\mathrm{T}} \delta X_i\right] \\
\leq {}& E|\delta X_i|^2 + \left(K E|\delta X_i|^2 + b_y E|\delta Y_i|^2\right) h^2 + 2k_b h E|\delta X_i|^2 \\
& + \left(\lambda_1 E|\delta X_i|^2 + \lambda_1^{-1} b_y E|\delta Y_i|^2\right) h + \left(\sigma_x E|\delta X_i|^2 + \sigma_y E|\delta Y_i|^2\right) h \\
= {}& [1 + (2k_b + \lambda_1 + \sigma_x + Kh)h]E|\delta X_i|^2 + \left[\left(\lambda_1^{-1} + h\right) b_y + \sigma_y\right] E|\delta Y_i|^2 h.
\end{aligned}
$$

Recall $A_1 = 2k_b + \lambda_1 + \sigma_x + Kh$, $A_2 = (\lambda_1^{-1} + h)b_y + \sigma_y$, $E|\delta X_0|^2 = 0$. By induction we can obtain that, for $0 \leq n \leq N$,

$$
E|\delta X_n|^2 \leq A_2 \sum_{i=0}^{n-1} e^{A_1(n-i-1)h} E|\delta Y_i|^2 h.
$$

Similarly, from Eq. (4.10), for any $\lambda_2 > 0$, we have

$$
\begin{aligned}
& E|\delta Y_{i+1}|^2 \\
\geq {}& E|\delta Y_i|^2 + \int_{t_i}^{t_{i+1}} E|\delta Z_t|^2 \, \mathrm{d}t \\
& - 2h E\left[\left(f\left(t_i, X_i^{1,\pi}, Y_i^{1,\pi}, Z_i^{1,\pi}\right) - f\left(t_i, X_i^{1,\pi}, Y_i^{2,\pi}, Z_i^{1,\pi}\right)\right)^{\mathrm{T}} \delta Y_i\right] \\
& - 2h E\left[\left(f\left(t_i, X_i^{1,\pi}, Y_i^{2,\pi}, Z_i^{1,\pi}\right) - f\left(t_i, X_i^{2,\pi}, Y_i^{2,\pi}, Z_i^{2,\pi}\right)\right)^{\mathrm{T}} \delta Y_i\right] \\
\geq {}& E|\delta Y_i|^2 + \int_{t_i}^{t_{i+1}} E|\delta Z_t|^2 \, \mathrm{d}t - 2k_f h E|\delta Y_i|^2 \\
& - \left[\lambda_2 E|\delta Y_i|^2 + \lambda_2^{-1}\left(f_x E|\delta X_i|^2 + f_z E|\delta Z_i|^2\right)\right] h.
\end{aligned}
$$

(4.11)

To deal with the integral term in (4.11), we apply Lemma 2 to (4.6) and (4.8) and get

$$
\delta Z_i = \frac{1}{h} E\left[\int_{t_i}^{t_{i+1}} \delta Z_t \, \mathrm{d}t \,\middle|\, \mathcal{F}_{t_i}\right],
$$

which implies, by the Cauchy inequality,

$$
\begin{aligned}
E|\delta Z_i|^2 h = \sum_{k=1}^{d} E|(\delta Z_i)_k|^2 h &= \sum_{k=1}^{d} \frac{1}{h} E \left| E \left[\int_{t_i}^{t_{i+1}} (\delta Z_t)_k \, dt | \mathcal{F}_{t_i} \right] \right|^2 \\
&\leq \sum_{k=1}^{d} \frac{1}{h} E \left| \int_{t_i}^{t_{i+1}} (\delta Z_t)_k \, dt \right|^2 \leq \sum_{k=1}^{d} \int_{t_i}^{t_{i+1}} E|(\delta Z_t)_k|^2 \, dt \\
&= \int_{t_i}^{t_{i+1}} E|\delta Z_t|^2 \, dt,
\end{aligned}
$$

where $(\cdot)_k$ denotes the k-th component of the vector. Plugging it into (4.11) gives us

$$
E|\delta Y_{i+1}|^2 \geq [1 - (2k_f + \lambda_2)h]E|\delta Y_i|^2 + \left(1 - f_z\lambda_2^{-1}\right) E|\delta Z_i|^2 h - f_x\lambda_2^{-1}E|\delta X_i|^2 h. \quad (4.12)
$$

Then for any $\lambda_2 \geq f_z$ and sufficiently small h satisfying $(2k_f + \lambda_2)h < 1$, we have

$$
E|\delta Y_i|^2 \leq [1 - (2k_f + \lambda_2)h]^{-1} \left[E|\delta Y_{i+1}|^2 + f_x\lambda_2^{-1}E|\delta X_i|^2 h \right].
$$

Recall $A_3 = -h^{-1}\ln[1 - (2k_f + \lambda_2)h]$, $A_4 = f_x\lambda_2^{-1}[1 - (2k_f + \lambda_2)h]^{-1}$. By induction we obtain that, for $0 \leq n \leq N$,

$$
E|\delta Y_n|^2 \leq e^{A_3(N-n)h} E|\delta Y_N|^2 + A_4 \sum_{i=n}^{N-1} e^{A_3(i-n)h} E|\delta X_i|^2 h.
$$

$\square$

Now we are ready to prove Theorem 1, whose precise statement is given below. Note that its conditions are satisfied if any of the five cases in the weak coupling and monotonicity conditions holds.

Theorem 1′ *Suppose Assumptions* 1, 2, 3, *and* 4 *hold true and there exist* $\lambda_1 > 0$, $\lambda_2 \geq f_z$ *such that* $\overline{A_0} < 1$, *where*

$$
\begin{aligned}
\overline{A_1} &:= 2k_b + \lambda_1 + \sigma_x, \\
\overline{A_2} &:= b_y\lambda_1^{-1} + \sigma_y, \\
\overline{A_3} &:= 2k_f + \lambda_2, \\
\overline{A_4} &:= f_x\lambda_2^{-1}, \\
\overline{A_0} &:= \overline{A_2}\frac{1 - e^{-(\overline{A_1}+\overline{A_3})T}}{\overline{A_1} + \overline{A_3}} \left\{ g_x e^{(\overline{A_1}+\overline{A_3})T} + \overline{A_4}\frac{e^{(\overline{A_1}+\overline{A_3})T} - 1}{\overline{A_1} + \overline{A_3}} \right\}.
\end{aligned}
\quad (4.13)
$$

Then there exists a constant $C > 0$, *depending on* $E|\xi|^2$, $\mathcal{L}$, T, λ_1, *and* λ_2, *such that for sufficiently small* h,

$$
\sup_{t \in [0,T]} \left(E|X_t - \hat{X}_t^\pi|^2 + E|Y_t - \hat{Y}_t^\pi|^2 \right) + \int_0^T E|Z_t - \hat{Z}_t^\pi|^2 \, dt \leq C \left[h + E|g\left(X_T^\pi\right) - Y_T^\pi|^2 \right],
$$

$$(4.14)$$

where $\hat{X}_t^\pi = X_{t_i}^\pi$, $\hat{Y}_t^\pi = Y_{t_i}^\pi$, $\hat{Z}_t^\pi = Z_{t_i}^\pi$ *for* $t \in [t_i, t_{i+1})$.

Remark 5 *The above theorem also implies the coercivity of the objective function* (2.4) *used in the deep BSDE method. Formally speaking, the coercivity means that if $\sum_{i=0}^{N-1} E|Z_{t_i}^\pi|^2 + E|Y_0^\pi|^2 \to +\infty$, we have $E|g\left(X_T^\pi\right) - Y_T^\pi|^2 \to +\infty$, which is a direct result from Theorem* 1'.

Remark 6 *If any of the weak coupling and monotonicity conditions introduced in Assumption* 3 *holds to a sufficient extent, there must exist λ_1, λ_2 satisfying the conditions in Theorem* 1'. *We discuss the 5 cases in what follows.*

1. *Suppose all other constants and $\lambda_1 > 0, \lambda_2 \geq f_z$ are fixed, if $T > 0$ is sufficiently small, then the second factor of $\overline{A_0}$ could be sufficiently close to 0 such that $\overline{A_0} < 1$.*
2. *Suppose all other constants and $\lambda_1 > 0, \lambda_2 \geq f_z$ are fixed, if $b_y \geq 0$ and $\sigma_y \geq 0$ are sufficiently small, then $\overline{A_2} \geq 0$ could be sufficiently small such that $\overline{A_0} < 1$.*
3. *Suppose all other constants and $\lambda_1 > 0, \lambda_2 \geq f_z$ are fixed, if $f_x \geq 0$ and $g_x \geq 0$ are sufficiently small, then $\overline{A_4}$ and thus the last factor in $\overline{A_0}$ could be sufficiently close to 0 such that $\overline{A_0} < 1$.*
4. *Suppose all constants except k_f and $\lambda_2 > 0$ are fixed. Let $\overline{A_1}' := \overline{A_1} + \overline{A_3} = 2k_b + 2k_f + \sigma_x + \lambda_1 + \lambda_2$ and rewrite $\overline{A_0}$ as*

$$\overline{A_0} = \overline{A_2}\left\{ g_x \frac{e^{\overline{A_1}'T} - 1}{\overline{A_1}'} + \overline{A_4}\frac{e^{\overline{A_1}'T} + e^{-\overline{A_1}'T} - 2}{\left(\overline{A_1}'\right)^2} \right\}.$$

It is straightforward to check that there exists a negative constant C_1 such that when $\overline{A_1}' \leq C_1$, $\left(e^{\overline{A_1}'T} - 1\right)/\overline{A_1}' < 1/\left(2\overline{A_2}g_x\right)$. By the definition of $\overline{A_1}'$, if k_f is sufficiently negative, there exists $\lambda_2 \geq f_x$ such that $\overline{A_1}' = C_1$ and λ_2 is sufficiently large to ensure

$$\overline{A_2}\,\overline{A_4}\frac{e^{C_1T} + e^{-C_1T} - 2}{C_1^2} = \frac{f_x\overline{A_2}\left(e^{C_1T} + e^{-C_1T} - 2\right)}{\lambda_2 C_1^2} < \frac{1}{2}.$$

Combining these two estimates gives $\overline{A_0} < 1$.
5. *Noting that k_b and k_f play the same role in $\overline{A_1}'$, we use the same argument as above to show that when k_b is sufficiently negative, there exists $\lambda_2 \geq f_x$ such that $\overline{A_0} < 1$.*

Proof From the proof of this theorem and throughout the remainder of the paper, we use C to generally denote a constant that only depends on $E|\xi|^2, \mathscr{L}$, and T, whose value may change from line to line when there is no need to distinguish. We also use $C(\cdot)$ to generally denote a constant depending on $E|\xi|^2, \mathscr{L}, T$ and the constants represented by $\cdot$.

We use the same notations as Lemma 1. Let $X_{t_i}^{\pi,1} = X_{t_i}^\pi, Y_{t_i}^{\pi,1} = Y_{t_i}^\pi, Z_{t_i}^{\pi,1} = Z_{t_i}^\pi$ (defined in system (2.3)) and $X_{t_i}^{\pi,2} = \overline{X}_{t_i}^\pi, Y_{t_i}^{\pi,2} = \overline{Y}_{t_i}^\pi, Z_{t_i}^{\pi,2} = \overline{Z}_{t_i}^\pi$ (defined in system (3.3)). It can be easily checked that both $\left(\left\{X_{t_i}^{\pi,j}\right\}_{0\leq i\leq N}, \left\{Y_{t_i}^{\pi,j}\right\}_{0\leq i\leq N}\right.$,

$\left\{ Z_{t_i}^{\pi,j} \right\}_{0 \le i \le N-1}$, $j = 1, 2$ satisfy the system of Eq. (4.3). Our proof strategy is to use Lemma 1 to bound the difference between two solutions through the objective function $E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2$. This allows us to apply Theorem 5 to derive the desired estimates.

To begin with, note that for any $\lambda_3 > 0$, the RMS-GM inequality yields

$$E|\delta Y_N|^2 = E|g\left(\overline{X}_T^{\pi}\right) - Y_T^{\pi}|^2 \le \left(1 + \lambda_3^{-1}\right) E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2 + g_x(1 + \lambda_3) E|\delta X_N|^2.$$

Let

$$P = \max_{0 \le n \le N} e^{-A_1 nh} E|\delta X_n|^2, \quad S = \max_{0 \le n \le N} e^{A_3 nh} E|\delta Y_n|^2.$$

Lemma 1 tells us

$$e^{-A_1 nh} E|\delta X_n|^2 \le A_2 \sum_{i=0}^{n-1} e^{-A_1(i+1)h} E|\delta Y_i|^2 h \le A_2 S \sum_{i=0}^{n-1} e^{-A_1(i+1)h - A_3 ih} h,$$

and

$$e^{A_3 nh} E|\delta Y_n|^2$$

$$\le e^{A_3 T} E|\delta Y_N|^2 + A_4 \sum_{i=n}^{N-1} e^{A_3 ih} E|\delta X_i|^2 h$$

$$\le e^{A_3 T} \left[\left(1 + \lambda_3^{-1}\right) E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2 + g_x(1 + \lambda_3) E|\delta X_N|^2 \right] + A_4 \sum_{i=n}^{N-1} e^{A_3 ih} E|\delta X_i|^2 h$$

$$\le e^{A_3 T} \left(1 + \lambda_3^{-1}\right) E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2 + \left[g_x(1 + \lambda_3) e^{(A_1 + A_3)T} + A_4 \sum_{i=n}^{N-1} e^{(A_1 + A_3)ih} h \right] P.$$

Therefore by definition of P and S, we have

$$P \le A_2 h e^{-A_1 h} \frac{e^{-(A_1 + A_3)T} - 1}{e^{-(A_1 + A_3)h} - 1} S,$$

$$S \le e^{A_3 T} \left(1 + \lambda_3^{-1}\right) E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2 + \left[g_x(1 + \lambda_3) e^{(A_1 + A_3)T} + A_4 h \frac{e^{(A_1 + A_3)T} - 1}{e^{(A_1 + A_3)h} - 1} \right] P.$$

Consider the function

$$A(h) = A_2 h e^{-A_1 h} \frac{e^{-(A_1 + A_3)T} - 1}{e^{-(A_1 + A_3)h} - 1} \left[g_x(1 + \lambda_3) e^{(A_1 + A_3)T} + A_4 h \frac{e^{(A_1 + A_3)T} - 1}{e^{(A_1 + A_3)h} - 1} \right].$$

When $A(h) < 1$, we have

$$P \le [1 - A(h)]^{-1} e^{A_3 T} \left(1 + \lambda_3^{-1}\right) A_2 h e^{-A_1 h} \frac{e^{-(A_1 + A_3)T} - 1}{e^{-(A_1 + A_3)h} - 1} E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2,$$

$$S \le [1 - A(h)]^{-1} e^{A_3 T} \left(1 + \lambda_3^{-1}\right) E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2.$$

Let

$$\overline{P} = \max_{0 \le n \le N} e^{-\overline{A_1} nh} E|\delta X_n|^2, \quad \overline{S} = \max_{0 \le n \le N} e^{\overline{A_3} nh} E|\delta Y_n|^2. \tag{4.15}$$

Recall

$$\lim_{h \to 0} A_i = \overline{A_i}, \quad i = 1, 2, 3, 4,$$

and note that

$$\lim_{h\to 0} A(h) = \overline{A_2}\frac{1 - e^{-(\overline{A_1}+\overline{A_3})T}}{\overline{A_1}+\overline{A_3}}\left[g_x(1+\lambda_3)e^{(\overline{A_1}+\overline{A_3})T} + \overline{A_4}\frac{e^{(\overline{A_1}+\overline{A_3})T}-1}{\overline{A_1}+\overline{A_3}}\right].$$

When $\overline{A_0} < 1$, comparing $\lim_{h\to 0} A(h)$ and $\overline{A_0}$, we know that, for any $\epsilon > 0$, there exists $\lambda_3 > 0$ and sufficiently small h such that

$$\overline{P} \le (1+\epsilon)\left[1 - \overline{A_0}\right]^{-1}\overline{A_2}e^{\overline{A_3}T}\left(1+\lambda_3^{-1}\right)\frac{1 - e^{-(\overline{A_1}+\overline{A_3})T}}{\overline{A_1}+\overline{A_3}}E|g\left(X_T^\pi\right) - Y_T^\pi|^2, \tag{4.16}$$

$$\overline{S} \le (1+\epsilon)\left[1 - \overline{A_0}\right]^{-1}e^{\overline{A_3}T}\left(1+\lambda_3^{-1}\right)E|g\left(X_T^\pi\right) - Y_T^\pi|^2. \tag{4.17}$$

By fixing $\epsilon = 1$ and choosing suitable λ_3, we obtain our error estimates of $E|\delta X_n|^2$ and $E|\delta Y_n|^2$ as

$$\max_{0\le n\le N} E|\delta X_n|^2 \le e^{\overline{A_1}T\vee 0}\overline{P} \le C(\lambda_1,\lambda_2)E|g\left(X_T^\pi\right) - Y_T^\pi|^2, \tag{4.18}$$

$$\max_{0\le n\le N} E|\delta Y_n|^2 \le e^{(-\overline{A_3}T)\vee 0}\overline{S} \le C(\lambda_1,\lambda_2)E|g\left(X_T^\pi\right) - Y_T^\pi|^2. \tag{4.19}$$

To estimate $E|\delta Z_n|^2$, we consider estimate (4.12), in which λ_2 can take any value no smaller than f_z. If $f_z \ne 0$, we choose $\lambda_2 = 2f_z$ and obtain

$$\frac{1}{2}E|\delta Z_i|^2 h \le \frac{f_x}{2f_z}E|\delta X_i|^2 h + E|\delta Y_{i+1}|^2 - [1 - (2k_f + 2f_z)h]E|\delta Y_i|^2.$$

Summing from 0 to $N-1$ gives us

$$\sum_{i=0}^{N-1} E|\delta Z_i|^2 h \le \frac{f_x T}{f_z}\max_{0\le n\le N} E|\delta X_n|^2 + [4(k_f + f_z)T \vee 0 + 2]\max_{0\le n\le N} E|\delta Y_n|^2$$

$$\le C(\lambda_1,\lambda_2)E|g\left(X_T^\pi\right) - Y_T^\pi|^2. \tag{4.20}$$

The case $f_z = 0$ can be dealt with similarly by choosing $\lambda_2 = 1$ and the same type of estimate can be derived. Finally, combining estimates (4.18), (4.19) and (4.20) with Theorem 5, we prove the statement in Theorem 1'. $\qquad\square$

5 An upper bound for the minimized objective function

We prove Theorem 2 in this section. We first state three useful lemmas. Theorem 2', as a detailed statement of Theorem 2, and Theorem 6, as an variation of Theorem 2' under stronger conditions, are then provided, followed by their proofs. The proofs of three lemmas are given at the end of the section.

The main process we analyze is (2.3). Lemma 3 gives an estimate of the final distance $E|g\left(X_T^\pi\right) - Y_T^\pi|^2$ provided by (2.3) in terms of the deviation between the approximated variables Y_0^π, $Z_{t_i}^\pi$ and the true solutions.

Lemma 3 *Suppose Assumptions* 1, 2, *and* 3 *hold true. Let* $X_T^\pi, Y_0^\pi, Y_T^\pi$, $\{Z_{t_i}^\pi\}_{0 \le i \le N-1}$ *be defined as in system* (2.3) *and* $\tilde{Z}_{t_i} = h^{-1} E\left[\int_{t_i}^{t_{i+1}} Z_t \, dt | \mathcal{F}_{t_i}\right]$. *Given* $\lambda_4 > 0$, *there exists a constant* $C > 0$ *depending on* $E|\xi|^2$, $\mathcal{L}$, T, *and* λ_4, *such that for sufficiently small* h,

$$E|g\left(X_T^\pi\right) - Y_T^\pi|^2 \le (1 + \lambda_4) H_{min} \sum_{i=0}^{N-1} E|\delta \tilde{Z}_{t_i}|^2 h + C\left[h + E|Y_0 - Y_0^\pi|^2\right],$$

where $\delta \tilde{Z}_{t_i} = \tilde{Z}_{t_i} - Z_{t_i}^\pi$, $H(x) = (1 + \sqrt{g_x})^2 e^{(2K + 2Kx^{-1} + x)T}\left(1 + f_z x^{-1}\right)$, *and* $H_{min} = \min_{x \in R^+} H(x)$.

Lemma 3 is close to Theorem 2, except that $\tilde{Z}_{t_i}$ is not a function of $X_{t_i}^\pi$ and $Y_{t_i}^\pi$ defined in (2.3). To bridge this gap, we need the following two lemmas. First, similar to the proof of Theorem 1', an estimate of the distance between the process defined in (2.3) and the process defined in (3.3) is also needed here. Lemma 4 is a general result to serve this purpose, providing an estimate of the difference between two backward processes driven by different forward processes.

Lemma 4 *Let* $X_{t_i}^{\pi,j} \in L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P})$ *for* $0 \le i \le N$, $j = 1, 2$. *Suppose* $\left\{Y_{t_i}^{\pi,j}\right\}_{0 \le i \le N}$ *and* $\left\{Z_{t_i}^{\pi,j}\right\}_{0 \le i \le N-1}$ *satisfy*

$$\begin{cases} Y_T^{\pi,j} = g\left(X_T^{\pi,j}\right), \\ Z_{t_i}^{\pi,j} = \dfrac{1}{h} E\left[Y_{t_{i+1}}^{\pi,j} \Delta W_i | \mathcal{F}_{t_i}\right], \\ Y_{t_i}^{\pi,j} = E\left[Y_{t_{i+1}}^{\pi,j} + f\left(t_i, X_{t_i}^{\pi,j}, Y_{t_i}^{\pi,j}, Z_{t_i}^{\pi,j}\right) h | \mathcal{F}_{t_i}\right], \end{cases} \tag{5.1}$$

for $0 \le i \le N-1$, $j = 1, 2$. *Let* $\delta X_i = X_{t_i}^{\pi,1} - X_{t_i}^{\pi,2}$, $\delta Z_i = Z_{t_i}^{\pi,1} - Z_{t_i}^{\pi,2}$, *then for any* $\lambda_7 > f_z$, *and sufficiently small* h, *we have*

$$\sum_{i=0}^{N-1} E|\delta Z_i|^2 h \le \frac{\lambda_7 (e^{-A_5 T} \vee 1)}{\lambda_7 - f_z} \left\{ g_x e^{A_5 T - A_5 h} E|\delta X_N|^2 + \frac{f_x}{\lambda_7} \sum_{i=0}^{N-1} e^{A_5 i h} E|\delta X_i|^2 h \right\},$$

where $A_5 := -h^{-1} \ln[1 - (2k_f + \lambda_7)h]$.

Lemma 5 shows that, similar to the nonlinear Feynman–Kac formula, the discrete stochastic process defined in (2.3) can also be linked to some deterministic functions.

Lemma 5 *Let* $\left\{X_{t_i}^\pi\right\}_{0 \le i \le N}$, $\left\{Y_{t_i}^\pi\right\}_{0 \le i \le N}$ *be defined in* (2.3). *When* $h < 1/\sqrt{K}$, *there exist deterministic functions* $U_i^\pi : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$, $V_i^\pi : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^d$ *for* $0 \le i \le N$ *such that* $Y_{t_i}^{\pi,'} = U_i^\pi\left(X_{t_i}^\pi, Y_{t_i}^\pi\right)$, $Z_{t_i}^{\pi,'} = V_i^\pi\left(X_{t_i}^\pi, Y_{t_i}^\pi\right)$ *satisfy*

$$\begin{cases} Y_{t_N}^{\pi,'} = g\left(X_{t_N}^{\pi}\right), \\ Z_{t_i}^{\pi,'} = \dfrac{1}{h} E\left[Y_{t_{i+1}}^{\pi,'} \Delta W_i | \mathcal{F}_{t_i}\right], \\ Y_{t_i}^{\pi,'} = E\left[Y_{t_{i+1}}^{\pi,'} + f\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi,'}, Z_{t_i}^{\pi,'}\right) h | \mathcal{F}_{t_i}\right], \end{cases} \quad (5.2)$$

for $0 \le i \le N - 1$. If b and σ are independent of y, then there exist deterministic functions $U_i^{\pi} : \mathbb{R}^m \to \mathbb{R}$, $V_i^{\pi} : \mathbb{R}^m \to \mathbb{R}^d$ for $0 \le i \le N$ such that $Y_{t_i}^{\pi,'} = U_i^{\pi}\left(X_{t_i}^{\pi}\right)$, $Z_{t_i}^{\pi,'} = V_i^{\pi}\left(X_{t_i}^{\pi}\right)$ satisfy (5.2).

Now we are ready to prove Theorem 2, with a precise statement given below. Like Theorem 1', the conditions below are satisfied if any of the five cases of the weak coupling and monotonicity conditions holds to certain extent.

Theorem 2′ *Suppose Assumptions 1, 2, 3, and 4 hold true. Given any $\lambda_1, \lambda_3 > 0$, $\lambda_2 \ge f_z$, and $\lambda_7 > f_z$, let $\overline{A_i}$, ($i = 1, 2, 3, 4$) be defined in (4.4) and*

$$\overline{A_5} := \lambda_7 + 2k_f,$$

$$\overline{A_0}' := \overline{A_2} \frac{1 - e^{-(\overline{A_1}+\overline{A_3})T}}{\overline{A_1}+\overline{A_3}} \left\{ g_x(1+\lambda_3)e^{(\overline{A_1}+\overline{A_3})T} + \overline{A_4} \frac{e^{(\overline{A_1}+\overline{A_3})T} - 1}{\overline{A_1}+\overline{A_3}} \right\},$$

$$\overline{B_0} := H_{min}\overline{A_2}e^{\overline{A_3}T} \frac{1 - e^{-(\overline{A_1}+\overline{A_3})T}}{\overline{A_1}+\overline{A_3}} \left[1 - \overline{A_0}'\right]^{-1} \left(1 + \lambda_3^{-1}\right) \quad (5.3)$$

$$\times \frac{\lambda_7\left(e^{-\overline{A_5}T} \vee 1\right)}{\lambda_7 - f_z} \left\{ g_x e^{(\overline{A_1}+\overline{A_5})T} + \frac{f_x}{\lambda_7} \frac{e^{(\overline{A_1}+\overline{A_5})T} - 1}{\overline{A_1}+\overline{A_5}} \right\}.$$

If there exist $\lambda_1, \lambda_2, \lambda_3, \lambda_7$ satisfying $\overline{A_0}' < 1$ and $\overline{B_0} < 1$, then there exists a constant C depending on $E|\xi|^2$, $\mathscr{L}$, T, $\lambda_1, \lambda_2, \lambda_3$, and λ_7, such that for sufficiently small h,

$$E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2 \le C \left\{ h + E|Y_0 - Y_0^{\pi}|^2 + \sum_{i=0}^{N-1} E|E\left[\tilde{Z}_{t_i} | X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right] - Z_{t_i}^{\pi}|^2 h \right\}, \quad (5.4)$$

where $\tilde{Z}_{t_i} = h^{-1}E\left[\int_{t_i}^{t_{i+1}} Z_t \, dt | \mathcal{F}_{t_i}\right]$. If Z_t is càdlàg, we can replace $\tilde{Z}_{t_i}$ with Z_{t_i}. If b and σ are independent of y, we can replace $E\left[\tilde{Z}_{t_i} | X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right]$ with $E\left[\tilde{Z}_{t_i} | X_{t_i}^{\pi}\right]$.

Remark 7 *If we take the infimum within the domains of Y_0^{π} and $Z_{t_i}^{\pi}$ on both sides, we recover the original statement in Theorem 2.*

Remark 8 *If any of the weak coupling and monotonicity conditions introduced in Assumption 3 holds to a sufficient extent, there must exist $\lambda_1, \lambda_2, \lambda_3, \lambda_7$ satisfying the conditions in Theorem 2'. The arguments are very similar to those provided in Remark 6. Hence, we omit the details here for the sake of brevity.*

Proof Using Lemma 3 with $\lambda_4 > 0$, we obtain

$$E|g(X_T^\pi) - Y_T^\pi|^2 \le (1+\lambda_4)H_{\min}\sum_{i=0}^{N-1} E|\delta\tilde{Z}_{t_i}|^2 h + C(\lambda_4)\left[h + E|Y_0 - Y_0^\pi|^2\right].$$
(5.5)

Splitting the term $\delta\tilde{Z}_{t_i} = \tilde{Z}_{t_i} - Z_{t_i}^\pi$ and applying the generalized mean inequality, we have (recall $\overline{Z}_{t_i}^\pi$ is defined in Theorem 5)

$$
\begin{aligned}
&E|\delta\tilde{Z}_{t_i}|^2 \\
&\le (1+\lambda_4)E|\overline{Z}_{t_i}^\pi - E\left[\overline{Z}_{t_i}^\pi | X_{t_i}^\pi, Y_{t_i}^\pi\right]|^2 \\
&\quad + (1+\lambda_4^{-1})\left\{E|\left(\tilde{Z}_{t_i} - \overline{Z}_{t_i}^\pi\right) - E\left[\left(\tilde{Z}_{t_i} - \overline{Z}_{t_i}^\pi\right) | X_{t_i}^\pi, Y_{t_i}^\pi\right]\right. \\
&\qquad\qquad \left. + (E[\tilde{Z}_{t_i}|X_{t_i}^\pi, Y_{t_i}^\pi] - Z_{t_i}^\pi)|^2\right\} \\
&\le (1+\lambda_4)E|\overline{Z}_{t_i}^\pi - E\left[\overline{Z}_{t_i}^\pi | X_{t_i}^\pi, Y_{t_i}^\pi\right]|^2 \\
&\quad + 3\left(1+\lambda_4^{-1}\right)\left\{E|\tilde{Z}_{t_i} - \overline{Z}_{t_i}^\pi|^2 + E|E\left[\left(\tilde{Z}_{t_i} - \overline{Z}_{t_i}^\pi\right) | X_{t_i}^\pi, Y_{t_i}^\pi\right]|^2\right. \\
&\qquad\qquad \left. + E|E\left[\tilde{Z}_{t_i}|X_{t_i}^\pi, Y_{t_i}^\pi\right] - Z_{t_i}^\pi|^2\right\} \\
&\le (1+\lambda_4)E|\overline{Z}_{t_i}^\pi - E\left[\overline{Z}_{t_i}^\pi | X_{t_i}^\pi, Y_{t_i}^\pi\right]|^2 \\
&\quad + 3\left(1+\lambda_4^{-1}\right)\left\{2E|\tilde{Z}_{t_i} - \overline{Z}_{t_i}^\pi|^2 + E|E\left[\tilde{Z}_{t_i}|X_{t_i}^\pi, Y_{t_i}^\pi\right] - Z_{t_i}^\pi|^2\right\}.
\end{aligned}
$$
(5.6)

From Eqs. (3.2) and (3.4), we know that

$$
\begin{aligned}
\sum_{i=0}^{N-1} E|\tilde{Z}_{t_i} - \overline{Z}_{t_i}^\pi|^2 h &\le 2\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} [E|Z_t - \tilde{Z}_{t_i}|^2 + E|Z_t - \overline{Z}_{t_i}^\pi|^2 \mathrm{d}t \\
&= 2\int_0^T [E|Z_t - \tilde{Z}_{t_i}|^2 + E|Z_t - \overline{Z}_{t_i}^\pi|^2 \mathrm{d}t \\
&\le C\left(1 + E|\xi|^2\right)h.
\end{aligned}
$$
(5.7)

Plugging estimates (5.6) and (5.7) into (5.5) gives us

$$
\begin{aligned}
&E|g(X_T^\pi) - Y_T^\pi|^2 \\
&\le (1+\lambda_4)^2 H_{\min}\sum_{i=0}^{N-1} E|\overline{Z}_{t_i}^\pi - E\left[\overline{Z}_{t_i}^\pi | X_{t_i}^\pi, Y_{t_i}^\pi\right]|^2 h \\
&\quad + C(\lambda_4)\left\{h + E|Y_0 - Y_0^\pi|^2 + \sum_{i=0}^{N-1} E|E\left[\tilde{Z}_{t_i}|X_{t_i}^\pi, Y_{t_i}^\pi\right] - Z_{t_i}^\pi|^2 h\right\}.
\end{aligned}
$$
(5.8)

It remains to estimate the term $\sum_{i=0}^{N-1} E|\overline{Z}_{t_i}^\pi - E\left[\overline{Z}_{t_i}^\pi | X_{t_i}^\pi, Y_{t_i}^\pi\right]|^2 h$, to which we intend to apply Lemma 4. Let $X_{t_i}^{\pi,1} = X_{t_i}^\pi$ and $X_{t_i}^{\pi,2} = \overline{X}_{t_i}^\pi$. The associated $Z_{t_i}^{\pi,1}$ and $Z_{t_i}^{\pi,2}$ are then defined according to Eq. 5.1. Note that $Z_{t_i}^{\pi,2} = \overline{Z}_{t_i}^\pi$ but $Z_{t_i}^{\pi,1}$ is

not necessarily equal to $Z_{t_i}^{\pi}$, due to the possible violation of the terminal condition. From Lemma 5, we know $Z_{t_i}^{\pi,1}$ can be represented as $V_i^{\pi}\left(X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right)$ with V_i^{π} being a deterministic function. By the property of conditional expectation, we have

$$E|\overline{Z}_{t_i}^{\pi} - E\left[\overline{Z}_{t_i}^{\pi}|X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right]|^2 \le E|\overline{Z}_{t_i}^{\pi} - V_i\left(X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right)|^2,$$

for any V_i. Therefore we have the estimate

$$\begin{aligned}
&\sum_{i=0}^{N-1} E|\overline{Z}_{t_i}^{\pi} - E\left[\overline{Z}_{t_i}^{\pi}|X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right]|^2 h \le \sum_{i=0}^{N-1} E|\delta Z_i|^2 h \\
&\le \frac{\lambda_7(e^{-A_5 T} \vee 1)}{\lambda_7 - f_z}\left\{g_x e^{A_5 T - A_5 h} E|\delta X_N|^2 + \frac{f_x}{\lambda_7}\sum_{i=0}^{N-1} e^{A_5 i h} E|\delta X_i|^2 h\right\}.
\end{aligned} \tag{5.9}$$

Recall that $\delta X_i = X_{t_i}^{\pi} - \overline{X}_{t_i}^{\pi}$, $\delta Z_i = Z_{t_i}^{\pi,1} - \overline{Z}_{t_i}^{\pi}$. Similar to the derivation of estimate (4.16) (using a given $\lambda_3 > 0$ without final specification) in the proof of Theorem 1', when $\overline{A_0}' < 1$, we have

$$\overline{P} \le (1 + \lambda_4)\overline{A_2}e^{\overline{A_3}T}\frac{1 - e^{-(\overline{A_1} + \overline{A_3})T}}{\overline{A_1} + \overline{A_3}}[1 - \overline{A_0}']^{-1}\left(1 + \lambda_3^{-1}\right) E|Y_T^{\pi} - g(X_T^{\pi})|^2, \tag{5.10}$$

in which $\overline{P} = \max_{0 \le n \le N} e^{-\overline{A_1} n h} E|\delta X_i|^2$. Plugging (5.10) into (5.9), and then into (5.8), we get

$$\sum_{i=0}^{N-1} E|\delta Z_i|^2 h \le \frac{\lambda_7\left(e^{-A_5 T} \vee 1\right)\overline{P}}{\lambda_7 - f_z}\left\{g_x e^{(\overline{A_1} + A_5)T - A_5 h} + \frac{f_x}{\lambda_7}\sum_{i=0}^{N-1} e^{(\overline{A_1} + A_5)i h} h\right\},$$

and

$$\begin{aligned}
&E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2 \\
&\le (1 + \lambda_4)^3 B(h) E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2 \\
&\quad + C(\lambda_4)\left\{h + E|Y_0 - Y_0^{\pi}|^2 + \sum_{i=0}^{N-1} E|E\left[\tilde{Z}_{t_i}|X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right] - Z_{t_i}^{\pi}|^2 h\right\},
\end{aligned} \tag{5.11}$$

for sufficiently small h. Here $B(h)$ is defined as

$$\begin{aligned}
B(h) = {}&H_{\min}\overline{A_2}e^{\overline{A_3}T}\frac{1 - e^{-(\overline{A_1} + \overline{A_3})T}}{\overline{A_1} + \overline{A_3}}\left[1 - \overline{A_0}'\right]^{-1}\left(1 + \lambda_3^{-1}\right) \\
&\times \frac{\lambda_7\left(e^{-A_5 T} \vee 1\right)}{\lambda_7 - f_z}\left\{g_x e^{(\overline{A_1} + A_5)T - A_5 h} + \frac{f_x}{\lambda_7}\sum_{i=0}^{N-1} e^{(\overline{A_1} + A_5)i h} h\right\}.
\end{aligned}$$

The forms of inequalities (5.4) and (5.11) are already very close. When $\lim_{h \to 0} B(h) = \overline{B_0} < 1$, there exists $\lambda_4 > 0$ such that for sufficiently small h, we have $1 - (1 + \lambda_4)^3 B(h) > \frac{1}{2}(1 - \overline{B_0})$. Rearranging the term $E|g\left(X_T^{\pi}\right) - Y_T^{\pi}|^2$ in inequality (5.11) yields our final estimate. $\qquad\square$

We shall briefly discuss how the universal approximation theorem can be applied based on Theorem 2'. For instance, Theorem 2.1 in Arora et al. (2018) states that every continuous and piecewise linear function with m-dimensional input can be represented by a deep neural network with rectified linear units and at most $\lceil 1 + \log_2(m + 1) \rceil$ depth. Now we view Y_0 as a target function with input ξ and $E\left[\tilde{Z}_{t_i} | X_{t_i}^\pi, Y_{t_i}^\pi\right]$ as another target function with input $\left(X_{t_i}^\pi, Y_{t_i}^\pi\right)$. Since $E|Y_0|^2 < +\infty$ and $E|E\left[\tilde{Z}_{t_i} | X_{t_i}^\pi, Y_{t_i}^\pi\right]|^2 \leq E|\tilde{Z}_{t_i}|^2 < +\infty$, we know that both target functions can be approximated in the L^2 sense by continuous and piecewise linear functions with arbitrary accuracy. Then the aforementioned statement implies that the two target functions can be approximated by two neural networks with rectified linear units and at most $\lceil 1 + \log_2(m + 1) \rceil$ depth, although the width might go to infinity as the approximation error decreases to 0. Therefore, according to Theorem 2', there exist good neural networks such that the value of the objective function is small.

Note that there still exist some concerns about the result in Theorem 2'. First, the function $E\left[\tilde{Z}_{t_i} | X_{t_i}^\pi, Y_{t_i}^\pi\right]$ changes when $Z_{t_j}^\pi$ changes for $j < i$. Second, the function may depend on $Y_{t_i}^\pi$. Even if the FBSDEs are decoupled so that the above two concerns do not exist, we know nothing a priori about the property of $E\left[\tilde{Z}_{t_i} | X_{t_i}^\pi, Y_{t_i}^\pi\right]$. In the next theorem, we replace $E\left[\tilde{Z}_{t_i} | X_{t_i}^\pi, Y_{t_i}^\pi\right]$ with $\sigma^{\mathrm{T}}\left(t_i, X_{t_i}^\pi, u\left(t_i, X_{t_i}^\pi\right)\right) \nabla_x u\left(t_i, X_{t_i}^\pi\right)$, which can resolve these problems. However, meanwhile we require more regularity for the coefficients of the FBSDEs.

Theorem 6 *Suppose Assumptions 1, 2, 3 and 4 and the assumptions in Theorem 3 hold true. Let u be the solution of corresponding quasilinear PDEs (2.6) and L be the squared Lipschitz constant of $\sigma^{\mathrm{T}}(t, x, u(t, x)) \nabla_x u(t, x)$ with respect to x. With the same notations of Theorem 2', when $\overline{A_0}' < 1$ and*

$$\overline{B_0}' := H_{min} L \overline{A_2} e^{\overline{A_3}T} \frac{\left(e^{\overline{A_1}T} - 1\right)\left(1 - e^{-(\overline{A_1} + \overline{A_3})T}\right)}{\overline{A_1}(\overline{A_1} + \overline{A_3})} \left[1 - \overline{A_0}'\right]^{-1}\left(1 + \lambda_3^{-1}\right) < 1,$$

there exists a constant $C > 0$ depending on $E|\xi|^2$, T, $\mathscr{L}$, L, λ_1, λ_2, and λ_3, such that for sufficiently small h,

$$E|g(X_T^\pi) - Y_T^\pi|^2 \leq C\left\{h + E|Y_0 - Y_0^\pi|^2 + \sum_{i=0}^{N-1} E|f_i\left(X_{t_i}^\pi\right) - Z_{t_i}^\pi|^2 h\right\}, \quad (5.12)$$

where $f_i(x) = \sigma^{\mathrm{T}}(t_i, x, u(t_i, x)) \nabla_x u(t_i, x)$.

Proof By Theorem 3, we have $Z_{t_i} = f_i(X_{t_i})$, in which X_t is the solution of

$$X_t = \xi + \int_0^t b(s, X_s, u(s, X_s)) \, \mathrm{d}s + \int_0^t \sigma(s, X_s, u(s, X_s)) \, \mathrm{d}W_s.$$

Using Lemma 3 again with $\lambda_4 > 0$ gives us

$$E|g\left(X_T^\pi\right) - Y_T^\pi|^2 \le (1 + \lambda_4) H_{\min} \sum_{i=0}^{N-1} E|\delta \tilde{Z}_{t_i}|^2 h + C(\lambda_4) \left[h + E|Y_0 - Y_0^\pi|^2\right].$$

Given the continuity of $\sigma^{\mathrm{T}}(t, x, u(t, x)) \nabla_x u(t, x)$, we know Z_t admits a continuous version. Hence the term $\tilde{Z}_{t_i}$ in $\delta \tilde{Z}_{t_i} = \tilde{Z}_{t_i} - Z_{t_i}^\pi$ can be replaced with Z_{t_i}, i.e.,

$$E|g\left(X_T^\pi\right) - Y_T^\pi|^2 \le (1 + \lambda_4) H_{\min} \sum_{i=0}^{N-1} E|Z_{t_i} - Z_{t_i}^\pi|^2 h + C(\lambda_4) \left[h + E|Y_0 - Y_0^\pi|^2\right].$$

$$(5.13)$$

Similar to the arguments in inequalities (5.6) and (5.7), we have

$$
\begin{aligned}
& E|Z_{t_i} - Z_{t_i}^\pi|^2 \\
&\le \left(1 + \lambda_4^{-1}\right) E|f_i\left(X_{t_i}^\pi\right) - Z_{t_i}^\pi|^2 + (1 + \lambda_4) E|Z_{t_i} - f_i(X_{t_i}^\pi)|^2 \\
&\le \left(1 + \lambda_4^{-1}\right) E|f_i\left(X_{t_i}^\pi\right) - Z_{t_i}^\pi|^2 + (1 + \lambda_4) L E|X_{t_i} - X_{t_i}^\pi|^2 \\
&\le \left(1 + \lambda_4^{-1}\right) E|f_i\left(X_{t_i}^\pi\right) - Z_{t_i}^\pi|^2 \\
&\quad + (1 + \lambda_4) L \left[(1 + \lambda_4) E|X_{t_i}^\pi - \overline{X}_{t_i}^\pi|^2 + \left(1 + \lambda_4^{-1}\right) E|X_{t_i} - \overline{X}_{t_i}^\pi|^2\right] \\
&\le (1 + \lambda_4)^2 L E|X_{t_i}^\pi - \overline{X}_{t_i}^\pi|^2 + C(L, \lambda_4) \left\{E|f_i\left(X_{t_i}^\pi\right) - Z_{t_i}^\pi|^2 + h\right\},
\end{aligned}
$$

where the last equality uses the convergence result (3.4). Plugging it into (5.13), we have

$$
\begin{aligned}
E|g(X_T^\pi) - Y_T^\pi|^2 &\le (1 + \lambda_4)^3 H_{\min} L \sum_{i=0}^{N-1} E|X_{t_i}^\pi - \overline{X}_{t_i}^\pi|^2 h \\
&\quad + C(L, \lambda_4) \left\{h + E|Y_0 - Y_0^\pi|^2 + \sum_{i=0}^{N-1} E|f_i\left(X_{t_i}^\pi\right) - Z_{t_i}^\pi|^2 h\right\}
\end{aligned}
$$

$$(5.14)$$

for sufficiently small h.

We employ the estimate (5.10) again to rewrite inequality (5.14) as

$$
\begin{aligned}
E|g\left(X_T^\pi\right) - Y_T^\pi|^2 &\le (1 + \lambda_4)^4 \widetilde{B}(h) E|g\left(X_T^\pi\right) - Y_T^\pi|^2 \\
&\quad + C(L, \lambda_4) \left\{h + E|Y_0 - Y_0^\pi|^2 + \sum_{i=0}^{N-1} E|f_i\left(X_{t_i}^\pi\right) - Z_{t_i}^\pi|^2 h\right\},
\end{aligned}
$$

$$(5.15)$$

where

$$\widetilde{B}(h) = H_{\min} L \overline{A_2} e^{\overline{A_3} T} \frac{1 - e^{-(\overline{A_1} + \overline{A_3}) T}}{\overline{A_1} + \overline{A_3}} \left[1 - \overline{A_0}'\right]^{-1} \left(1 + \lambda_3^{-1}\right) \sum_{i=0}^{N-1} e^{i \overline{A_1} h} h.$$

Arguing in the same way as that in the proof of Theorem 2', when $\tilde{B}(h)$ is strictly bounded above by 1 for sufficiently small h, we can choose λ_4 small enough and rearrange the terms in inequality (5.15) to obtain the result in inequality (5.12). $\qquad\square$

Remark 9 *The Lipschitz constant used in Theorem 6 may be further estimated a priori. Denote the Lipschitz constant of function f with respect to x as $L_x(f)$, and the bound of function f as $M(f)$. Loosely speaking, we have*

$$L_x(\sigma^{\mathrm{T}}(t, x, u(t, x))\nabla_x u(t, x)) \leq M(\sigma)L_x(\nabla_x u) + M(\nabla_x u)[L_x(\sigma) + L_y(\sigma)L_x(u)].$$

Here $L_x(u) = M(\nabla_x u(t, x))$ can be estimated from the first point of Theorem 4 and $L(\nabla_x u(t, x)) = M(\nabla_{xx} u)$ can be estimated through the Schauder estimate (see, e.g., (Ma and Yong 2007, Chapter 4, Lemma 2.1)). Note that the resulting estimate may depend on the dimension d.

5.1 Proof of Lemmas

Proof of Lemma 3 We construct continuous processes X_t^{π}, Y_t^{π} as follows. For $t \in [t_i, t_{i+1})$, let

$$X_t^{\pi} = X_{t_i}^{\pi} + b\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right)(t - t_i) + \sigma\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right)(W_t - W_{t_i}),$$

$$Y_t^{\pi} = Y_{t_i}^{\pi} - f\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}, Z_{t_i}^{\pi}\right)(t - t_i) + \left(Z_{t_i}^{\pi}\right)^{\mathrm{T}}(W_t - W_{t_i}).$$

From system (2.3), we see this definition also works at t_{i+1}. We are interested in again the estimates of the following terms

$$\delta X_t = X_t - X_t^{\pi}, \quad \delta Y_t = Y_t - Y_t^{\pi}, \quad \delta Z_t = Z_t - Z_{t_i}^{\pi}, \quad t \in [t_i, t_{i+1}).$$

For $t \in [t_i, t_{i+1})$, let

$$\delta b_t = b(t, X_t, Y_t) - b\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right),$$
$$\delta \sigma_t = \sigma(t, X_t, Y_t) - \sigma\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}\right),$$
$$\delta f_t = f(t, X_t, Y_t, Z_t) - f\left(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}, Z_{t_i}^{\pi}\right).$$

By definition,

$$d(\delta X_t) = \delta b_t \, dt + \delta \sigma_t \, dW_t,$$
$$d(\delta Y_t) = -\delta f_t \, dt + (\delta Z_t)^{\mathrm{T}} \, dW_t.$$

Then by Itô's formula, we have

$$d|\delta X_t|^2 = [2(\delta b_t)^{\mathrm{T}}\delta X_t + \|\delta \sigma_t\|^2]dt + 2(\delta X_t)^{\mathrm{T}}\delta \sigma_t \, dW_t,$$
$$d|\delta Y_t|^2 = [-2(\delta f_t)^{\mathrm{T}}\delta Y_t + |\delta Z_t|^2] \, dt + 2\delta Y_t(\delta Z_t)^{\mathrm{T}} \, dW_t.$$

Thus,

$$E|\delta X_t|^2 = E|\delta X_{t_i}|^2 + \int_{t_i}^{t} E\left[2(\delta b_s)^{\mathrm{T}}\delta X_s + \|\delta \sigma_s\|^2\right] ds,$$

$$E|\delta Y_t|^2 = E|\delta Y_{t_i}|^2 + \int_{t_i}^{t} E\left[-2(\delta f_s)^{\mathrm{T}}\delta Y_s + |\delta Z_s|^2\right] ds.$$

For any $\lambda_5, \lambda_6 > 0$, using Assumptions 1 and 2 and the RMS-GM inequality, we have

$$
\begin{aligned}
E|\delta X_t|^2 \\
\leq E|\delta X_{t_i}|^2 &+ \int_{t_i}^t [\lambda_5 E|\delta X_s|^2 + \lambda_5^{-1} E|\delta b_s|^2 + E\|\delta\sigma_s\|^2]\,\mathrm{d}s \\
\leq E|\delta X_{t_i}|^2 &+ \lambda_5 \int_{t_i}^t E|\delta X_s|^2\,\mathrm{d}s + \int_{t_i}^t K\left(\lambda_5^{-1}+1\right)|s-t_i|\,\mathrm{d}s \\
&+ \int_{t_i}^t \left[(K\lambda_5^{-1}+\sigma_x)E|X_s - X_{t_i}^\pi|^2 + \left(b_y\lambda_5^{-1}+\sigma_y\right) E|Y_s - Y_{t_i}^\pi|^2 \right]\,\mathrm{d}s.
\end{aligned}
\tag{5.16}
$$

By the RMS-GM inequality, we also have

$$
E|X_s - X_{t_i}^\pi|^2 \leq (1+\epsilon_1)E|\delta X_{t_i}|^2 + \left(1+\epsilon_1^{-1}\right)E|X_s - X_{t_i}|^2,
\tag{5.17}
$$

$$
E|Y_s - Y_{t_i}^\pi|^2 \leq (1+\epsilon_2)E|\delta Y_{t_i}|^2 + \left(1+\epsilon_2^{-1}\right)E|Y_s - Y_{t_i}|^2,
\tag{5.18}
$$

in which we choose $\epsilon_1 = \lambda_6\left(K\lambda_5^{-1}+\sigma_x\right)^{-1}$ and $\epsilon_2 = \lambda_6\left(b_y\lambda_5^{-1}+\sigma_y\right)^{-1}$. The path regularity in Theorem 4 tells us

$$
\sup_{s\in[t_i,t_{i+1}]} \left(E|X_s - X_{t_i}|^2 + E|Y_s - Y_{t_i}|^2 \right) \leq Ch.
\tag{5.19}
$$

Plugging inequalities (5.17), (5.18), (5.19) into (5.16) with simplification, we obtain

$$
\begin{aligned}
E|\delta X_t|^2 \leq {}& [1+\left(K\lambda_5^{-1}+\sigma_x+\lambda_6\right)h]E|\delta X_{t_i}|^2 + \lambda_5 \int_{t_i}^t E|\delta X_s|^2\,\mathrm{d}s \\
&+ \left(b_y\lambda_5^{-1}+\sigma_y+\lambda_6\right) E|\delta Y_{t_i}|^2 h + C(\lambda_5,\lambda_6)h^2.
\end{aligned}
\tag{5.20}
$$

Then, by Grönwall inequality, we have

$$
\begin{aligned}
E|\delta X_{t_{i+1}}|^2 \\
\leq {}& e^{\lambda_5 h}\left\{\left[1+\left(K\lambda_5^{-1}+\sigma_x+\lambda_6\right)h\right]E|\delta X_{t_i}|^2 \right. \\
&\left. + \left(b_y\lambda_5^{-1}+\sigma_y+\lambda_6\right)E|\delta Y_{t_i}|^2 h + C(\lambda_5,\lambda_6)h^2\right\} \\
\leq {}& e^{A_6 h}E|\delta X_{t_i}|^2 + e^{\lambda_5 h}\left(b_y\lambda_5^{-1}+\sigma_y+\lambda_6\right)E|\delta Y_{t_i}|^2 h + C(\lambda_5,\lambda_6)h^2 \\
\leq {}& e^{A_6 h}E|\delta X_{t_i}|^2 + A_7 E|\delta Y_{t_i}|^2 h + C(\lambda_5,\lambda_6)h^2,
\end{aligned}
\tag{5.21}
$$

where $A_6 := K\lambda_5^{-1}+\sigma_x+\lambda_5+\lambda_6$, $A_7 := b_y\lambda_5^{-1}+\sigma_y+2\lambda_6$, and h is sufficiently small.

Similarly, with the same type of estimates in (5.16) and (5.20), for any $\lambda_5, \lambda_6 > 0$, we have

$$E|\delta Y_t|^2$$
$$\leq E|\delta Y_{t_i}|^2 + \int_{t_i}^t \left[\lambda_5 E|\delta Y_s|^2 + \lambda_5^{-1} E|\delta f_s|^2 + E|\delta Z_s|^2 \right] ds$$
$$\leq E|\delta Y_{t_i}|^2 + \lambda_5 \int_{t_i}^t E|\delta Y_s|^2 \, ds + \int_{t_i}^t K\lambda_5^{-1}|s - t_i| \, ds$$
$$+ \int_{t_i}^t \lambda_5^{-1} \left[f_x E|X_s - X_{t_i}^\pi|^2 + K E|Y_s - Y_{t_i}^\pi|^2 \right] ds + \left(1 + f_z \lambda_5^{-1} \right) \int_{t_i}^t E|\delta Z_s|^2 \, ds$$
$$\leq \left[1 + \left(K\lambda_5^{-1} + \lambda_6 \right) h \right] E|\delta Y_{t_i}|^2 + \lambda_5 \int_{t_i}^t E|\delta Y_s|^2 \, ds + \left(f_x \lambda_5^{-1} + \lambda_6 \right) E|\delta X_{t_i}^\pi|^2 h$$
$$+ \left(1 + f_z \lambda_5^{-1} \right) \int_{t_i}^t E|\delta Z_s|^2 \, ds + C(\lambda_5, \lambda_6)h^2.$$

Arguing in the same way of (5.21), by Grönwall inequality, for sufficiently small h, we have

$$E|\delta Y_{t_{i+1}}|^2$$
$$\leq e^{A_8 h} E|\delta Y_{t_i}|^2 + A_9 E|\delta X_{t_i}|^2 h + (1 + f_z \lambda_5^{-1} + \lambda_6) \int_{t_i}^t E|\delta Z_s|^2 \, ds + C(\lambda_5, \lambda_6)h^2,$$

with $A_8 := K\lambda_5^{-1} + \lambda_5 + \lambda_6$, $A_9 := f_x \lambda_5^{-1} + 2\lambda_6$. Choosing $\epsilon_3 = \left(1 + f_z \lambda_5^{-1} + \lambda_6 \right)^{-1} \lambda_6$ and using

$$\int_{t_i}^{t_{i+1}} E|\delta Z_t|^2 \, dt \leq (1 + \epsilon_3) E|\delta \tilde{Z}_{t_i}|^2 h + \left(1 + \epsilon_3^{-1} \right) E_z^i,$$

where $\delta \tilde{Z}_{t_i} = \tilde{Z}_{t_i} - Z_{t_i}^\pi$ and $E_z^i = \int_{t_i}^{t_{i+1}} E|Z_t - \tilde{Z}_{t_i}|^2 \, dt$, we furthermore obtain

$$E|\delta Y_{t_{i+1}}|^2 \leq e^{A_8 h} E|\delta Y_{t_i}|^2 + A_9 E|\delta X_{t_i}|^2 h + A_{10} E|\delta \tilde{Z}_{t_i}|^2 h + C(\lambda_5, \lambda_6) \left(h^2 + E_z^i \right), \tag{5.22}$$

with $A_{10} := 1 + f_z \lambda_5^{-1} + 2\lambda_6$.

Define
$$M_i = \max \left\{ E|\delta X_i|^2, E|\delta Y_i|^2 \right\}, \quad 0 \leq i \leq N.$$

Combining inequalities (5.21) and (5.22) together yields

$$M_{i+1}$$
$$\leq \left(e^{\max\{A_6, A_8\}h} + \max\{A_7, A_9\}h \right) M_i + A_{10} E|\delta \tilde{Z}_{t_i}|^2 h + C(\lambda_5, \lambda_6) \left(h^2 + E_z^i \right)$$
$$\leq e^{(\max\{A_6, A_8\} + \max\{A_7, A_9\})h} M_i + A_{10} E|\delta \tilde{Z}_{t_i}|^2 h + C(\lambda_5, \lambda_6) \left(h^2 + E_z^i \right).$$

Letting $A_{11} := \max\{A_6, A_8\} + \max\{A_7, A_9\}$, we have

$$M_{i+1} \leq e^{A_{11} h} M_i + A_{10} E|\delta \tilde{Z}_{t_i}|^2 h + C(\lambda_5, \lambda_6) \left(h^2 + E_z^i \right). \tag{5.23}$$

We start from $M_0 = E|Y_0 - Y_0^\pi|^2$ and apply inequality (5.23) repeatedly to obtain

$$M_N \le A_{10} e^{A_{11} T} \sum_{i=0}^{N-1} E|\delta \tilde{Z}_{t_i}|^2 h + C(\lambda_5, \lambda_6) \left[h + E|Y_0 - Y_0^\pi|^2 \right], \qquad (5.24)$$

in which for the last term we use the fact $\sum_{i=0}^{N-1} E_z^i \le Ch$ from inequality (3.2). Note that

$$A_{10} = 1 + f_z \lambda_5^{-1} + 2\lambda_6,$$
$$A_{11} \le 2K + 2K\lambda_5^{-1} + \lambda_5 + 3\lambda_6.$$

Given any $\lambda_4 > 0$, we can choose λ_6 small enough such that

$$\left(1 + f_z \lambda_5^{-1} + 2\lambda_6 \right) e^{A_{11} T} \le (1 + \lambda_4) \left(1 + f_z \lambda_5^{-1} \right) e^{\left(2K + 2K\lambda_5^{-1} + \lambda_5 \right) T}.$$

This condition and inequality (5.24) together give us

$$
\begin{aligned}
M_N \le{} & (1 + \lambda_4) \left(1 + f_z \lambda_5^{-1} \right) e^{\left(2K + 2K\lambda_5^{-1} + \lambda_5 \right) T} \sum_{i=0}^{N-1} E|\delta \tilde{Z}_{t_i}|^2 h \\
& + C(\lambda_4, \lambda_5) \left[h + E|Y_0 - Y_0^\pi|^2 \right].
\end{aligned}
\qquad (5.25)
$$

Finally, by decomposing the objective function, we have

$$
\begin{aligned}
& E|g\left(X_T^\pi\right) - Y_T^\pi|^2 \\
={} & E|g\left(X_T^\pi\right) - g(X_T) + Y_T - Y_T^\pi|^2 \\
\le{} & \left(1 + (\sqrt{g_x})^{-1} \right) E|g\left(X_T^\pi\right) - g(X_T)|^2 + (1 + \sqrt{g_x}) E|\delta Y_N|^2 \\
\le{} & (g_x + \sqrt{g_x}) E|\delta X_N|^2 + (1 + \sqrt{g_x}) E|\delta Y_N|^2 \\
\le{} & (1 + \sqrt{g_x})^2 M_N.
\end{aligned}
\qquad (5.26)
$$

We complete our proof by combining inequalities (5.25), (5.26) and choosing $\lambda_5 = \text{argmin}_{x \in \mathbb{R}^+} H(x)$. □

Proof of Lemma 4 We use the same notations as in the proof of Lemma 1. As derived in (4.12), for any $\lambda_7 > f_z \ge 0$, we have

$$E|\delta Y_{i+1}|^2 \ge [1 - (2k_f + \lambda_7)h]E|\delta Y_i|^2 + \left(1 - f_z \lambda_7^{-1} \right) E|\delta Z_i|^2 h - f_x \lambda_7^{-1} E|\delta X_i|^2 h. \qquad (5.27)$$

Multiplying both sides of (5.27) by $e^{A_5 ih}(e^{-A_5 T} \vee 1)/ \left(1 - f_z \lambda_7^{-1} \right)$ gives us

$$
\begin{aligned}
& \frac{\lambda_7 \left(e^{-A_5 T} \vee 1 \right)}{\lambda_7 - f_z} \left\{ e^{A_5 ih} E|\delta Y_{i+1}|^2 - e^{A_5 (i-1)h} E|\delta Y_i|^2 + e^{A_5 ih} \frac{f_x}{\lambda_7} E|\delta X_i|^2 h \right\} \\
& \ge e^{A_5 ih} \left(e^{-A_5 T} \vee 1 \right) E|\delta Z_i|^2 h \\
& \ge E|\delta Z_i|^2 h.
\end{aligned}
\qquad (5.28)
$$

Summing (5.28) up from $i = 0$ to $N - 1$, we obtain

$$\sum_{i=0}^{N-1} E|\delta Z_i|^2 h \leq \frac{\lambda_7 \left(e^{-A_5 T} \vee 1\right)}{\lambda_7 - f_z} \left\{ e^{A_5 T - A_5 h} E|\delta Y_N|^2 + \frac{f_x}{\lambda_7} \sum_{i=0}^{N-1} e^{A_5 i h} E|\delta X_i|^2 h \right\}.$$
(5.29)

Note that $E|\delta Y_N|^2 \leq g_x E|\delta X_N|^2$ by Assumption 1. Plugging it into (5.29), we arrive at the desired result. □

Proof of Lemma 5 We prove by induction backwardly. Let $Z_{t_N}^{\pi,'} = 0$ for convenience. It is straightforward to see that the statement holds for $i = N$. Assume the statement holds for $i = k + 1$. For $i = k$, we know $Y_{t_{k+1}}^{\pi,'} = U_{k+1}\left(X_{t_{k+1}}^{\pi}, Y_{t_{k+1}}^{\pi}\right)$. Recalling the definition of $\left\{X_{t_i}^{\pi}\right\}_{0 \leq i \leq N}$, $\left\{Y_{t_i}^{\pi}\right\}_{0 \leq i \leq N}$ in (2.3), we can rewrite $Y_{t_{k+1}}^{\pi,'} = \overline{U}_k\left(X_{t_k}^{\pi}, Y_{t_k}^{\pi}, \Delta W_k\right)$, with $\overline{U}_k : \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ being a deterministic function. Note $Z_{t_k}^{\pi,'} = h^{-1} E\left[\overline{U}_k\left(X_{t_k}^{\pi}, Y_{t_k}^{\pi}, \Delta W_k\right) \Delta W_k | \mathcal{F}_{t_k}\right]$. Since ΔW_k is independent of $\mathcal{F}_{t_k}$, there exists a deterministic function $V_k^{\pi} : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^d$ such that $Z_{t_k}^{\pi,'} = V_k^{\pi}\left(X_{t_k}^{\pi}, Y_{t_k}^{\pi}\right)$.

Next we consider $Y_{t_k}^{\pi,'}$. Let $H_k = L^2(\Omega, \sigma(X_{t_k}^{\pi}, Y_{t_k}^{\pi}), \mathbb{P})$, where $\sigma\left(X_{t_k}^{\pi}, Y_{t_k}^{\pi}\right)$ denotes the σ-algebra generated by $X_{t_k}^{\pi}, Y_{t_k}^{\pi}$. We know H_k is a Banach space and another equivalent representation is

$$H_k = \left\{Y = \phi\left(X_{t_k}^{\pi}, Y_{t_k}^{\pi}\right) \mid \phi \text{ is measurable and } E|Y|^2 < \infty\right\}.$$

Consider the following map defined on H_k:

$$\Phi_k(Y) = E\left[Y_{t_{k+1}}^{\pi,'} + f\left(t_k, X_{t_k}^{\pi}, Y, Z_{t_k}^{\pi,'}\right) h | \mathcal{F}_{t_k}\right].$$

By Assumption 3, $\Phi_k(Y)$ is square-integrable. Furthermore, following the same argument for $Z_{t_k}^{\pi,'}$, $\Phi_k(Y)$ can also be represented as a deterministic function of $X_{t_k}^{\pi}, Y_{t_k}^{\pi}$. Hence, $\Phi_k(Y) \in H_k$. Note that Assumption 1 implies $E|\Phi_k(Y_1) - \Phi_k(Y_2)|^2 \leq Kh^2 E|Y_1 - Y_2|^2$. Therefore Φ_k is a contraction map on H_k when $h < 1/\sqrt{K}$. By the Banach fixed-point theorem, there exists a unique fixed-point $Y^* = \phi_k^*\left(X_{t_k}^{\pi}, Y_{t_k}^{\pi}\right) \in H_k$ satisfying $Y^* = \Phi_k(Y^*)$. We choose $U_k^{\pi} = \phi_k^*$ to validate the statement for $Y_{t_k}^{\pi,'}$.

When b and σ are independent of y, all of the arguments above can be made similarly with U_i^{π}, V_i^{π} also being independent of Y. □

6 Numerical examples

6.1 General setting

In this section, we illustrate the proposed numerical scheme by solving two high-dimensional coupled FBSDEs adapted from the literature. The common setting for these two numerical examples is as follows. We assume $d = m = 100$, that

is, $X_t, Z_t, W_t \in \mathbb{R}^{100}$. Assume ξ is deterministic and we are interested in the approximation error of Y_0, which is also a deterministic scalar.

We use $N - 1$ fully-connected feedforward neural networks to parameterize ϕ_i^π, $i = 0, 1, \ldots, N-1$. Each of the neural networks has 2 hidden layers with dimension $d + 10$. The input has dimension $d + 1$ ($X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$) and the output has dimension d. In practice, one can of course choose X_i only as the input. We additionally test this input for the two examples and find no difference in terms of the relative error of Y_0 (up to second decimal place). We use the rectifier function (ReLU) as the activation function and adopt batch normalization (Ioffe and Szegedy 2015) right after each matrix multiplication and before activation. We employ the Adam optimizer (Kingma and Ba 2015) to optimize the parameters with batch-size being 64. The loss function is computed based on 256 validation sample paths. We initialize all the parameters using a uniform or normal distribution and run each experiment 5 times to report the average result.

6.2 Example 1

The first problem is adapted from (Milstein and Tretyakov 2006; Huijskens et al. 2016), in which the original spatial dimension of the problem is 1. We consider the following coupled FBSDEs

$$
\begin{cases}
X_{j,t} = x_{j,0} + \int_0^t \dfrac{X_{j,s}\left(1 + X_{j,s}^2\right)}{(2 + X_{j,s})^3}\, ds \\
\qquad + \int_0^t \dfrac{1 + X_{j,s}^2}{2 + X_{j,s}^2} \sqrt{\dfrac{1 + 2Y_s^2}{1 + Y_s^2 + \exp\left(-\frac{2|X_s|^2}{d(s+5)}\right)}}\, dW_{j,s}, \quad j = 1, \ldots, d, \\
Y_t = \exp\left(-\dfrac{|X_T|^2}{d(T+5)}\right) \\
\qquad + \int_t^T a(s, X_s, Y_s) + \sum_{j=1}^d b(s, X_{j,s}, Y_s) Z_{j,s}\, ds - \int_t^T (Z_s)^{\mathrm{T}}\, dW_s,
\end{cases}
\tag{6.1}
$$

where $X_{j,t}, Z_{j,t}, W_{j,t}$ denote the j-th components of X_t, Y_t, W_t, and the coefficient functions are given as

$$
a(t, x, u) = \frac{1}{d(t+5)} \exp\left(-\frac{|x|^2}{d(t+5)}\right)
$$

$$
\times \sum_{j=1}^d \left\{ \frac{4x_j^2\left(1 + x_j^2\right)}{\left(2 + x_j^2\right)^3} + \frac{\left(1 + x_j^2\right)^2}{\left(2 + x_j^2\right)^2} - \frac{2x_j^2\left(1 + x_j^2\right)^2}{d(t+5)\left(2 + x_j^2\right)^2} - \frac{x_j^2}{t+5} \right\},
$$

$$
b(t, x_j, u) = \frac{x_j}{\left(2 + x_j^2\right)^2} \sqrt{\frac{1 + u^2 + \exp\left(-\frac{|x|^2}{d(t+5)}\right)}{1 + 2u^2}}.
$$

It can be verified by Itô's formula that the Y part of the solution of (6.1) is given by

$$Y_t = \exp\left(-\frac{|X_t|^2}{d(t+5)}\right).$$

Let $\xi = (1, 1, \ldots, 1)$ (100-dimensional), $T = 5, N = 160$. The initial guess of Y_0 is generated from a uniform distribution on the interval $[2, 4]$ while the true value of $Y_0 \approx 0.81873$. We train 25000 steps with an exponential decay learning rate that decays every 100 steps, with the starting learning rate being 1e-2 and ending learning rate being 1e-5. Figure 1 illustrates the mean of the loss function and relative approximation error of Y_0 against the number of iteration steps. All runs converged and the average final relative error of Y_0 is 0.39%.

6.3 Example 2

The second problem is adapted from (Bender and Zhang 2008), in which the spatial dimension is originally tested up to 10. The coupled FBSDEs are given by

$$\begin{cases} X_{j,t} = x_{j,0} + \int_0^t \sigma Y_s \, dW_{j,s}, \quad j = 1, \ldots, d, \\[2mm] Y_t = D \sum_{j=1}^d \sin(X_{j,T}) \\[2mm] \quad + \int_t^T -rY_s + \frac{1}{2} e^{-3r(T-s)\sigma^2} \left(D \sum_{j=1}^d \sin(X_{j,s})\right)^3 ds - \int_t^T (Z_s)^\mathsf{T} \, dW_s, \end{cases}$$

(6.2)

where $\sigma > 0, r, D$ are constants. One can easily check by Itô's formula that the Y part of the solution of (6.2) is

$$Y_t = e^{-r(T-t)} D \sum_{j=1}^d \sin(X_{j,t}).$$
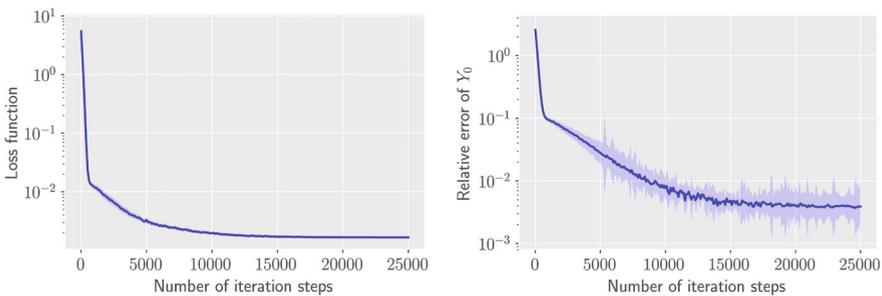
Fig. 1 Loss function (left) and relative approximation error of Y_0 (right) against the number of iteration steps in the case of Example 1 (100-dimensional). The proposed deep BSDE method achieves a relative error of size 0.39%. The shaded area depicts the mean $\pm$ the standard deviation of the associated quantity in 5 runs
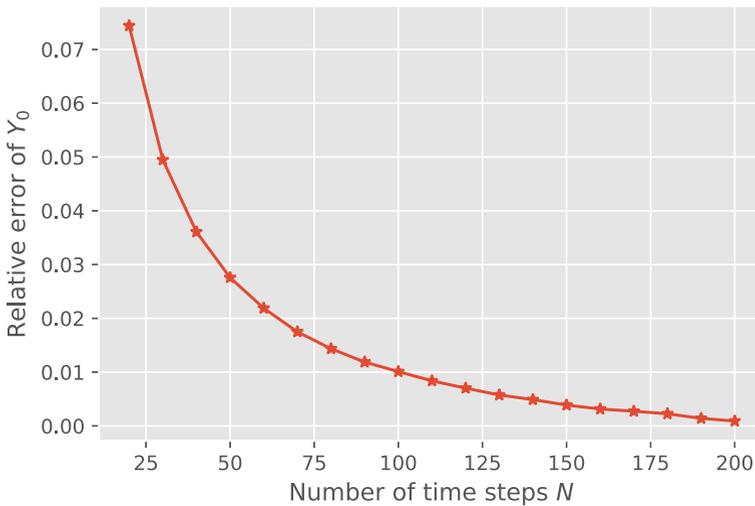
Fig. 2 Relative approximation error of Y_0 against the time step size h in the case of Example 2 (100-dimensional). The proposed deep BSDE method achieves a relative error of size 0.09% when $N = 200$ ($h = 0.005$)

Let $\xi = (\pi/2, \pi/2, \ldots, \pi/2)$ (100-dimensional), $T = 1, r = 0.1, \sigma = 0.3, D = 0.1$. The initial guess of Y_0 is generated from a uniform distribution on the interval [0, 1] while the true value of $Y_0 \approx 9.04837$. We train 5000 steps with an exponential decay learning rate that decays every 100 steps, with the starting learning rate being 1e-2 and the ending learning rate being 1e-3. When $h = 0.005$ ($N = 200$), the relative approximation error of Y_0 is 0.09%. Furthermore, we test the influence of the time partition by choosing different values of N. In all cases, the training converged, and we plot in Fig. 2 the mean of relative error of Y_0 against the number of time steps N. It is clearly shown that the error decreases as N increases (h decreases).

Authors' contributions

Both authors read and approved the final manuscript.

Funding

Availability of data and materials

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Competing interests

References

Antonelli, F.: Backward-forward stochastic differential equations. Ann. Appl. Probab. **3**, 777–793 (1993)

Arora, R., Basu, A., Mianjy, P., Mukherjee, A.: Understanding deep neural networks with rectified linear units (2018). In: Proceedings of the International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=B1J_rgWRW

Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans Inf. Theory. **39**(3), 930–945 (1993)

Beck, C., E, W., Jentzen, A.: Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations (2017). arXiv preprint arXiv:170905963

Bellman, R.E.: Dynamic Programming. Princeton University Press, USA (1957)

Bender, C., Steiner, J.: Least-squares Monte Carlo for backward SDEs. In: Carmona, R., Del Moral, P., Hu, P., Oudjane, N. (eds.), pp. 257–289. Numerical Methods in Finance. Springer Proceedings in Mathematics, vol 12. Springer, Berlin (2012)

Bender, C., Zhang, J.: Time discretization and Markovian iteration for coupled FBSDEs. Ann. Appl. Probab. **18**(1), 143–177 (2008)

Berner, J., Grohs, P., Jentzen, A.: Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations (2018). arXiv preprint arXiv:180903062

Bölcskei, H., Grohs, P., Kutyniok, G., Petersen, P.: Optimal approximation with sparsely connected deep neural networks (2017). arXiv preprint arXiv:170501714

Bouchard, B., Ekeland, I., Touzi, N.: On the Malliavin approach to Monte Carlo approximation of conditional expectations. Finance Stoch. **8**(1), 45–71 (2004)

Bouchard, B., Touzi, N.: Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations. Stoch. Process. Appl. **111**(2), 175–206 (2004)

Cohen, N., Sharir, O., Shashua, A.: On the expressive power of deep learning: A tensor analysis. In: Feldman, V., Rakhlin, A., Shamir, O. (eds.), pp. 698–728. 29th Annual Conference on Learning Theory, vol. 49. PMLR, Columbia University, New York (2016). pp. 698–728. http://proceedings.mlr.press/v49/cohen16.html

Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control. Signal. Syst. **2**(4), 303–314 (1989)

E, W., Han, J., Jentzen, A.: Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. Commun. Math. Stat. **5**(4), 349–380 (2017)

E, W., Hutzenthaler, M., Jentzen, A., Kruse, T.: On multilevel Picard numerical approximations for high-dimensional nonlinear parabolic partial differential equations and high-dimensional nonlinear backward stochastic differential equations. J. Sci. Comput. **79**(3), 1534–1571 (2019)

Eldan, R., Shamir, O.: The power of depth for feedforward neural networks. In: Feldman, V., Rakhlin, A., Shamir, O. (eds.), pp. 907–940. 29th Annual Conference on Learning Theory, vol. 49. PMLR, Columbia University, New York (2016). pp. 907–940. http://proceedings.mlr.press/v49/eldan16.html

Funahashi, K.I.: On the approximate realization of continuous mappings by neural networks. Neural Netw. **2**(3), 183–192 (1989)

Grohs, P., Hornung, F., Jentzen, A., von Wurstemberger, P.: A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations (2018). arXiv preprint arXiv:180902362

Han, J., Hu, R.: Deep fictitious play for finding Markovian Nash equilibrium in multi-agent games (2019). arXiv preprint arXiv:191201809

Han, J., Jentzen, A., E, W.: Solving high-dimensional partial differential equations using deep learning. Proc. Natl. Acad. Sci. **115**(34), 8505–8510 (2018)

Han, J., Lu, J., Zhou, M.: Solving high-dimensional eigenvalue problems using deep neural networks: A diffusion Monte Carlo like approach (2020). arXiv preprint arXiv:200202600

Henry-Labordere, P.: Counterparty risk valuation: A marked branching diffusion approach (2012). Available at SSRN 1995503. https://arxiv.org/abs/1203.2369

Henry-Labordere, P., Oudjane, N., Tan, X., Touzi, N., Warin, X., et al.: Branching diffusion representation of semilinear PDEs and Monte Carlo approximation, pp. 184–210. In: Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, Institut Henri Poincaré, vol. 55. Institut Henri Poincaré, Paris (2019). pp. 184–210. https://projecteuclid.org/euclid.aihp/1547802399

Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)

Huijskens, T., Ruijter, M., Oosterlee, C.: Efficient numerical Fourier methods for coupled forward–backward SDEs. J. Comput. Appl. Math. **296**, 593–612 (2016)

Hutzenthaler, M., Jentzen, A., Kruse, T., et al.: Multilevel Picard iterations for solving smooth semilinear parabolic heat equations (2016). arXiv preprint arXiv:160703295

Hutzenthaler, M., Jentzen, A., Kruse, T., Nguyen, T.A.: A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations (2020). arXiv preprint arXiv:190110854

Hutzenthaler, M., Jentzen, A., Kruse, T., Nguyen, T.A., von Wurstemberger, P.: Overcoming the curse of dimensionality in the numerical approximation of semilinear parabolic partial differential equations (2018). arXiv preprint arXiv:180701212

Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning, Vol. 37. JMLR.org, Lille (2015). pp. 448–456

Jentzen, A., Salimova, D., Welti, T.: 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 (2018). arXiv preprint arXiv:180907321

Kingma, D., Ba, J.: Adam: a method for stochastic optimization (2015). In: Proceedings of the International Conference on Learning Representations (ICLR)

Liang, S., Srikant, R.: Why deep neural networks for function approximation? (2017). In: Proceedings of the International Conference on Learning Representations (ICLR)

Ma, J., Protter, P., Yong, J.: Solving forward-backward stochastic differential equations explicitly–a four step scheme. Probab. Theory Relat. Fields. **98**(3), 339–359 (1994)

Ma, J., Yong, J.: Forward-Backward Stochastic Differential Equations and their Applications. Springer, Berlin Heidelberg (2007)

Mhaskar, H.N., Poggio, T.: Deep vs. shallow networks: An approximation theory perspective. Anal. Appl. **14**(06), 829–848 (2016)

Milstein, G., Tretyakov, M.: Numerical algorithms for forward-backward stochastic differential equations. SIAM J. Sci. Comput. **28**(2), 561–582 (2006)

Pardoux, E., Peng, S.: Backward stochastic differential equations and quasilinear parabolic partial differential equations, pp. 200–217. Springer, Berlin (1992)

Pardoux, E., Tang, S.: Forward-backward stochastic differential equations and quasilinear parabolic PDEs. Prob. Theory Relat. Fields. **114**(2), 123–150 (1999)

Zhang, J.: A numerical scheme for BSDEs. Ann. Appl. Prob. **14**(1), 459–488 (2004)