Generalized convergence of the deep BSDE method: a step towards fully-coupled FBSDEs and applications in stochastic control

Balint Negyesi[a,*], Zhipeng Huang[b], Cornelis W. Oosterlee[b]

[a]*Delft Institute of Applied Mathematics (DIAM), Delft University of Technology, P.O. Box 5031, 2600 GA, Delft, The Netherlands*
[b]*Mathematical Institute, Utrecht University, Postbus 80010, 3508 TA, Utrecht, The Netherlands*

Abstract

We are concerned with high-dimensional coupled FBSDE systems approximated by the deep BSDE method of Han et al. (2018). It was shown by Han and Long (2020) that the errors induced by the deep BSDE method admit a posteriori estimate depending on the loss function, whenever the backward equation only couples into the forward diffusion through the Y process. We generalize this result to fully-coupled drift coefficients, and give sufficient conditions for convergence under standard assumptions. The resulting conditions are directly verifiable for any equation. Consequently, unlike in earlier theory, our convergence analysis enables the treatment of FBSDEs stemming from stochastic optimal control problems. In particular, we provide a theoretical justification for the non-convergence of the deep BSDE method observed in recent literature, and present direct guidelines for when convergence can be guaranteed in practice. Our theoretical findings are supported by several numerical experiments in high-dimensional settings.

Keywords:
deep BSDE, coupled FBSDE, posteriori estimate, convergence

Contents

*Corresponding author
Email address: `b.negyesi@tudelft.nl` (Balint Negyesi)

1. Introduction

In this paper, we are concerned with the numerical approximation of a system of coupled forward-backward stochastic differential equations (FBSDE) over a finite time interval $[0,T]$

$$\begin{cases} X_t = x_0 + \int_0^t b(s, X_s, Y_s, Z_s)\mathrm{d}s + \int_0^t \sigma(s, X_s, Y_s)\mathrm{d}W_s, \\ Y_t = g(X_T) + \int_t^T f(s, X_s, Y_s, Z_s)\mathrm{d}s - \int_t^T Z_s \mathrm{d}W_s, \end{cases} \tag{1}$$

where $b:[0,T]\times\mathbb{R}^d\times\mathbb{R}^q\times\mathbb{R}^{q\times m}\to\mathbb{R}^d$, $\sigma:[0,T]\times\mathbb{R}^d\times\mathbb{R}^q\to\mathbb{R}^{d\times m}$, $f:[0,T]\times\mathbb{R}^d\times\mathbb{R}^q\times\mathbb{R}^{q\times m}\to\mathbb{R}^q$ and $g:\mathbb{R}^d\to\mathbb{R}^q$ are all deterministic mappings. The equation is given on a complete probability space $(\Omega,\mathcal{F},\mathbb{P})$, over which $\{W_t\}_{0\le T}$ is a standard m-dimensional Brownian motion, $\mathcal{F}:=\{\mathcal{F}_t\}_{0\le t\le T}$ its natural filtration and augmented by the usual $\mathbb{P}$-null sets. A triple of $(\mathbb{R}^d\times\mathbb{R}^q\times\mathbb{R}^{q\times m})$ valued, $\mathcal{F}_t$ adapted stochastic processes $\{(X_t,Y_t,Z_t)\}_{0\le t\le T}$ is a solution if (1) holds $\mathbb{P}$ almost surely and the processes satisfy natural integrability conditions, see [1, 2].

The study of FBSDEs was initiated by the seminal paper of Pardoux and Peng [3], and then extended to coupled equations by Antonelli [4]. Equations like (1) subsequently attracted widespread attention due to their inherent connections with systems of second-order quasi-linear partial differential equations (PDE) established by non-linear extension to the Feynman-Kac lemma, see e.g. [2, 1] and theorem 1 below. This probabilistic representation casts FBSDEs to be the natural framework to model a wide range of problems arising in mathematical finance, physics, biology and stochastic control. The well-posedness of (1) has been rigorously studied and established under by now standard assumptions, see e.g. [2, 5, 6, 7] and the references therein. Most of such results rely either on classical solutions of the corresponding quasi-linear PDE with high regularity or some abstract conditions heuristically associated with monotonicity, small time duration or weak coupling. In the rest of the article, we consider the setting where (1) is well-posed and admits a unique strong solution triple.

Solving FBSDEs analytically is seldom possible and one usually has to resort to numerical approximations. In the decoupled framework where b,σ in (1) do not depend on Y,Z, one can detach the solution of X, and use the resulting discrete time approximation in order to approximate the solution pair of the BSDE by sequences of backward, recursive conditional expectations, we refer to [8, 9, 10, 11, 12]. In the coupled framework, things become more subtle due to the interdependence from the backward equation into the forward diffusion. Inspired by [13], classical approaches usually consist of *decoupling* the forward diffusion by means of deterministic mappings, and iteratively converging to the unique decoupling field related to the associated quasi-linear PDE, see e.g. [13, 14, 15, 16, 17, 18]. A common challenge across the aforementioned classical references is the setting of high-dimensionality. In fact, whenever either d,q or m are large, these methods suffer from the curse dimensionality and become intractable in high dimensions. Such settings arise naturally, for instance, in portfolio allocation or climate risk management. A recently emerging branch of numerical algorithms called *deep BSDE* methods and pioneered by [19, 20] addressed this gap, and has shown remarkable empirical performance in terms of tackling high-dimensional FBSDEs and associated quasi-linear PDEs. These approaches were first developed for decoupled equations, see e.g. [19, 20, 21, 22, 23], and then extended to the coupled framework [24, 25]. For an overview, we refer to the recent survey [26].

Motivated by their outstanding empirical performance, serious efforts have been made in order to establish convergence guarantees for such deep BSDE algorithms. Originally the pioneering paper of Han and Long [24] managed to show a *posteriori bound* which depends on the objective loss functional of the machine learning algorithm. Their result was later extended in [27] to the case of non-Lipschitz continuous drift coefficients, and in [28] to the vector-valued framework in the context of stochastic optimal control. These works all have in common that they relied on the assumption of a narrower class of FBSDEs, in fact they considered a special case of (1) where only the Y process enters the dynamics of X, and not Z. Consequently, these convergence results were inapplicable in the context a wide range of stochastic control problems, for instance formulated through the dynamic programming principle [29], where coupling occurs in Z. In particular, these works could not provide a theoretical explanation for the phenomena observed in [25], where they found empirical evidence for the non-convergence of the deep BSDE method for FBSDEs stemming from stochastic control.

The main motivation of the present paper is to address this gap in the literature. In fact, we extend the convergence analysis of [24] to the case of fully-coupled drift coefficients as in (1). To the best of our understanding, this is the first convergence result in the literature which allows for Z coupling. The main challenge is to handle the error estimate of the X process with extra Z coupling, which we control by the new estimates established in lemma 2. This enables us to derive our main result, stated in theorem 3, which is a posteriori error estimate similar to [24]. In particular, our work enjoys several important features:

- we recover the results of [24] in the limit case of no Z coupling;
- our result is applicable for a more general class of FBSDEs, including, but not limited to, the ones obtained for stochastic control problems stemming either from dynamic programming [25] or the stochastic maximum principle [30, 28], due to the coupling of not just Y but also Z in the forward process;
- given a particular FBSDE, we can check whether or not it satisfies the convergence conditions through a directly verifiable approach.

The paper is organized as follows. In section 2, we give the discrete time approximation scheme of the deep BSDE algorithm with Z coupling in the drift b. Section 3 contains our main result stated in theorem 3. Thereafter the abstract sufficient conditions of convergence are analyzed in section 4. In particular, we show that the assumptions of theorem 3 hold under heuristic interpretations such as *weak coupling* or *small time duration*. Additionally, we get earlier convergence results [24] as a limit case of our more general theory. Finally, we demonstrate our theoretical contributions by several numerical experiments on high-dimensional FBSDEs in section 5. These simulations confirm and showcase our theoretical findings.

2. The deep BSDE algorithm

In this section we formulate the deep BSDE algorithm for FBSDEs as in (1), naturally extending [19, 20, 24] to the framework of Z coupling in b. For the rest of the paper we denote the Frobenius norm by $\|x\|$ for any $x \in \mathbb{R}^{i\times j}$, not to be confused with the matrix 2-norm $\|x\|_2$. Without loss of generality, we work with an equidistant time partition $\pi := \{t_i, i = 0, \dots, N | 0 = t_0 < t_1 < \dots <$

$t_N = T\}$ with $h = T/N$ and study the following discretization

$$\inf_{\varphi_0 \in \mathcal{N}_0^Y(\theta_0^Y), \zeta_i \in \mathcal{N}_i^Z(\theta_i^Z)} \mathbb{E}\left[\left\| g\left(X_{t_N}^\pi\right) - Y_{t_N}^\pi \right\|^2 \right], \tag{2a}$$

$$\text{s.t.} \begin{cases} X_0^\pi = x_0, \\ Y_0^\pi = \varphi_0(x_0; \theta_0^Y), \\ X_{t_{i+1}}^\pi = X_{t_i}^\pi + b\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi, Z_{t_i}^\pi\right) h + \sigma\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi\right) \Delta W_i, \\ Z_{t_i}^\pi = \zeta_i(X_{t_i}^\pi; \theta_i^Z), \\ Y_{t_{i+1}}^\pi = Y_{t_i}^\pi - f\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi, Z_{t_i}^\pi\right) h + Z_{t_i}^\pi \Delta W_i, \end{cases} \tag{2b}$$

for $i = 0, \ldots, N-1$, where we put $\Delta W_i \coloneqq W_{t_{i+1}} - W_{t_i}$. In doing so, the numerical solution of a coupled FBSDE (1) is reformulated into a stochastic optimization problem consisting of the minimization of an objective functional (2a) subject to the Euler-Maruyama discretization (2b). As in the continuous limit the loss functional (2a) attains 0 at the unique solution triple $\{(X_t, Y_t, Z_t)\}_{0 \le t \le T}$ of (1) while also satisfying (2b), it is expected that for sufficiently large N and sufficiently wide function spaces $\mathcal{N}_0^Y, \mathcal{N}_i^Z$, the solution of (2a)-(2b) is a good discrete time approximation of (1). Motivated by universal approximation arguments, see e.g. [31], we set $\mathcal{N}_0^Y(\theta_0^Y), \mathcal{N}_i^Z(\theta_i^Z)$ to be spaces of deep neural networks parametrized by θ_0^Y, θ_i^Z for $i = 0, \ldots, N-1$. Subsequently, the goal of the deep BSDE method is to solve this non-linear, constrained optimization problem through the training of deep neural networks. Hence, we seek to find $\varphi_0(x_0; \theta_0^Y) \in \mathcal{N}_0^Y$ and $\zeta_i(X_{t_i}^\pi; \theta_i^Z) \in \mathcal{N}_i^Z$ that approximate Y_0 and Z_{t_i} sufficiently well. The resulting pseudo-code for the complete deep BSDE method is collected in algorithm 1, its implementation is discussed in section 5.

Algorithm 1 Deep BSDE algorithm

Input: Initial parameters $(\theta_0^{Y,(0)}, \theta_i^{Z,(0)})$, learning rate η; batch size M; number of iterations K.
Data: Simulated Brownian increments $\{\Delta W_{t_i}^{(k)}\}_{0 \le i \le N-1, 1 \le k \le K}$
Output: Discrete time approximations $\{(\hat{X}_{t_i}^\pi, \hat{Y}_{t_i}^\pi, \hat{Z}_{t_i}^\pi)\}_{i=0,\ldots,N}$
for $k = 1$ to K **do** ▷ Euler-Maruyama (2b)
    $X_{t_0}^{\pi,(k)} = x_0$, $Y_{t_0}^{\pi,(k)} = \varphi_0(x_0; \theta_0^{Y,(k-1)})$
    for $i = 0$ to $N-1$ **do**
        $Z_{t_i}^{\pi,(k)} = \zeta_i(X_{t_i}^{\pi,(k)}; \theta_i^{Z,(k-1)})$
        $X_{t_{i+1}}^{\pi,(k)} = X_{t_i}^{\pi,(k)} + b(t_i, X_{t_i}^{\pi,(k)}, Y_{t_i}^{\pi,(k)}, Z_{t_i}^{\pi,(k)})h + \sigma(t_i, X_{t_i}^{\pi,(k)}, Y_{t_i}^{\pi,(k)})\Delta W_{t_i}^{(k)}$
        $Y_{t_{i+1}}^{\pi,(k)} = Y_{t_i}^{\pi,(k)} - f(t_i, X_{t_i}^{\pi,(k)}, Y_{t_i}^{\pi,(k)}, Z_{t_i}^{\pi,(k)})h + Z_{t_i}^{\pi,(k)} \Delta W_{t_i}^{(k)}$
    end for
    Loss $= \frac{1}{M} \sum_{j=1}^{M} \|g(X_{t_N}^{\pi,(k)}) - Y_{t_N}^{\pi,(k)}\|^2$ ▷ empirical (2a)
    $(\theta_0^{Y,(k)}, \theta_0^{Z,(k)}, \ldots, \theta_{N-1}^{Z,(k)}) = (\theta_0^{Y,(k-1)}, \theta_0^{Z,(k-1)}, \ldots, \theta_{N-1}^{Z,(k-1)}) - \eta \nabla \text{Loss}$ ▷ SGD
end for
$(\hat{X}_i^\pi, \hat{Y}_{t_i}^\pi, \hat{Z}_{t_i}^\pi) = (X_{t_i}^{\pi,(K+1)}, Y_{t_i}^{\pi,(K+1)}, Z_{t_i}^{\pi,(K+1)}), \quad i = 0, \ldots, N-1$

3. Convergence analysis

This section is dedicated to our generalized convergence analysis for the deep BSDE method reviewed in section 2 and the discrete scheme (2b) for (1). In particular, we will show that the approximation errors of the numerical solution to the FBSDE are bounded by the simulation error of the objective function corresponding to (2a) which could be arbitrarily small due to the universal approximation theorem. The convergence analysis follows a similar strategy as that of [24]. We first introduce the standing assumptions and review some useful result. Throughout the paper we use the notation $\mathbb{E}_i[\cdot] \coloneqq \mathbb{E}[\cdot|\mathcal{F}_i]$.

Assumption 1. *There exist constants k^b and k^f, that are possibly negative, such that*

$$\begin{aligned}(b(t,x_1,y,z)-b(t,x_2,y,z))^\top \Delta x \le k^b\|\Delta x\|^2,\\ (f(t,x,y_1,z)-f(t,x,y_2,z))^\top \Delta y \le k^f\|\Delta y\|^2.\end{aligned}$$

Assumption 2. *b,σ,f,g are uniformly Lipschitz continuous with respect to (x,y,z). In particular, there are non-negative constants such that*

$$\begin{aligned}\|b(t,x_1,y_1,z_1)-b(t,x_2,y_2,z_2)\|^2 &\le L_x^b\|\Delta x\|^2+L_y^b\|\Delta y\|^2+L_z^b\|\Delta z\|^2,\\ \|\sigma(t,x_1,y_1)-\sigma(t,x_2,y_2)\|^2 &\le L_x^\sigma\|\Delta x\|^2+L_y^\sigma\|\Delta y\|^2,\\ \|f(t,x_1,y_1,z_1)-f(t,x_2,y_2,z_2)\|^2 &\le L_x^f\|\Delta x\|^2+L_y^f\|\Delta y\|^2+L_z^f\|\Delta z\|^2,\\ \|g(x_1)-g(x_2)\|^2 &\le L_x^g\|\Delta x\|^2.\end{aligned}$$

Assumption 3. *$g(0),b(t,0,0,0),f(t,0,0,0)$ and $\sigma(t,0,0)$ are bounded for $t\in[0,T]$.*

Notice that assumption 2 implies 1 with $k^b,k^f\ge 0$. The reason for allowing for negativity shall be made clear by the forthcoming convergence result, see points ((3)d), ((3)e) in section 4 below. For convenience, we use $\mathscr{L}$ to denote the set of all the constants mentioned above and assume L is the upper bound of $\mathscr{L}$.

Next, we introduce the following system of quasi-linear parabolic PDEs associated with FBSDE (1),

$$\begin{cases}\partial_t\nu^i+\dfrac{1}{2}\partial_{xx}\nu^i:\sigma\sigma^\top(t,x,\nu)+\partial_x\nu^i b\left(t,x,\nu,\partial_x\nu\sigma(t,x,\nu)\right)+f^i\left(t,x,\nu,\partial_x\nu\sigma(t,x,\nu)\right)=0,\\ \nu(T,x)=g(x),\quad \forall i=1,\cdots,q,\end{cases}\tag{3}$$

The following assumption is needed in order to guarantee convergence of the implicit Euler-Maruyama scheme in theorem 2.

Assumption 4. *The PDE (3) has a classical solution ν with bounded derivatives $\partial_x\nu$ and $\partial^2_{xx}\nu$, and σ is bounded.*

The non-linear Feynman-Kac lemma, stated below, establishes the connection between (3) and (1).

Theorem 1 (Feynman-Kac)**.** *Under Assumptions 2, 3 and 4, the FBSDE (1) has a unique solution (X,Y,Z), and it holds that for $t\in[0,T]$,*

$$Y_t=\nu\left(t,X_t\right),\quad Z_t=\partial_x\nu\left(t,X_t\right)\sigma\left(t,X_t,\nu\left(t,X_t\right)\right).$$

Remark 1. *The proof of this theorem can be found in [1, pp. 185-186]. Similar to other numerical methods for coupled FBSDEs, we use this theorem to decouple the original FBSDE (1) in order to be able to exploit standard results from the decoupled FBSDE literature.*

In addition to the assumptions above, we need Hölder-continuity in time for the convergence of the implicit scheme for (1), as stated below.

Assumption 5. *b, σ, f are uniformly Hölder-$\frac{1}{2}$-continuous with respect to t.*

Theorem 2 (Convergence of the implicit scheme)**.** *Under Assumptions 2, 3, 4 and 5, for sufficiently small h, the following discrete-time equation* $(0 \le i \le N-1)$

$$\begin{cases} \bar{X}_0^\pi = x_0, \\ \bar{X}_{t_{i+1}}^\pi = \bar{X}_{t_i}^\pi + b\left(t_i, \bar{X}_{t_i}^\pi, \bar{Y}_{t_i}^\pi, \bar{Z}_{t_i}^\pi\right) h + \sigma\left(t_i, \bar{X}_{t_i}^\pi, \bar{Y}_{t_i}^\pi\right) \Delta W_i, \\ \bar{Y}_T^\pi = g\left(\bar{X}_T^\pi\right), \\ \bar{Z}_{t_i}^\pi = \frac{1}{h}\mathbb{E}_i\left[\bar{Y}_{t_{i+1}}^\pi \Delta W_i^\top\right], \\ \bar{Y}_{t_i}^\pi = \mathbb{E}_i\left[\bar{Y}_{t_{i+1}}^\pi + f\left(t_i, \bar{X}_{t_i}^\pi, \bar{Y}_{t_i}^\pi, \bar{Z}_{t_i}^\pi\right) h\right], \end{cases} \tag{4}$$

has a solution $\{(\bar{X}_{t_i}^\pi, \bar{Y}_{t_i}^\pi, \bar{Z}_{t_i}^\pi)\}_{i=0,\ldots,N}$, such that $\bar{X}_{t_i}^\pi \in L^2\left(\Omega, \mathcal{F}_{t_i}, \mathbb{P}\right)$ and

$$\sup_{t\in[0,T]} \left(\mathbb{E}[\left\|X_t - \bar{X}_t^\pi\right\|^2] + \mathbb{E}[\left\|Y_t - \bar{Y}_t^\pi\right\|^2]\right) + \int_0^T \mathbb{E}[\left\|Z_t - \bar{Z}_t^\pi\right\|^2]\mathrm{d}t \le C\left(1 + \mathbb{E}\|x_0\|^2\right) h,$$

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

Remark 2. *This result follows from theorem 5.3.1 and theorem 5.3.3 in [1]. In particular, therein the results are stated for the explicit scheme, but the same convergence result holds also for the implicit scheme as mentioned in [1, Remark 5.3.2 (ii)].*

Remark 3. *It is worth noting that theorem 2 is derived differently from [24] or [17], as in our setting the drift function b has an extra argument Z. Consequently, we do not require the weak and monotonicity conditions for theorem 2.*

Recall the classical Euler scheme in (2b). Taking conditional expectations of the discrete equation of $Y_{t_{i+1}}^\pi$, and of the same equation multiplied by $(\Delta W_i)^\top$, we obtain a formulation that does not include the objective functional (2a)

$$\begin{cases} X_0^\pi = x_0, \\ X_{t_{i+1}}^\pi = X_{t_i}^\pi + b\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi, Z_{t_i}^\pi\right) h + \sigma\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi\right) \Delta W_i, \\ Z_{t_i}^\pi = \frac{1}{h}\mathbb{E}_i\left[Y_{t_{i+1}}^\pi \Delta W_i^\top\right], \\ Y_{t_i}^\pi = \mathbb{E}_i\left[Y_{t_{i+1}}^\pi + f\left(t_i, X_{t_i}^\pi, Y_{t_i}^\pi, Z_{t_i}^\pi\right) h\right]. \end{cases} \tag{5}$$

With formulation (5) in hand, we can derive the following apriori estimate bounding the difference between two solutions of it.

Lemma 1. *For $j = 1, 2$, suppose $\left(\left\{X_{t_i}^{\pi,j}\right\}_{0\le i\le N}, \left\{Y_{t_i}^{\pi,j}\right\}_{0\le i\le N}, \left\{Z_{t_i}^{\pi,j}\right\}_{0\le i\le N-1}\right)$ are two solutions of (5), with $X_{t_i}^{\pi,j}, Y_{t_i}^{\pi,j} \in L^2\left(\Omega, \mathcal{F}_{t_i}, \mathbb{P}\right), 0 \le i \le N$. For any $\lambda_1 > 0, \lambda_2 > L_z^f$, and sufficiently small h, denote*

$$K_1 \coloneqq 2k^b + \lambda_1 + L_x^\sigma + L_x^b h, \quad K_2 \coloneqq \left(\lambda_1^{-1} + h\right) L_y^b + L_y^\sigma, \quad K_3 \coloneqq -\frac{\ln\left(1 - \left(2k^f + \lambda_2\right) h\right)}{h}, \tag{6}$$
$$K_4 \coloneqq \frac{L_x^f}{\left(1 - \left(2k^f + \lambda_2\right) h\right) \lambda_2}, \quad C_1 \coloneqq L_z^b(h + \lambda_1^{-1}).$$

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

$$\mathbb{E}\left[\|\delta X_i\|^2\right] \le K_2 h \sum_{i=0}^{n-1} e^{K_1(n-i-1)h}\mathbb{E}\left[\|\delta Y_i\|^2\right] + C_1 h \sum_{i=0}^{n-1} e^{K_1(n-i-1)h}\mathbb{E}\left[\|\delta Z_i\|^2\right], \tag{7}$$

$$\mathbb{E}\left[\|\delta Y_i\|^2\right] \le e^{K_3(N-n)h}\mathbb{E}\left[\|\delta Y_N\|^2\right] + K_4 \sum_{i=n}^{N-1} e^{K_3(i-n)h}\mathbb{E}\left[\|\delta X_i\|^2\right] h. \tag{8}$$

Proof. Let us define

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

$$\delta f_i := f\left(t_i, X_{t_i}^{\pi,1}, Y_{t_i}^{\pi,1}, Z_{t_i}^{\pi,1}\right) - f\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,2}, Z_{t_i}^{\pi,2}\right).$$

Then we have

$$\delta X_{i+1} = \delta X_i + \delta b_i h + \delta\sigma_i \Delta W_i, \tag{9}$$

$$\delta Y_i = \mathbb{E}_i\left[\delta Y_{i+1} + \delta f_i h\right], \tag{10}$$

and (5) also gives

$$\delta Z_i = \frac{1}{h}\mathbb{E}_i\left[\delta Y_{i+1}\Delta W_i^\top\right]. \tag{11}$$

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

$$\delta Y_{i+1} = \mathbb{E}_i\left[\delta Y_{i+1}\right] + \int_{t_i}^{t_{i+1}} \delta Z_t \mathrm{d}W_t,$$

which together with (10) implies

$$\delta Y_{i+1} = \delta Y_i - \delta f_i h + \int_{t_i}^{t_{i+1}} \delta Z_t \mathrm{d}W_t. \tag{12}$$

From (9) and (12), noting that $\delta X_i, \delta Y_i, \delta b_i, \delta\sigma_i$ and δf_i are all $\mathcal{F}_{t_i}$ measurable, and $\mathbb{E}_i[\Delta W_i] = 0$, $\mathbb{E}_i[\int_{t_i}^{t_{i+1}} Z_t \mathrm{d}W_t] = 0$, we have

$$\mathbb{E}\left[\|\delta X_{i+1}\|^2\right] = \mathbb{E}\left[\|\delta X_i + \delta b_i h\|^2\right] + h\mathbb{E}\left[\|\delta\sigma_i\|^2\right],$$

$$\mathbb{E}\left[\|\delta Y_{i+1}\|^2\right] = \mathbb{E}\left[\|\delta Y_i - \delta f_i h\|^2\right] + \int_{t_i}^{t_{i+1}} \mathbb{E}\left[\|\delta Z_t\|^2\right] \mathrm{d}t,$$

where we also used a Fubini argument. We proceed in steps, controlling each of the terms above.

Step 1. *Estimate for* δX_n. By assumptions 1, 2, and the root-mean-square and geometric mean inequality (RMS-GM inequality), we have that for any $\lambda_1 > 0$

$$\begin{aligned}
\mathbb{E}\left[\|\delta X_{i+1}\|^2\right] &= \mathbb{E}\left[\|\delta X_i\|^2\right] + \mathbb{E}\left[\|\delta b_i\|^2\right] h^2 + h\mathbb{E}\left[\|\delta\sigma_i\|^2\right] \\
&\quad + 2h\mathbb{E}\left[\left(b\left(t_i, X_{t_i}^{\pi,1}, Y_{t_i}^{\pi,1}, Z_{t_i}^{\pi,1}\right) - b\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,1}, Z_{t_i}^{\pi,1}\right)\right)^\top \delta X_i\right] \\
&\quad + 2h\mathbb{E}\left[\left(b\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,1}, Z_{t_i}^{\pi,1}\right) - b\left(t_i, X_{t_i}^{\pi,2}, Y_{t_i}^{\pi,2}, Z_{t_i}^{\pi,2}\right)\right)^\top \delta X_i\right] \\
&\le \left(1 + \left(2k^b + \lambda_1 + L_x^\sigma + L_x^b h\right) h\right) \mathbb{E}\left[\|\delta X_i\|^2\right] \\
&\quad + \left(\left(\lambda_1^{-1} + h\right) L_y^b + L_y^\sigma\right) \mathbb{E}\left[\|\delta Y_i\|^2\right] h + \left(L_z^b h + \lambda_1^{-1} L_z^b\right) \mathbb{E}\left[\|\delta Z_i\|^2\right] h.
\end{aligned}$$

Recalling the definition of C_1, K_1, K_2 from (6), we subsequently gather

$$\mathbb{E}\left[\|\delta X_{i+1}\|^2\right] \leq (1+K_1h)\mathbb{E}\left[\|\delta X_i\|^2\right] + K_2h\mathbb{E}\left[\|\delta Y_i\|^2\right] + C_1h\mathbb{E}\left[\|\delta Z_i\|^2\right].$$

Notice that $\mathbb{E}\left[\|\delta X_0\|^2\right] = 0$, and thus by induction, we have that for any $1 \leq n \leq N$,

$$\begin{aligned}
\mathbb{E}\left[\|\delta X_n\|^2\right] \leq & \prod_{i=0}^{n-1}(1+K_1h)\mathbb{E}\left[\|\delta X_0\|^2\right] + \sum_{i=0}^{n-1}(1+K_1h)^{n-1-i}K_2\mathbb{E}\left[\|\delta Y_i\|^2\right]h \\
& + \sum_{i=0}^{n-1}(1+K_1h)^{n-1-i}C_1\mathbb{E}\left[\|\delta Z_i\|^2\right]h \\
\leq & K_2h\sum_{i=0}^{n-1}e^{K_1(n-i-1)h}\mathbb{E}\left[\|\delta Y_i\|^2\right] + C_1h\sum_{i=0}^{n-1}e^{K_1(n-i-1)h}\mathbb{E}\left[\|\delta Z_i\|^2\right],
\end{aligned}$$

where we used the inequality $(1+x) \leq e^x$ for $\forall x \in \mathbb{R}$. We remark that due to the coupling of Z in the drift, the last term of the right-hand side above is not present in [24].

Step 2. *Estimate for δY_n.* We employ a similar approach as in step 1. Using assumption 2 and the RMS-GM inequality, we obtain for any $\lambda_2 > 0$,

$$\begin{aligned}
\mathbb{E}\left[\|\delta Y_{i+1}\|^2\right] \geq & \mathbb{E}\left[\|\delta Y_i\|^2\right] + \int_{t_i}^{t_{i+1}}\mathbb{E}\left[\|\delta Z_t\|^2\right]\mathrm{d}t \\
& - 2h\mathbb{E}\left[\left(f\left(t_i, X_i^{1,\pi}, Y_i^{1,\pi}, Z_i^{1,\pi}\right) - f\left(t_i, X_i^{1,\pi}, Y_i^{2,\pi}, Z_i^{1,\pi}\right)\right)^\top \delta Y_i\right] \\
& - 2h\mathbb{E}\left[\left(f\left(t_i, X_i^{1,\pi}, Y_i^{2,\pi}, Z_i^{1,\pi}\right) - f\left(t_i, X_i^{2,\pi}, Y_i^{2,\pi}, Z_i^{2,\pi}\right)\right)^\top \delta Y_i\right] \\
\geq & \mathbb{E}\left[\|\delta Y_i\|^2\right] + \int_{t_i}^{t_{i+1}}\mathbb{E}\left[\|\delta Z_t\|^2\right]\mathrm{d}t - 2k^f h\mathbb{E}\left[\|\delta Y_i\|^2\right] \\
& - \left(\lambda_2\mathbb{E}\left[\|\delta Y_i\|^2\right] + \lambda_2^{-1}\left(L_x^f\mathbb{E}\left[\|\delta X_i\|^2\right] + L_z^f\mathbb{E}\left[\|\delta Z_i\|^2\right]\right)\right)h.
\end{aligned} \tag{13}$$

To deal with the integral term in the last inequality, we derive the following relation via Ito's isometry, (12) and (11)

$$\delta Z_i = \frac{1}{h}\mathbb{E}_i\left[\int_{t_i}^{t_{i+1}}\delta Z_t \mathrm{d}t\right].$$

Then by the Jensen- and Cauchy-Schwartz inequalities and the Fubini theorem, we derive a lower bound for the integral term

$$\begin{aligned}
\mathbb{E}\left[\|\delta Z_i\|^2\right]h = & \sum_{j=1}^{q}\sum_{k=1}^{m}\mathbb{E}\left[(\delta Z_i)_{j,k}^2\right]h = \sum_{j=1}^{q}\sum_{k=1}^{m}\frac{1}{h}\mathbb{E}\left[\left(\mathbb{E}_i\left[\int_{t_i}^{t_{i+1}}(\delta Z_t)_{j,k}\,\mathrm{d}t\right]\right)^2\right] \\
\leq & \sum_{j=1}^{q}\sum_{k=1}^{m}\frac{1}{h}\mathbb{E}\left[\left(\int_{t_i}^{t_{i+1}}(\delta Z_t)_{j,k}\,\mathrm{d}t\right)^2\right] \\
\leq & \sum_{j=1}^{q}\sum_{k=1}^{m}\int_{t_i}^{t_{i+1}}\mathbb{E}\left[(\delta Z_t)_{j,k}^2\right]\mathrm{d}t = \int_{t_i}^{t_{i+1}}\mathbb{E}\left[\|\delta Z_t\|^2\right]\mathrm{d}t,
\end{aligned} \tag{14}$$

where $(\cdot)_{j,k}$ denotes the (j,k)-entry of the matrix. Combining (13) with (14) gives

$$\mathbb{E}\left[\|\delta Y_{i+1}\|^2\right] \geq \left(1-\left(2k^f+\lambda_2\right)h\right)\mathbb{E}\left[\|\delta Y_i\|^2\right]+\left(1-L_z^f\lambda_2^{-1}\right)\mathbb{E}\left[\|\delta Z_i\|^2\right]h \\ - L_x^f\lambda_2^{-1}\mathbb{E}\left[\|\delta X_i\|^2\right]h. \tag{15}$$

Then for any $\lambda_2 > L_z^f \geq 0$, and sufficiently small h satisfying $\left(2k^f+\lambda_2\right)h < 1$, this implies

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

Recalling the definitions of K_3, K_4 in (6), we subsequently gather by induction that for any $0 \leq n \leq N-1$

$$\mathbb{E}\left[\|\delta Y_n\|^2\right] \leq e^{K_3(N-n)h}\mathbb{E}\left[\|\delta Y_N\|^2\right]+K_4\sum_{i=n}^{N-1}e^{K_3(i-n)h}\mathbb{E}\left[\|\delta X_i\|^2\right]h.$$

We remark that this estimate coincides with the one of [24, Lemma 1].

□

Due to the coupling of Z in the drift coefficient of the forward diffusion, we need an additional estimate to handle the extra $\mathbb{E}[\|\delta Z_i\|^2]$ term in the estimate for $\mathbb{E}[\|\delta X_i\|^2]$. One of our main contributions is to establish the following lemma for this purpose.

Lemma 2. *Under the setting of lemma 1, for any $\lambda_3 > 2mL_z^f$ and sufficiently small h, let us define*

$$C_2 := 2((h+\lambda_3^{-1})L_y^f+\lambda_3), \quad C_3 := 2(h+\lambda_3^{-1}), \quad C_4 := (1-mC_3L_z^f)^{-1}m. \tag{16}$$

Then we have $C_4 > 0$, furthermore the following estimates also hold

$$\mathbb{E}\left[\|\delta Y_i\|^2\right] \leq (1+C_2h)\mathbb{E}\left[\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\|^2\right]+C_3hL_x^f\mathbb{E}\left[\|\delta X_i\|^2\right]+C_3hL_z^f\mathbb{E}\left[\|\delta Z_i\|^2\right], \tag{17}$$

$$h\sum_{i=0}^{N-1}\mathbb{E}[\|\delta Z_i\|^2] \leq C_4\left(\sum_{i=1}^{N-1}C_2h\mathbb{E}\left[\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\|^2\right]+C_3hL_x^f\mathbb{E}\left[\|\delta X_i\|^2\right]\right)+C_4\mathbb{E}\left[\|\delta Y_N\|^2\right] \tag{18}$$

Proof. We take the squares of both sides of (10) and use the ϵ-Young inequality to get

$$\|\delta Y_i\|^2 \leq (1+\lambda_3h)\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\|^2+(1+(\lambda_3h)^{-1})\|h\delta f_i\|^2,$$

which holds for any $\lambda_3 > 0$, independent of $h > 0$. Taking expectations on both sides and using the Lipschitz continuity of f established by assumption 2 yields

$$\begin{aligned}\mathbb{E}\left[\|\delta Y_i\|^2\right] &\leq (1+\lambda_3h)\mathbb{E}\left[\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\|^2\right]+\left(h+\lambda_3^{-1}\right)h\mathbb{E}\left[\|\delta f_i\|^2\right]\\ &\leq (1+\lambda_3h)\mathbb{E}\left[\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\|^2\right]+(h+\lambda_3^{-1})h\left(L_x^f\mathbb{E}\left[\|\delta X_i\|^2\right]+L_y^f\mathbb{E}[\|\delta Y_i\|^2]\right.\\ &\qquad \left.+L_z^f\mathbb{E}\left[\|\delta Z_i\|^2\right]\right).\end{aligned}$$

Therefore by a rearrangement

$$\begin{aligned}(1-(h+\lambda_3^{-1})hL_y^f)\mathbb{E}\left[\|\delta Y_i\|^2\right] \leq& (1+\lambda_3h)\mathbb{E}\left[\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\|^2\right]\\ &+(h+\lambda_3^{-1})hL_x^f\mathbb{E}\left[\|\delta X_i\|^2\right]+\left(h+\lambda_3^{-1}\right)hL_z^f\mathbb{E}\left[\|\delta Z_i\|^2\right].\end{aligned}$$

Consequently, for any $\lambda_3 > 0$ and sufficiently small h, we obtain the following estimate, for $i = 0, 1, \ldots, N-1$

$$\mathbb{E}\left[\|\delta Y_i\|^2\right] \leq (1 + C_2 h)\mathbb{E}\left[\|\mathbb{E}_i[\delta Y_{i+1}]\|^2\right] + C_3 h L_x^f \mathbb{E}\left[\|\delta X_i\|^2\right] + C_3 h L_z^f \mathbb{E}\left[\|\delta Z_i\|^2\right], \tag{19}$$

where we used the definitions in (16). This proves (17).

Next, we derive the estimate for Z. Recalling the definition in (11), we get

$$h\delta Z_i = \mathbb{E}_i\left[\delta Y_{i+1} \Delta W_i^\top\right] = \mathbb{E}_i\left[\left(\delta Y_{i+1} - \mathbb{E}_i\left[\delta Y_{i+1}\right]\right) \Delta W_i^\top\right].$$

Taking the Frobenius norm on both sides and applying the Cauchy–Schwartz inequality then yields

$$\begin{aligned} h\left\|\delta Z_i\right\| &= \left\|\mathbb{E}_i\left[\left(\delta Y_{i+1} - \mathbb{E}_i\left[\delta Y_{i+1}\right]\right) \Delta W_i^\top\right]\right\| \\ &\leq \left(\mathbb{E}_i\left[\left\|\delta Y_{i+1} - \mathbb{E}_i\left[\delta Y_{i+1}\right]\right\|^2\right]\right)^{\frac{1}{2}} \left(\mathbb{E}_i\left[\left\|\Delta W_i^\top\right\|^2\right]\right)^{\frac{1}{2}} \\ &= \left(\mathbb{E}_i\left[\left\|\delta Y_{i+1} - \mathbb{E}_i\left[\delta Y_{i+1}\right]\right\|^2\right]\right)^{\frac{1}{2}} (hm)^{\frac{1}{2}}, \end{aligned}$$

which leads to

$$h\mathbb{E}\left[\|\delta Z_i\|^2\right] \leq m\mathbb{E}\left[\left\|\delta Y_{i+1} - \mathbb{E}_i\left[\delta Y_{i+1}\right]\right\|^2\right] = m\left(\mathbb{E}\left[\|\delta Y_{i+1}\|^2\right] - \mathbb{E}\left[\left\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\right\|^2\right]\right).$$

Summing both sides from 0 to $N-1$ and using the estimate (19), we gather

$$\begin{aligned} h\sum_{i=0}^{N-1} \mathbb{E}\left[\|\delta Z_i\|^2\right] &\leq m\sum_{i=0}^{N-1} \mathbb{E}\left[\|\delta Y_{i+1}\|^2\right] - \mathbb{E}\left[\left\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\right\|^2\right] \\ &= m\sum_{i=1}^{N-1} \mathbb{E}\left[\|\delta Y_i\|^2\right] + \mathbb{E}\left[\|\delta Y_N\|^2\right] - \mathbb{E}\left[\left\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\right\|^2\right] - \mathbb{E}\left[\left\|\mathbb{E}_0\left[\delta Y_1\right]\right\|^2\right] \\ &\leq m\left(\sum_{i=1}^{N-1} \mathbb{E}\left[\|\delta Y_i\|^2\right] - \mathbb{E}\left[\left\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\right\|^2\right]\right) + m\mathbb{E}\left[\|\delta Y_N\|^2\right] \\ &\leq m\left(\sum_{i=1}^{N-1} C_2 h\mathbb{E}\left[\left\|\mathbb{E}_i\left[\delta Y_{i+1}\right]\right\|^2\right] + C_3 h L_x^f \mathbb{E}\left[\|\delta X_i\|^2\right] + C_3 h L_z^f \mathbb{E}\left[\|\delta Z_i\|^2\right]\right) \\ &\quad + m\mathbb{E}\left[\|\delta Y_N\|^2\right]. \end{aligned}$$

Recalling C_4 in (16), it is easy to check that for any $\lambda_3 > 2mL_z^f$ and sufficiently small $h > 0$, we have $C_4 > 0$ and therefore

$$h\sum_{i=0}^{N-1} \mathbb{E}\left[\|\delta Z_i\|^2\right] \leq C_4\left(\sum_{i=1}^{N-1} C_2 h\mathbb{E}\left[\|\mathbb{E}_i[\delta Y_{i+1}]\|^2\right] + C_3 h L_x^f \mathbb{E}\left[\|\delta X_i\|^2\right]\right) + C_4\mathbb{E}\left[\|\delta Y_N\|^2\right].$$

□

Remark 4. *As shown transparently in the proofs of lemma 1 and 2, the constants C_j, $j = 1, 2, 3, 4$ appear because of the Z coupling in the drift. Conversely, the constants K_j, $j = 1, 2, 3, 4$ are present even in the less general case of only Y coupling, and they are consistent with [24]. In order to emphasize the difference, we denoted these by different letters.*

With these auxiliary results, and particularly lemma 2, we are ready to state our main result, a posteriori error estimate, generalizing the convergence of the deep BSDE method.

Theorem 3 (Convergence of the deep BSDE method)**.** *Suppose Assumptions 1-5 hold, and define*

$$\bar{B} := e^{\max(-K_1T,0)}\bar{C}_1\bar{C}_4\bar{C}_3\frac{e^{\bar{K}_1T}-1}{\bar{K}_1} + e^{\max(-K_1T,0)}\bar{C}_1\bar{C}_4L_x^g(1+\lambda_4)e^{\bar{K}_1T}, \tag{20}$$

$$\begin{aligned}\bar{A} := &\left(L_x^g(1+\lambda_4)e^{(\bar{K}_1+\bar{K}_3)T} + \frac{\bar{K}_4}{\bar{K}_1+\bar{K}_3}\left(e^{(\bar{K}_1+\bar{K}_3)T}-1\right)\right) \\ &\times\left(1-\bar{B}\right)^{-1}\left(\bar{K}_2\frac{1-e^{-(\bar{K}_1+\bar{K}_3)T}}{\bar{K}_1+\bar{K}_3} + e^{\max(-K_1T,0)}\bar{C}_1\bar{C}_4\bar{C}_2\frac{1-e^{-\bar{K}_3T}}{\bar{K}_3}\right),\end{aligned} \tag{21}$$

where $\bar{K}_j := \lim_{h\to 0} K_j, \bar{C}_j := \lim_{h\to 0} C_j$ *for* $j = 1,2,3,4$*. If*

$$\inf_{\lambda_1>0,\lambda_2>L_z^f,\lambda_3>2mL_z^f,\lambda_4>0} \max(\bar{B},\bar{A}) < 1, \tag{22}$$

then there exists a constant $C > 0$*, depending only on* $\mathbb{E}[\|x_0\|^2], \mathscr{L}, T, \lambda_1, \lambda_2, \lambda_3$ *and* λ_4*, such that for sufficiently small* h

$$\sup_{t\in[0,T]}\left(\mathbb{E}[\|X_t-\hat{X}_t^\pi\|^2] + \mathbb{E}[\|Y_t-\hat{Y}_t^\pi\|^2]\right) + \int_0^T \mathbb{E}[\|Z_t-\hat{Z}_t^\pi\|^2]\mathrm{d}t \le C\left(h+\mathbb{E}[\|g\left(X_T^\pi\right)-Y_T^\pi\|^2]\right), \tag{23}$$

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* $t \in [t_i, t_{i+1})$*.*

Proof. Let $X_{t_i}^{\pi,1} = X_{t_i}^\pi, Y_{t_i}^{\pi,1} = Y_{t_i}^\pi, Z_{t_i}^{\pi,1} = Z_{t_i}^\pi$ given by the Euler scheme (2b), and $X_{t_i}^{\pi,2} = \bar{X}_{t_i}^\pi, Y_{t_i}^{\pi,2} = \bar{Y}_{t_i}^\pi, Z_{t_i}^{\pi,2} = \bar{Z}_{t_i}^\pi$ given by implicit scheme (4). Both of them solve (5), and therefore we can apply lemma 1 to bound their differences. In what follows, we use the same notations as in the proof of lemma 1.

First, using the RMS-GM inequality, for any $\lambda_4 > 0$ we get

$$\mathbb{E}\left[\|\delta Y_N\|^2\right] \equiv \mathbb{E}\left[\|g(\bar{X}_T^\pi)-Y_T^\pi\|^2\right] \le (1+\lambda_4^{-1})\mathbb{E}\left[\|g(X_T^\pi)-Y_T^\pi\|^2\right] + L_x^g(1+\lambda_4)\mathbb{E}\left[\|\delta X_N\|^2\right]. \tag{24}$$

Let

$$\mathcal{X} := \max_{0\le n\le N} e^{-K_1nh}\mathbb{E}\left[\|\delta X_n\|^2\right], \qquad \mathcal{Y} := \max_{0\le n\le N} e^{K_3nh}\mathbb{E}\left[\|\delta Y_n\|^2\right]. \tag{25}$$

From estimate (8) in lemma 1, we derive the following by multiplying with e^{K_3nh} on both sides

$$\begin{aligned}&e^{K_3nh}\mathbb{E}\left[\|\delta Y_n\|^2\right] \\ \le& e^{K_3T}\mathbb{E}\left[\|\delta Y_N\|^2\right] + K_4\sum_{i=n}^{N-1} e^{K_3ih}\mathbb{E}\left[\|\delta X_i\|^2\right]h \\ \le& e^{K_3T}\left(\left(1+\lambda_4^{-1}\right)\mathbb{E}\left[\|g(X_T^\pi)-Y_T^\pi\|^2\right] + L_x^g(1+\lambda_4)\mathbb{E}\left[\|\delta X_N\|^2\right]\right) + K_4\sum_{i=n}^{N-1} e^{K_3ih}\mathbb{E}\left[\|\delta X_i\|^2\right]h \\ \le& e^{K_3T}(1+\lambda_4^{-1})\mathbb{E}\left[\|g(X_T^\pi)-Y_T^\pi\|^2\right] + \left(L_x^g(1+\lambda_4)e^{(K_1+K_3)T} + K_4\sum_{i=n}^{N-1} e^{(K_1+K_3)ih}h\right)\mathcal{X},\end{aligned}$$

where we used the definition of $\mathcal{X}$ in (25) and the estimate (24) in the last inequality. Maximizing over n subsequently yields

$$\mathcal{Y} \leq e^{K_3 T}(1+\lambda_4^{-1})\mathbb{E}\left[\|g(X_T^\pi) - Y_T^\pi\|^2\right] + \left(L_x^g(1+\lambda_4)e^{(K_1+K_3)T} + K_4 h \frac{e^{(K_1+K_3)T}-1}{e^{(K_1+K_3)h}-1}\right)\mathcal{X}. \tag{26}$$

We approach δX_n in the same manner, and from (7) collect

$$\begin{aligned} e^{-K_1 nh}\mathbb{E}\left[\|\delta X_n\|^2\right] &\leq K_2 \sum_{i=0}^{n-1} e^{-K_1(i+1)h}\mathbb{E}\left[\|\delta Y_i\|^2\right] h + C_1 \sum_{i=0}^{n-1} e^{-K_1(i+1)h}\mathbb{E}\left[\|\delta Z_i\|^2\right] h \\ &\leq K_2 \mathcal{Y} \sum_{i=0}^{n-1} e^{-K_1(i+1)h - K_3 ih} h + C_1 \sum_{i=0}^{n-1} e^{-K_1(i+1)h}\mathbb{E}\left[\|\delta Z_i\|^2\right] h. \end{aligned}$$

Additionally, from (17) and (18) we get

$$\begin{aligned} &C_1 \sum_{i=0}^{n-1} e^{-K_1 h(i+1)}\mathbb{E}\left[\|\delta Z_i\|^2\right] h \\ \leq &e^{\max(-K_1 T, 0)} C_1 \sum_{i=0}^{N-1} \mathbb{E}\left[\|\delta Z_i\|^2\right] h \\ \leq &e^{\max(-K_1 T, 0)} C_1 C_4 \left(\sum_{i=1}^{N-1} C_2 h \mathbb{E}\left[\|\delta Y_{i+1}\|^2\right] + \sum_{i=1}^{N-1} C_3 h L_x^f \mathbb{E}\left[\|\delta X_i\|^2\right] + \mathbb{E}\left[\|\delta Y_N\|^2\right]\right) \\ \leq &e^{\max(-K_1 T, 0)} C_1 C_4 h \left(C_2 \sum_{i=1}^{N-1} e^{-K_3(i+1)h}\mathcal{Y} + C_3 \sum_{i=1}^{N-1} e^{K_1 ih}\mathcal{X}\right) + e^{\max(-K_1 T, 0)} C_1 C_4 \mathbb{E}\left[\|\delta Y_N\|^2\right] \\ \leq &e^{\max(-K_1 T, 0)} C_1 C_4 h \left(C_2 e^{-K_3 h}\frac{e^{-K_3 T}-1}{e^{-K_3 h}-1}\mathcal{Y} + C_3 \frac{e^{K_1 T}-1}{e^{K_1 h}-1}\mathcal{X}\right) + e^{\max(-K_1 T, 0)} C_1 C_4 \mathbb{E}\left[\|\delta Y_N\|^2\right], \end{aligned}$$

where we recall the definitions in (6) and (16). Combining these inequalities and applying estimate (24), we obtain the following by maximizing over n

$$\begin{aligned} \mathcal{X} \leq &K_2 \mathcal{Y} h e^{-K_1 h}\frac{e^{-(K_1+K_3)T}-1}{e^{-(K_1+K_3)h}-1} + e^{\max(-K_1 T,0)} C_1 C_4 h C_2 e^{-K_3 h}\frac{e^{-K_3 T}-1}{e^{-K_3 h}-1}\mathcal{Y} \\ &+ e^{\max(-K_1 T,0)} C_1 C_4 h C_3 \frac{e^{K_1 T}-1}{e^{K_1 h}-1}\mathcal{X} + e^{\max(-K_1 T,0)} C_1 C_4 L_x^g (1+\lambda_4) e^{K_1 T}\mathcal{X} \\ &+ e^{\max(-K_1 T,0)} C_1 C_4 (1+\lambda_4^{-1})\mathbb{E}\left[\|g\left(X_T^\pi\right) - Y_T^\pi\|^2\right]. \end{aligned} \tag{27}$$

In order to simplify the expressions, we define

$$\begin{aligned} &A_1(h) \coloneqq e^{K_3 T}(1+\lambda_4^{-1}), \quad A_2(h) \coloneqq L_x^g(1+\lambda_4)e^{(K_1+K_3)T} + K_4 h \frac{e^{(K_1+K_3)T}-1}{e^{(K_1+K_3)h}-1}, \\ &A_3(h) \coloneqq K_2 h e^{-K_1 h}\frac{e^{-(K_1+K_3)T}-1}{e^{-(K_1+K_3)h}-1}, \quad A_4(h) \coloneqq e^{\max(-K_1 T,0)} C_1 C_4 C_2 h e^{-K_3 h}\frac{e^{-K_3 T}-1}{e^{-K_3 h}-1}, \\ &A_5(h) \coloneqq e^{\max(-K_1 T,0)} C_1 C_4 C_3 h \frac{e^{K_1 T}-1}{e^{K_1 h}-1}, \quad A_6(h) \coloneqq e^{\max(-K_1 T,0)} C_1 C_4 L_x^g (1+\lambda_4) e^{K_1 T}, \\ &A_7(h) \coloneqq e^{\max(-K_1 T,0)} C_1 C_4 (1+\lambda_4^{-1}). \end{aligned} \tag{28}$$

Consequently, (26) and (27) read as follows

$$\mathcal{Y} \le A_1(h)\mathbb{E}\left[\|g\left(X_T^\pi\right) - Y_T^\pi\|^2\right] + A_2(h)\mathcal{X}, \tag{29}$$

$$\mathcal{X} \le A_3(h)\mathcal{Y} + A_4(h)\mathcal{Y} + A_5(h)\mathcal{X} + A_6(h)\mathcal{X} + A_7(h)\mathbb{E}\left[\|g\left(X_T^\pi\right) - Y_T^\pi\|^2\right]. \tag{30}$$

Next, we solve (29)-(30) such that $\mathcal{Y}$ and $\mathcal{X}$ are both controlled by $\mathbb{E}[\|g(X_T^\pi) - Y_T^\pi\|^2]$. Let

$$B(h) \coloneqq A_5(h) + A_6(h), \quad A(h) \coloneqq A_2(h)\left(1 - A_5(h) - A_6(h)\right)^{-1}\left(A_3(h) + A_4(h)\right). \tag{31}$$

Whenever $B(h) < 1$, rearranging the terms in (30) yields

$$\mathcal{X} \le \left((1 - A_5(h) - A_6(h)\right)^{-1}\left(\left(A_3(h) + A_4(h)\right)\mathcal{Y} + A_7(h)\mathbb{E}\left[\|g\left(X_T^\pi\right) - Y_T^\pi\|^2\right]\right). \tag{32}$$

Additionally, if also $A(h) < 1$, we can derive the following by substituting (32) into (29)

$$\mathcal{Y} \le \left(1 - A(h)\right)^{-1}\left(A_1(h) + A_2(h)\left(1 - B(h)\right)^{-1}A_7(h)\right)\mathbb{E}\left[\|g\left(X_T^\pi\right) - Y_T^\pi\|^2\right]. \tag{33}$$

From (6) and (16), we directly collect the limits

$$\begin{aligned} &\bar{K}_1 = 2k^b + \lambda_1 + L_x^\sigma, \quad \bar{K}_2 = \lambda_1^{-1}L_y^b + L_y^\sigma, \quad \bar{K}_3 = 2k^f + \lambda_2, \quad \bar{K}_4 = L_x^f\lambda_2^{-1}, \\ &\bar{C}_1 = \lambda_1^{-1}L_z^b, \quad \bar{C}_2 = 2(\lambda_3^{-1}L_y^f + \lambda_3), \quad \bar{C}_3 = 2\lambda_3^{-1}, \quad \bar{C}_4 = \frac{m}{1 - 2mL_z^f\lambda_3^{-1}}. \end{aligned} \tag{34}$$

Consequently, from (28) we directly have

$$\begin{aligned} &\bar{A}_1 = e^{\bar{K}_3T}(1 + \lambda_4^{-1}), \quad \bar{A}_2 = L_x^g(1 + \lambda_4)e^{\left(\bar{K}_1 + \bar{K}_3\right)T} + \frac{\bar{K}_4}{\bar{K}_1 + \bar{K}_3}\left(e^{\left(\bar{K}_1 + \bar{K}_3\right)T} - 1\right), \\ &\bar{A}_3 = \bar{K}_2\frac{1 - e^{-\left(\bar{K}_1 + \bar{K}_3\right)T}}{\bar{K}_1 + \bar{K}_3}, \quad \bar{A}_4 = e^{\max(-\bar{K}_1T,0)}\bar{C}_1\bar{C}_4\bar{C}_2\frac{1 - e^{-\bar{K}_3T}}{\bar{K}_3}, \\ &\bar{A}_5 = e^{\max(-\bar{K}_1T,0)}\bar{C}_1\bar{C}_4\bar{C}_3\frac{e^{\bar{K}_1T} - 1}{\bar{K}_1}, \quad \bar{A}_6 = e^{\max(-\bar{K}_1T,0)}\bar{C}_1\bar{C}_4L_x^g(1 + \lambda_4)e^{\bar{K}_1T}, \\ &\bar{A}_7 = e^{\max(-\bar{K}_1T,0)}\bar{C}_1\bar{C}_4(1 + \lambda_4^{-1}), \end{aligned} \tag{35}$$

with the convention $\bar{A}_j = \lim_{h\to 0} A_j(h)$, $j = 1, \dots, 7$. If $\bar{K}_1 < 0$ the expressions above hold only for sufficiently small h such that $K_1 < 0$. Using the definitions in (31), it is straightforward to check that $\lim_{h\to 0} B(h) =: \bar{B}$ and $\lim_{h\to 0} A(h) =: \bar{A}$, given by (20) and (21), respectively. From (25), we get

$$\max_{0\le n\le N}\mathbb{E}\left[\|\delta X_n\|^2\right] \le e^{\max(K_1T,0)}\mathcal{X}, \quad \max_{0\le n\le N}\mathbb{E}\left[\|\delta Y_n\|^2\right] \le e^{\max(-K_3T,0)}\mathcal{Y}, \tag{36}$$

with K_1, K_3 defined in (6) both depending on h. When $\bar{B} < 1$ and $\bar{A} < 1$, we have that for any sufficiently small h (33) holds true. Hence, combining (33) with (36), we derive that for any sufficiently small h

$$\max_{0\le n\le N}\mathbb{E}\left[\|\delta Y_n\|^2\right] \le C\left(\lambda_1, \lambda_2, \lambda_3, \lambda_4\right)\mathbb{E}\left[\|g\left(X_T^\pi\right) - Y_T^\pi\|^2\right], \tag{37}$$

with a constant independent of h, depending only on the limits defined in (34), (20), (21), (35), and thus implicitly on $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. Similarly, when $\bar{B}, \bar{A} < 1$, combining (32) with (33), we also have that for any sufficiently small h

$$\max_{0\leq n\leq N} \mathbb{E}\left[\|\delta X_n\|^2\right] \leq C\left(\lambda_1, \lambda_2, \lambda_3, \lambda_4\right) \mathbb{E}\left[\|g\left(X_T^\pi\right) - Y_T^\pi\|^2\right], \tag{38}$$

for a constant determined by (35) which is independent of h.

In order to estimate $\mathbb{E}[\|\delta Z_n\|^2]$, we consider (15) from the proof of lemma 1. Notice that $1 - L_z^f/\lambda_2 > 0$ since we require $\lambda_2 > L_z^f \geq 0$, then by rearranging the terms in (15) we obtain

$$\begin{aligned}\mathbb{E}\left[\|\delta Z_i\|^2\right] h \leq \left(1 - L_z^f \lambda_2^{-1}\right)^{-1} &\left(L_x^f \lambda_2^{-1} \mathbb{E}\left[\|\delta X_i\|^2\right] h + \mathbb{E}\left[\|\delta Y_{i+1}\|^2\right] - \mathbb{E}\left[\|\delta Y_i\|^2\right]\right.\\ &\left. + \left(1 - \left(2k^f + \lambda_2\right) h\right) \mathbb{E}\left[\|\delta Y_i\|^2\right]\right).\end{aligned}$$

Summing from 0 to $N-1$ and taking the maximum on the right hand side, we gather

$$\begin{aligned}\sum_{i=0}^{N-1} \mathbb{E}\left[\|\delta Z_i\|^2\right] h &\leq \left(1 - L_z^f \lambda_2^{-1}\right)^{-1} \left(L_x^f \lambda_2^{-1} T \max_{0\leq n\leq N} \mathbb{E}\left[\|\delta X_n\|^2\right]\right.\\ &\qquad \left. + \left(\max\{(2k^f + \lambda_2)T, 0\} + 1\right) \max_{0\leq n\leq N} \mathbb{E}\left[\|\delta Y_n\|^2\right]\right)\\ &\leq C(\lambda_1, \lambda_2, \lambda_3, \lambda_4) \mathbb{E}\left[\|g\left(X_T^\pi\right) - Y_T^\pi\|^2\right],\end{aligned} \tag{39}$$

using (38), (37).

Finally, combining estimates (38), (37) and (39) with theorem 2, we prove our statement. □

4. Interpretation of the conditions in theorem 3

In this section we apply theorem 3 to special cases of FBSDEs and discuss how the conditions change depending on the coefficients in (1). Furthermore, we illustrate the role of the abstract conditions imposed by (22), and discuss several important heuristic settings under which they are satisfied.

(1) *Decoupled FBSDE*

In this case, $L_y^b = L_z^b = L_y^\sigma \equiv 0$, which immediately implies that $\bar{B} = \bar{A} \equiv 0$, since both $\bar{C}_1 = 0$ and $\bar{K}_2 = 0$. Estimates (32), (33) then reduce to

$$\mathcal{X} = 0, \qquad \mathcal{Y} \leq e^{K_3 T}(1 + \lambda_4^{-1}) \mathbb{E}\left[\|g\left(X_T^\pi\right) - Y_T^\pi\|^2\right].$$

Consequently, the total errors in the SDE reduce to those of the Euler-Maruyama discretization from (4), whereas for the BSDE part a posteriori error term remains in (23).

(2) *Coupled FBSDE with only Y coupling*

In this case $L_z^b = 0$ and therefore $\bar{C}_1 = 0$. We remark that we fully recover the result of [24] in this setting, in particular

$$\bar{B} \equiv 0, \quad \bar{A} = \left(L_x^g(1 + \lambda_4) e^{(\bar{K}_1 + \bar{K}_3)T} + \frac{\bar{K}_4}{\bar{K}_1 + \bar{K}_3}\left(e^{(\bar{K}_1 + \bar{K}_3)T} - 1\right)\right)\left(\bar{K}_2 \frac{1 - e^{-(\bar{K}_1 + \bar{K}_3)T}}{\bar{K}_1 + \bar{K}_3}\right),$$

where the condition $\bar{B} < 1$ becomes redundant and is automatically satisfied, whereas $\bar{A}$ has the same expression as the one derived in [24]. Moreover, we recover the weak and monotonicity conditions as in [24, remark 6] which guarantee $\bar{A} < 1$.

(3) *Coupled FBSDE in general as* (1)

This is the general setting we considered throughout this paper corresponding to (1). In order to guarantee that the conditions of theorem 3, and in particular (22) are satisfied we need certain requirements about T, the constants in $\mathcal{L}$, and choose λ_1, λ_2, λ_3 and λ_4 in an appropriate way. Recall that $\bar{B} \equiv \bar{B}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ and $\bar{A} \equiv \bar{A}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are functions of all λs, defined by (20) and (21), respectively. We divide the discussion into the following five cases which all have important physical interpretations.

(a) *Small time duration.* Suppose all other constants, λ_2, λ_3 and λ_4 are fixed. If $T > 0$ is sufficiently small, we can choose, for instance, $\lambda_1 = 1/\sqrt{T}$ which implies that $\bar{B}$ is sufficiently close to zero due to the factors $\bar{C}_1$ and $e^{\bar{K}_1 T} - 1$. Similarly $\bar{A}$ is sufficiently close to zero as well, due to the scaling factors $\bar{K}_2$, $1 - e^{-(\bar{K}_1 + \bar{K}_3)T}$, $\bar{C}_1$, $1 - e^{\bar{K}_3 T}$ in the last term of (21). Therefore (22) is satisfied for sufficiently small time durations T.

(b) *Weak coupling from BSDE to SDE.* Suppose all other constants, λ_1, λ_2, λ_3 and λ_4 are fixed. If $L_y^b > 0$, $L_z^b > 0$ and $L_y^\sigma > 0$ are sufficiently small, then so are the factors $\bar{C}_1$ and $\bar{K}_2$. Notice that $\bar{B}$ is scaled by $\bar{C}_1$, and for $\bar{A}$, the last term in (21) is scaled by both $\bar{C}_1$ and $\bar{K}_2$, and therefore both $\bar{A}$ and $\bar{B}$ are sufficiently close to zero and (22) holds.

(c) *Weak coupling from SDE to BSDE.* Suppose all other constants, λ_1, λ_2, λ_3 and λ_4 are fixed. If $L_x^g > 0$ and $L_x^f > 0$ are sufficiently small, and additionally $L_z^b > 0$ is sufficiently small as well, then both $\bar{B}$ and $\bar{A}$ could be sufficiently close to zero, due to the scaling factors $\bar{C}_1$ and L_x^g in $\bar{B}$, and L_x^g and $\bar{K}_4$ in the first term of (21) for $\bar{A}$. Consequently, (22) is satisfied.

(d) *Monotonicity in b.* Suppose all other constants, λ_3 and λ_4 are fixed, $k^b < 0$ is sufficiently negative and $L_z^b > 0$ is sufficiently small. We set $\lambda_1 = -2k^b - \epsilon > 0$ which implies $\bar{K}_1 = -\epsilon + L_x^\sigma < 0$ is fixed for any chosen $\epsilon > L_x^\sigma$. Then $\bar{B}$ could be sufficiently close to 0 since $\bar{K}_1 < 0$ is fixed and $\bar{C}_1 = \lambda_1^{-1} L_z^b$ could be sufficiently small. For $\bar{A}$, we directly compute from (21)

$$
\begin{aligned}
\bar{A} = \left(1 - \bar{B}\right)^{-1} \Bigg(& e^{\max(-K_1 T, 0)} \bar{C}_1 \bar{C}_4 \bar{C}_2 \frac{1 - e^{-\bar{K}_3 T}}{\bar{K}_3} \left(L_x^g (1 + \lambda_4^{-1}) e^{(\bar{K}_1 + \bar{K}_3)T} \right) \\
& + e^{\max(-K_1 T, 0)} \bar{C}_1 \bar{C}_4 \bar{C}_2 \frac{1 - e^{-\bar{K}_3 T}}{\bar{K}_3} \left(\frac{\bar{K}_4}{\bar{K}_1 + \bar{K}_3} \left(e^{(\bar{K}_1 + \bar{K}_3)T} - 1 \right) \right) \\
& + L_x^g (1 + \lambda_4^{-1}) \bar{K}_2 \frac{e^{(\bar{K}_1 + \bar{K}_3)T} - 1}{\bar{K}_1 + \bar{K}_3} + \bar{K}_2 \bar{K}_4 \frac{e^{(\bar{K}_1 + \bar{K}_3)T} + e^{-(\bar{K}_1 + \bar{K}_3)T} - 2}{(\bar{K}_1 + \bar{K}_3)^2} \Bigg).
\end{aligned} \tag{40}
$$

Let us first consider the last two terms in (40). Notice that we could fix some sufficiently large $\epsilon > L_x^\sigma$ and $\lambda_2 > L_z^f$ such that $\bar{K}_1 + \bar{K}_3 = -\epsilon + L_x^\sigma + 2k^f + \lambda_2 < 0$ is fixed and negative enough, and the penultimate term is sufficiently small because the fraction term is decreasing in $\bar{K}_1 + \bar{K}_3$ and λ_1 is sufficiently large. The last term could be sufficiently small as well, since $\bar{K}_1 + \bar{K}_3$ is fixed and $\bar{K}_4 = L_x^f / \lambda_2$ can be sufficiently small by choosing a large enough λ_2. For the remaining first two terms in (40), notice that $\bar{K}_1 + \bar{K}_3$ is fixed, and for the previously chosen λ_2, $\bar{K}_3$ is fixed as well, then both two terms are scaled by $\bar{C}_1$ with a sufficiently small L_z^b. Combining all these arguments we have $\bar{A} < 1$ and conclude that (22) is verified.

(e) *Monotonicity in f.* Suppose all other constants, λ_1, λ_3 and λ_4 are fixed. If $k^f < 0$ is sufficiently negative, and $L_z^b > 0$ is sufficiently small, then it is easy to see that $\bar{B}$ could be sufficiently close to 0 due to the scaling factor $\bar{C}_1 = \lambda_1^{-1} L_z^b$ and the fact that $\bar{B}$ does not depend on k^f and λ_2. To deal with $\bar{A}$, we set $\lambda_2 = -2k^f - \epsilon$, where $\epsilon > 0$ is chosen such that $\bar{K}_1 + \bar{K}_3 =$

$2k^b + \lambda_1 + L_x^\sigma - \epsilon < 0$ is negative enough and the penultimate term in (40) is sufficiently small. Since now $\bar{K}_1 + \bar{K}_3$ and $\bar{K}_3 = -\epsilon$ are fixed, the remaining three terms in $\bar{A}$ are all scaled by $\bar{C}_1$ and therefore we have $\bar{A} < 1$ guaranteeing (22).

(4) *Coupled FBSDE with b only depending on Z*

In this case $L_y^b = 0$. However, all constants $\bar{K}_j$, $\bar{C}_j$ for $j = 1, 2, 3, 4$ defined in (6), (16) are not zero in general, and therefore the conditions fall back under the general case discussed above. Even in a more special setting with $L_y^b = L_y^\sigma = L_y^f = k^f = 0$, i.e. when there is no Y coupling in neither the forward nor backward equation, the conditions do not seem to be easier to satisfy. Specifically, we have $\bar{K}_2 = 0$ in this special setting and consequently there will be one term less in $\bar{A}$ given by (21), but the expression (20) for $\bar{B}$ remains unchanged as it does not depend on $\bar{K}_2$. On one hand, this reduces some efforts due to the missing term in $\bar{A}$, for example, we do not take care of the last terms discussed in ((3)d) as they vanish. On the other hand, as $k^f = 0$ in this special case, we lost one possible way to make the conditions hold, i.e. ((3)e) does not apply anymore. In overall, we conclude that coupling in Z even special cases induces the need to treat the conditions in theorem 3 under the general framework established by our convergence result.

Remark 5. *The above five cases in (3) may be viewed as a generalization of the weak and monotonicity conditions stated in [24, 17]. One should note that, because of the extra Z coupling, we have to pay an extra price in establishing these five cases, e.g., we need to choose λs appropriately instead of fixing them as constants as in [24], and in particular, we require $L_z^b > 0$ to be sufficiently small for (3)c, (3)d and (3)e.*

Remark 6. *In [24, remark 2], it is claimed that it is general to consider drifts which only depend on X and Y but not on Z, since, by an appropriate change of probability measure, one can always reformulate* (1) *into an equivalent FBSDE whose drift is independent of Z yet its solution coincides with the same quasi-linear PDE. However, we find that there are several important issues with this statement both theoretically and numerically:*

(1) such approach would change the probability measure under which X is simulated, consequently the training of neural networks would be carried out in a different spatial region, and therefore the algorithm may have poor accuracy around the area of the domain of interest;

(2) it is common that the reformulated FBSDE does not satisfy the theoretical assumptions needed for convergence while the original one does. For instance, a linear z term in b would result in a quadratic term in the reformulated driver, which violates the assumptions of Lipschitz continuity. We illustrate this through Example 5.1 in section 5 below. Therefore, we believe our theory is a necessary generalization to [24] and it is applicable to a wider class of FBSDEs;

(3) with the same approach, one could remove the entire drift to the driver, and simulate a reformulated FBSDE with zero drift, but this, for the two reasons above, is rarely done in practice.

Finally, let us derive a lower bound for $\bar{B}$ in (20) by computing the infimum of $\bar{B}$ over all possible choices of λs. Notice that $\bar{B}$ does not depend on λ_2, decreases in λ_3 and increases in λ_4, and therefore we shall mainly look at λ_1. Let

$$\bar{B}_\ell := \inf_{\lambda_1 > 0, \lambda_3 > 2mL_z^f, \lambda_4 > 0} \bar{B}. \tag{41}$$

(1) If $\lambda_1 \geq \max(-2k^b - L_x^\sigma, 0)$, then $\bar{B}$ admits a unique stationary point along the λ_1 direction, and

$$\bar{B}_\ell = mL_z^b L_x^g e^{(2k^b + \lambda_1 + L_x^\sigma)T} T, \qquad \arg \inf_{\lambda_1, \lambda_3, \lambda_4} \bar{B} = (1/T, +\infty, 0).$$

(2) If $0 < \lambda_1 < \max(-2k^b - L_x^\sigma, 0)$, then $\bar{B}$ is convex in λ_1 but there is not stationary point in this range, and

$$\bar{B}_\ell = m\frac{L_z^b}{-2k^b - L_x^\sigma}L_x^g, \qquad \arg\inf_{\lambda_1,\lambda_3,\lambda_4} \bar{B} = (-2k^b - L_x^\sigma, +\infty, 0).$$

This result is particularly useful in practice and can serve as a preliminary test for the convergence of a given FBSDE. In fact, given an equation and all its relevant Lipschitz constants, if we find $\bar{B}_\ell \geq 1$, then we know that the conditions of theorem 3 cannot be satisfied and that the deep BSDE algorithm is less likely to converge. On the other hand, if $\bar{B}_\ell$ is less than 1 or in particular even close to 0, we may check if $\bar{A} < 1$, which can be solved efficiently by a numerical constrained minimization method ranging different λ_1, λ_2, λ_3 and λ_4.

5. Numerical experiments

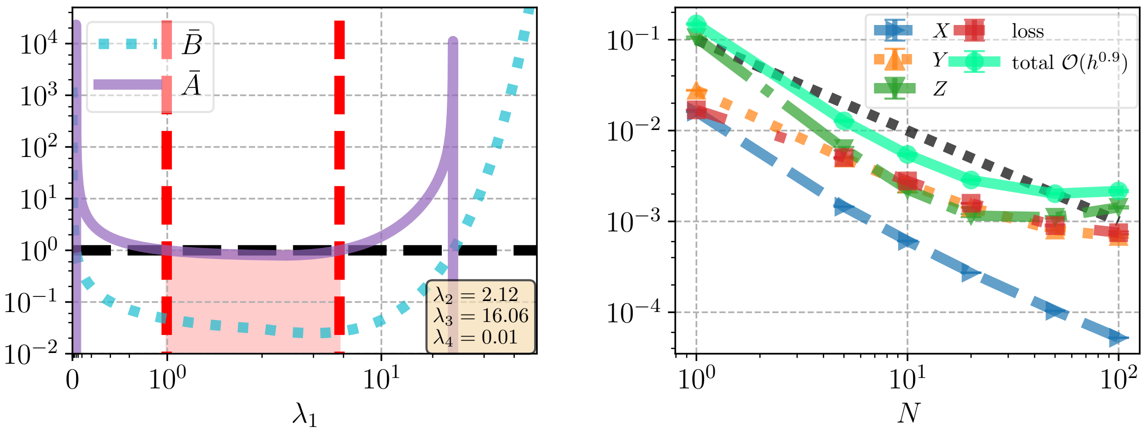
We implemented the deep BSDE method in TensorFlow 2.9. The errors correspond to the discretized version of the left hand side of (23) and are defined as follows

$$\text{error}(X) \coloneqq \max_{0\leq n\leq N} \mathbb{E}\left[\left\|\hat{X}_n^\pi - X_{t_n}\right\|^2\right], \quad \text{error}(Y) \coloneqq \max_{0\leq n\leq N} \mathbb{E}\left[\left\|\hat{Y}_n^\pi - Y_{t_n}\right\|^2\right],$$
$$\text{error}(Z) \coloneqq 1/N\sum_{n=0}^{N-1} \mathbb{E}\left[\left\|\hat{Z}_n^\pi - Z_{t_n}\right\|^2\right], \quad \text{total} = \text{error}(X) + \text{error}(Y) + \text{error}(Z).$$

We also define relative L^2 approximation errors by $\text{error}(X)/\mathbb{E}[\|X_{t_n}\|^2]$ and similarly for Y, Z, total. In what follows, the true expectations are approximated over a Monte Carlo sample of size 2^{12}. Given a classical solution to the corresponding quasi-linear PDE used for decoupling, we gather a reference solution to the associated FBSDE (1) by an Euler-Maruyama simulation with $N' = 10^4$ time steps, in order to guarantee that the time discretization error of the reference solution is negligible compared to the approximation errors incurred via the deep BSDE method. In all experiments below, we use neural networks with hyperbolic tangent activation, an input layer of width d, 2 hidden layers $30 + d$ neurons wide each, and an output layer of appropriate dimensions depending on the process approximated. As an optimization strategy, we use the Adam optimizer with default initializations and a learning rate schedule of exponential decay, starting from 10^{-2} with a decay rate of 10^{-2}. For a fixed N we perform 2^{14} SGD iterations, and for each iteration we take an independent sample of 2^{10} trajectories of the underlying Brownian motion. All experiments below were run on an Dell Alienware Aurora R10 machine, equipped with an AMD Ryzen 9 3950X CPU (16 cores, 64Mb cache, 4.7GHz) and an Nvidia GeForce RTX 3090 GPU (24Gb). In order to assess the inherent stochasticity of both the regression Monte Carlo method and the SGD iterations, we run each experiment 5 times and report on the mean and standard deviations of the resulting independent approximations. All operations were carried out with single precision.

5.1. Example 1

The following example is a modified version of the one in [17], where in order to demonstrate our theoretical extension we include Z coupling in the drift of the forward diffusion. The coefficients of



(a) $\bar{B}, \bar{A}$ as functions of λ_1 for given $(\lambda_2, \lambda_3, \lambda_4)$ in case of (42). Dotted vertical red lines mark the endpoints of the interval where $\bar{B}, \bar{A} < 1$, and the shaded red area the subset of the plane where the sufficient conditions of theorem 3 are satisfied.

(b) Convergence in N. Empirical convergence rate in labels. Dotted black line indicates the expected $\mathcal{O}(h)$ rate predicted by theorem 3.

Figure 1: Example 5.1. $T = 0.25, X_0 = (\pi/4, \dots, \pi/4)$.

the FBSDE system (1) read as follows

$$b(t,x,y,z) = \kappa_y \bar{\sigma} y \mathbf{1}_d + \kappa_z z^\top, \quad \sigma(t,x,y) = \bar{\sigma} y I_d, \quad g(x) = \sum_{i=1}^d \sin(x_i), \tag{42}$$

$$f(t,x,y,z) = -ry + 1/2 e^{-3r(T-t)} \bar{\sigma}^2 \Big(\sum_{i=1}^d \sin(x_i)\Big)^3 - \kappa_y \sum_{i=1}^d z_i - \kappa_z \bar{\sigma} e^{-3r(T-t)} \sum_{i=1}^d \sin(x_i) \sum_{i=1}^d \cos^2(x_i),$$

with $q = 1, d = m$. The analytical solution pair to the backward equation is given by

$$y(t,x) = e^{-r(T-t)} \sum_{i=1}^d \sin(x_i), \quad z_i(t,x) = e^{-2r(T-t)} \bar{\sigma} \Big(\sum_{j=1}^d \sin(x_j)\Big) \cos(x_i). \tag{43}$$

We note that the equation above falls under the theoretical assumptions of section 3. In particular, we get the following set of values for the corresponding constants $L_x^g = d, L_y^b = 2(\kappa_y \bar{\sigma})^2, L_z^b = 2\kappa_z^2, L_y^\sigma = d\bar{\sigma}^2, L_x^f = 3/2d(3\bar{\sigma}^2 d^2/2 + 2\kappa_z \bar{\sigma} d)^2, L_y^f = 18r^2, L_z^f = 3.6d\kappa_y^2, k^f = -r, k^b = L_x^\sigma = L_z^\sigma = L_x^b = 0$. The strength of coupling is determined by the values $\bar{\sigma}, r, \kappa_y, \kappa_z$. In order to satisfy the sufficient conditions of theorem 3, we put $r = 1, \bar{\sigma} = 0.1, \kappa_y = 10^{-1}, \kappa_z = 10^{-2}$ and $T = 0.25, X_0 = (\pi/4, \dots, \pi/4)$. Convergence results are collected in figure 1. Figure 1a shows that the conditions of theorem 3 are indeed satisfied, there exists a quadruple $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, such that $\bar{B}, \bar{A}$ defined by (21), (20) admit $\max(\bar{B}, \bar{A}) < 1$. In particular, for the fixed $\lambda_2, \lambda_3, \lambda_4$ we mark the interval of admissible λ_1s such that the sufficient conditions are satisfied within the shaded red area. Figure 1b depicts the convergence of the deep BSDE method. Its most important implications are as follows. The convergence is only guaranteed in a *posteriori* sense. In fact, as can be seen the convergence only shows the expected $\mathcal{O}(h)$ behavior whenever the loss function corresponding to the last term of (23) is dominated by the discretization error. In particular, for $N \in \{50, 100\}$ we see that the approximation errors of Z begin to stall and the total approximation errors are dominated by the loss function. This indicates that for very fine time grids one needs to make sure that the loss is

Table 1: Comparison on the convergence of the deep BSDE algorithm between (42) and (44). Numbers correspond to the mean(std.dev.) of the total approximation errors of 5 independent runs of the algorithm. $T = 0.25, X_0 = (\pi/4, \ldots, \pi/4)$.

N	1	5	10	20	50	100
total-Eq.(42)	1.7e−2(3e−3)	5.1e−3(3e−4)	2.8e−3(1e−4)	1.58e−3(5e−5)	9.0e−4(4e−5)	7.5e−4(6e−5)
total-Eq.(44)	1.8e−2(3e−3)	3e+2(2e+2)	2e+6(4e+6)	NaN	NaN	NaN

appropriately minimized when trying to recover discretization errors. Given the global minimization structure of the deep BSDE method the corresponding optimization problem becomes more difficult with an increased number of time steps. This demonstrates a clear trade-off between discretization and optimization, which is fully explained by theorem 3 and should be carefully considered in applications. Nevertheless, we get an empirical convergence rate of $\mathcal{O}(h^{0.9})$ for all time points, and accounting for the reasoning above we recover the predicted rate of our convergence analysis.

Furthermore, let us return to remark 6. In particular, Han and Long in [24, Remark 2] claim that the setting of Z independent drift is general, since, due to the connections with the associated quasi-linear PDEs, one can always move the Z dependence from the drift to the driver. In order to complement our arguments against this reasoning in remark 6, we provide a numerical demonstration of the points raised therein. One can derive a *reformulated* FBSDE system which is *decoupled in Z* and whose solution will coincide with (43). This equation has a modified drift and driver

$$\tilde{b}(t,x,y) = \kappa_y \bar{\sigma} y 1_d, \quad \tilde{f}(t,x,y,z) = f(t,x,y,z) + \kappa_z \|z\|^2/(\bar{\sigma}y), \tag{44}$$

whereas the rest of the coefficients remain the same as in (42). First, notice that even though (42) satisfies the theoretical assumptions of theorem 3, the reformulated FBSDE (44) does not. In particular, $\tilde{f}$ is not Lipschitz continuous in y, z which renders the results of theorem 3, or [24] as a limit case, inapplicable. This demonstrates point-(2) from remark 6. Nonetheless, as our convergence analysis only gives sufficient conditions one can still run the deep BSDE algorithm and find satisfactory results without theoretical guarantees. Table 1 shows that for equation (44) this is not case. Running the algorithm on the reformulated FBSDE (44) results in diverging errors. In fact, due to the singularity arising in the driver $\tilde{f}$, the backward equation blows up as N increases, which results in the forward equation also exploding due to the coupling. On the contrary, the original equation (42) with Z coupling converges as predicted by theorem 3 and also illustrated by figure 1. This observation demonstrates point-(1) from remark 6 and implies that the Lipschitz features in our analysis are crucial also in practice, in order to avoid such explosion of the coupled forward diffusion. This is in line with related results in the literature, see [17, pg.170]. Overall, we conclude that the framework of Z coupling in the drift cannot in general be circumvented neither theoretically nor numerically, and one needs to rely on our convergence result in theorem 3 instead.

5.2. Example 2

The following $d = 25$ dimensional example is related to a linear-quadratic stochastic control problem appearing in [25, example 3], which is defined by the following set of coefficients

$$\begin{aligned} M_x &= -\mathrm{diag}(1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1), \\ M_u &= (1,1,0.5,1,0,0,1,1,0.5,1,0,0,1,1,0.5,1,0,0,1,1,0.5,1,0,0,1)^\top, \\ M_c &= -M_x(-0.2,-0.1,0,0,0.1,0.2,-0.2,-0.1,0,0,0.1,0.2, \\ &\qquad -0.2,-0.1,0,0,0.1,0.2,-0.2,-0.1,0,0,0.1,0.2,-0.2)^\top, \end{aligned} \tag{45}$$

$$
\begin{aligned}
\Sigma &= \mathrm{diag}(0.15, 0.15, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25,\\
&\qquad 0.25, 0.15, 0.15, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25),\\
R_x &= 2\mathrm{diag}(25, 1, 25, 1, 25, 1, 25, 1, 25, 1, 25, 1, 25, 1, 25, 1, 25, 1, 25, 1, 25, 1, 25, 1, 25), \quad R_u = 2,\\
G &= 2\mathrm{diag}(25, 25, 25, 25, 25, 25, 1, 25, 1, 25, 1, 25, 25, 25, 25, 25, 25, 25, 1, 25, 1, 25, 1, 25, 1).
\end{aligned}
$$

One can derive an associated FBSDE system via either dynamic programming (DP) or the stochastic maximum principle (SMP), see e.g. [32]. The corresponding equations' coefficients in (1) then take the following form in case of DP

$$
\begin{aligned}
&b_{\mathrm{DP}}(t,x,y,z) = M_x x - M_u R_u^{-1} M_u^\top (z\Sigma^{-1})^\top, \quad \sigma_{\mathrm{DP}}(t,x,y) = \Sigma,\\
&f_{\mathrm{DP}}(t,x,y,z) = 1/2(x^\top R_x x + z\Sigma^{-1}(R_u^{-1}M_u^\top)^\top M_u^\top (z\Sigma^{-1})^\top), \quad g_{\mathrm{DP}}(x) = 1/2 x^\top G x,
\end{aligned} \tag{46}
$$

with $q = 1, d = m = 25$; and in case of SMP

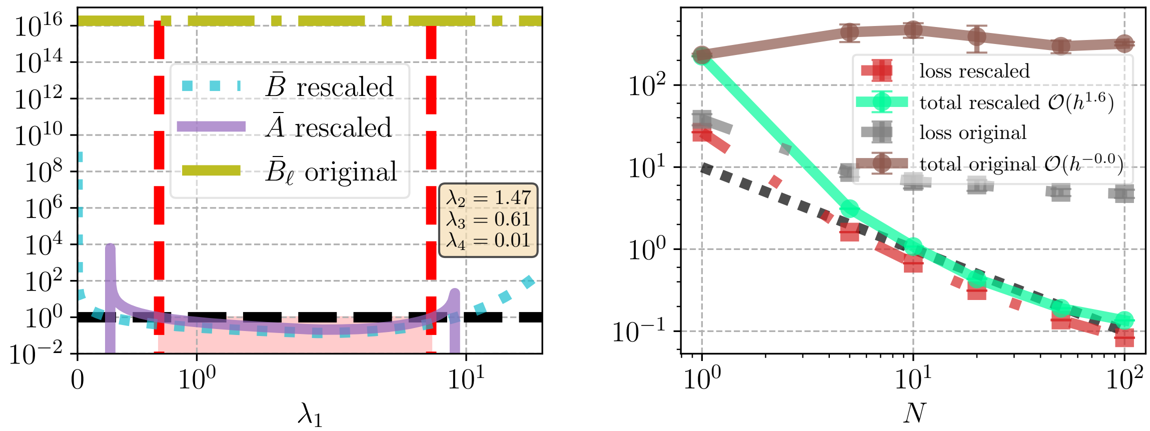
$$
\begin{aligned}
&b_{\mathrm{SMP}}(t,x,y,z) = M_x x + M_u R_u^{-1} M_u^\top y, \quad \sigma_{\mathrm{SMP}}(t,x,y) = \Sigma,\\
&f_{\mathrm{SMP}}(t,x,y,z) = -R_x x + M_x y, \quad g_{\mathrm{SMP}}(x) = -Gx,
\end{aligned} \tag{47}
$$

with $q = d = m = 25$. The main difference between the two formulations is that (46) leads to an FBSDE where coupling into b occurs through Z, whereas in (47) only through Y. Furthermore, the first equation gives a scalar-valued BSDE, whereas (47) is a vector-valued one. Both equations admit semi-analytical solutions given by the numerical resolution of a system of Ricatti ODEs. For details we refer to [28] and the references therein.

Remark 7. *Notice that the dynamic programming FBSDE* (46) *does not satisfy the Lipschitz conditions imposed in section 3. In fact, g_{DP}, f_{DP} in* (46) *are quadratic in x, z. Nevertheless, one can use a localization argument, and consider the equation over a compact domain such that the corresponding coefficients become Lipschitz continuous with a constant depending on the width of the domain. We choose truncation radiuses based on upper bounds for* 99% *quantiles of $\|X_t\|$, $\|Z_t\|$ computed over an independent Monte Carlo simulation consisting of 2^{20} paths using the semi-analytical reference solution. This results in a negligible truncation error and truncation radiuses $r_x = 1, r_z = 10$, with which we obtain a Lipschitz continuous approximation of g_{DP}, f_{DP} for which the constants in section 3 can be computed even in the case of* (46)*, and read as follows $L_x^g = r_x^2\|G\|_2^2/2, L_x^b = 2\|M_x\|_2^2, L_z^b = 2\|M_u R_u^{-1} M_u^\top (\Sigma^{-1})^\top\|_2^2, L_x^f = r_x^2\|R_x\|_2^2, L_z^f = r_z^2\|\Sigma^{-1}(R_u^{-1}M_u^\top)^\top M_u^\top (\Sigma^{-1})^\top\|_2^2, k^b = -1, L_y^b = L_x^\sigma = L_z^\sigma = L_y^f = k^f = 0$.*

5.2.1. Non-convergence of the deep BSDE method

Let us first focus on the FBSDE stemming from the dynamic programming principle. In [25] it was observed that the deep BSDE method does not converge for the FBSDE defined by (46) with coefficients as in (45) and $T = 1/2$, $X_0 = (0.1, \ldots, 0.1)$. Earlier convergence analyses such as [24] could not justify this phenomenon, as in (46) the coupling into the drift takes places via Z, which fell out of the framework of the aforementioned paper. Our generalization provided by theorem 3 enables the treatment of such equations, and in particular explains the non-convergence phenomenon. The problem lies in the *strength* of the coupling of Z into the forward diffusion. In order to demonstrate this, we consider two versions of (46), which differ in the coefficient M_u. One labelled as *"original"*, where the coefficients of the corresponding LQ problem are as in (45), and a *"rescaled"* version where M_u is replaced by $M_u/150$ and all other coefficients remain the same. Our findings are illustrated by figure 2. In particular, figure 2a depicts the contraction constants $\bar{B}, \bar{A}$ defined by (20), (21) appearing in theorem 3 for both versions. As can be seen, in case of



(a) $\bar{B}, \bar{A}$ as functions of λ_1 for given $(\lambda_2, \lambda_3, \lambda_4)$ in case of the "rescaled" version of (46). $\bar{B}_\ell$ as lower bound for $\bar{B}$ in case of the "original" version. Dotted vertical red lines mark the endpoints of the interval where $\bar{B}, \bar{A} < 1$, and the shaded red area the subset of the plane where the sufficient conditions of theorem 3 are satisfied.

(b) Convergence in N. Empirical convergence rates in labels. Dotted black line indicates the expected $\mathcal{O}(h)$ rate predicted by theorem 3.

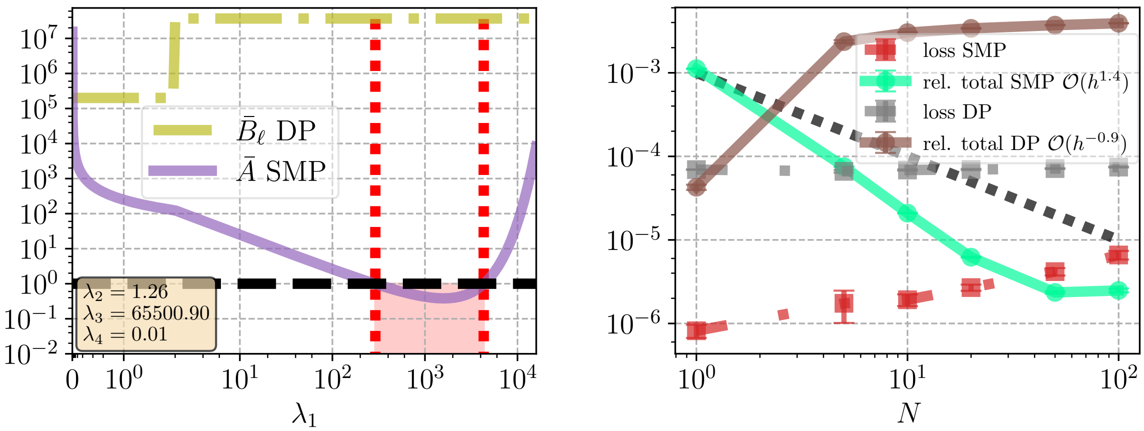
Figure 2: Example 5.2.1. Comparison of the Deep BSDE method on (46) between coefficients as in (45) (original) and M_u replaced by $M_u/150$ (rescaled). $T = 1/2, X_0 = (0.1, \dots, 0.1)$.

the "original" equation one gets a lower bound $\bar{B}_\ell$ defined by (41) which is of $\mathcal{O}(10^{16})$. In fact, this implies that the conditions of theorem 3 can never be satisfied for the original version of (46). However, as is also suggested by figure 2a, decreasing the strength of the coupling by the given rescaling of M_u we get an equation whose $\bar{B}, \bar{A}$ satisfy the sufficient conditions (22). Motivated by this, we collect the convergence of the total approximation errors in figure 2b. In line with the discussion above, we find that the deep BSDE method does not converge in the "original" case, whereas it does converge for the "rescaled" version, once we make sure that there exist appropriate contraction constants $\bar{B}, \bar{A} < 1$ such that the sufficient conditions of theorem 3 are satisfied. The empirical rate of convergence is of $\mathcal{O}(h^{1.6})$. Note that this example illustrates the weak coupling condition described in point (3)c of section 4. We emphasize that $\bar{B}$ is inherent to the Z coupling and it is the main novelty of our convergence analysis.

5.2.2. Stochastic control via DP or SMP

In [28] a convergence analysis has been given in the context of solving stochastic control problems with the deep BSDE method applied on the FBSDE system derived through the stochastic maximum principle (SMP) similar to (47). This analysis provided a natural extension to the works of Han and Long by extending [24] to vector-valued settings. In [25] it was found that for certain FBSDEs derived in the dynamic programming framework, such as (46), the Deep BSDE method does not converge. On the other hand, the authors of the present paper found in [28] that the same problems tackled by the stochastic maximum principle lead to an FBSDE (47) for which the deep BSDE method does converge to the unique solution triple. Our results in theorem 3 provide a natural explanation for these empirical findings. In fact, the problem lies in the Z coupling in (46), and in particular the value of the contraction constant $\bar{B}$ defined in (20). Conversely, (47) derived from the stochastic maximum principle has coupling only in Y, not in Z, leading to $\bar{B} \equiv 0$. In order to illustrate this, we ran the deep BSDE algorithm for both (46) and (47) with coefficients defined by (45), a short time horizon $T = 10^{-3}$ and $X_0 = (0.1, \dots, 0.1)$.

First, notice that only (47) satisfies the Lipschitz conditions imposed in section 3, with corre-



(a) $\bar{A}$ as function of λ_1 for given $(\lambda_2, \lambda_3, \lambda_4)$ in case of the SMP equation in (47). $\bar{B}_\ell$ as lower bound for $\bar{B}$ in case of the DP equation in (46). Dotted vertical red lines mark the endpoints of the interval where $\bar{A} < 1$, and the shaded red area the subset of the plane where the sufficient conditions of theorem 3 are satisfied.

(b) Convergence in N. Empirical convergence rates in labels. Dotted black line indicates the expected $\mathcal{O}(h)$ rate predicted by theorem 3.

Figure 3: Example 5.2.2. Comparison between the FBSDEs derived via dynamic programming (46) and the stochastic maximum principle (47) approximated by the deep BSDE method. Coefficients as in (45), $T = 10^{-3}, X_0 = (0.1, \dots, 0.1)$.

sponding constants $L_x^g = \|G\|_2^2, L_x^b = 2\|M_x\|_2^2, L_y^b = 2\|M_u R_u^{-1} M_u^\top\|_2^2, L_x^f = 2\|R_x\|_2^2, L_y^f = 2\|M_x\|_2^2$, $k^f = k^b = -1, L_z^b = L_x^\sigma = L_y^\sigma = L_z^\sigma = L_z^f = 0$. Hence, only the SMP formulation guarantees direct applicability of the convergence results. Nevertheless, with the localization argument in remark 7 one can find an accurate Lipschitz continuous approximation of the DP equation (46) for which the sufficient conditions can also be checked. More importantly, (46) and (47) also differ in the way of coupling. Namely, in case of the former, Z couples into the forward diffusion, whereas in case of the latter only Y does. This in particular implies that besides the different Lipschitz constants, the two equations also differ in terms of the sufficient conditions of (23). In fact, for the SMP equation (47), there is no coupling in Z which implies $\bar{B} \equiv 0$, see also discussion in section 4. Hence (47) reduces to the theoretical framework of [28, 24] under which it is sufficient for $\bar{A} < 1$ to hold. On the other hand, (46) couples through Z, which in light of theorem 3 implies that $\bar{B}, \bar{A} < 1$ need to hold simultaneously, leading to stronger conditions to hold.

The above discussion is illustrated by figure 3. From figure 3a, we find that in case of the dynamic programming equation (46) $\bar{B}$ admits a lower bound $\bar{B}_\ell$ defined by (41) which is of $\mathcal{O}(10^5)$. In particular, this implies that the dynamic programming formulation can never satisfy the sufficient conditions imposed by theorem 3. On the contrary, under the SMP formulation we find a range of $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ such that $\bar{A} < 1$ meaning that the convergence criteria are met. Motivated by these conditions, the convergence of the deep BSDE method is collected in figure 3b for both equations. In line with our previous observations, we find that the method converges with an empirical rate of $\mathcal{O}(h^{1.4})$ for the SMP equation (47), which in this particular case is even faster than the rate predicted by [24, 28]. On the other hand, for the FBSDE derived via dynamic programming we find that the deep BSDE does not converge. As shown by figure 3b this is not due to the posteriori nature of the estimate in theorem 3, as the errors are growing even as the loss functional decreases. The critical phenomenon is indeed the coupling in Z, and the extra conditions it imposes as predicted by theorem 3. These findings complement our earlier convergence results in the context of stochastic control tackled by the deep BSDE method and the stochastic maximum principle [28]. In particular, our generalization in theorem 3 explains our empirical findings in [28] suggesting that for a large class of stochastic control problems, deriving an associated FBSDE through SMP leads to a system which is more tractable by deep BSDE formulations. The crucial property here is the lack of Z coupling leading to milder conditions to ensure convergence according to our new convergence result in theorem 3.

6. Conclusion

In this paper a generalized proof for the convergence of the deep BSDE method was given. Our main contributions can be summarized as follows. We extended the convergence analysis of [24] to FBSDEs with fully-coupled drift coefficients. Such an extension is essential in practice as it enables the treatment of FBSDEs stemming from stochastic optimal control problems. Our theory provides a unified framework and, in particular, includes earlier results from the literature as limit cases. Due to the extra Z coupling, the final posteriori error estimate stated in theorem 3 requires an additional condition expressed by (22). These sufficient conditions are directly verifiable for any equation, and as shown in section 4, they are satisfied under heuristic settings such as weak coupling, short time duration or monotonicity. Moreover, as demonstrated in section 5.2.1, our theory explains the non-convergence of the deep BSDE method observed in recent literature, and provides direct guidelines to avoid such issues and ensure convergence in practice. Several numerical experiments were presented for high-dimensional equations, which support and highlight key features of the theoretical findings.

Acknowledgement. The first author acknowledges financial support from the Peter Paul Peterich Foundation via the TU Delft University Fund. The second author would like to thank the China Scholarship Council (CSC) for the financial support.

References

[1] Jianfeng Zhang and Jianfeng Zhang. *Backward stochastic differential equations*. Springer, 2017.

[2] Jin Ma and Jiongmin Yong. *Forward-Backward Stochastic Differential Equations and their Applications*, volume 1702 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[3] E. Pardoux and S. Peng. Backward stochastic differential equations and quasilinear parabolic partial differential equations. In Boris L. Rozovskii and Richard B. Sowers, editors, *Stochastic Partial Differential Equations and Their Applications*, pages 200–217, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg.

[4] Fabio Antonelli. Backward-Forward Stochastic Differential Equations. *The Annals of Applied Probability*, 3(3):777–793, 1993. Publisher: Institute of Mathematical Statistics.

[5] Etienne Pardoux and Shanjian Tang. Forward-backward stochastic differential equations and quasilinear parabolic pdes. *Probability theory and related fields*, 114:123–150, 1999.

[6] François Delarue. On the existence and uniqueness of solutions to FBSDEs in a non-degenerate case. *Stochastic Processes and their Applications*, 99(2):209–286, June 2002.

[7] F. Delarue and G. Guatteri. Weak existence and uniqueness for forward–backward SDEs. *Stochastic Processes and their Applications*, 116(12):1712–1742, December 2006.

[8] Emmanuel Gobet, Jean-Philippe Lemor, and Xavier Warin. A regression-based Monte Carlo method to solve backward stochastic differential equations. *The Annals of Applied Probability*, 15(3):2172–2202, 2005. Publisher: The Institute of Mathematical Statistics.

[9] Bruno Bouchard and Nizar Touzi. Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations. *Stochastic Processes and their Applications*, 111(2):175–206, June 2004.

[10] Christian Bender and Robert Denk. A forward scheme for backward sdes. *Stochastic processes and their applications*, 117(12):1793–1812, 2007.

[11] Christian Bender and Jessica Steiner. Least-Squares Monte Carlo for Backward SDEs. In René A Carmona, Pierre Del Moral, Peng Hu, and Nadia Oudjane, editors, *Numerical Methods in Finance*, pages 257–289, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[12] M J Ruijter and C W Oosterlee. A Fourier Cosine Method for an Efficient Computation of Solutions to BSDEs. *SIAM Journal on Scientific Computing*, 37(2):A859–A889, 2015.

[13] Jin Ma, Philip Protter, and Jiongmin Yong. Solving forward-backward stochastic differential equations explicitly — a four step scheme. *Probability Theory and Related Fields*, 98(3):339–359, September 1994.

[14] Jim Douglas, Jin Ma, and Philip Protter. Numerical Methods for Forward-Backward Stochastic Differential Equations. *The Annals of Applied Probability*, 6(3):940–968, 1996. Publisher: Institute of Mathematical Statistics.

[15] Jaksa Cvitanic and Jianfeng Zhang. The Steepest Descent Method for Forward-Backward SDEs. *Electronic Journal of Probability*, 10(none):1468–1495, January 2005. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

[16] François Delarue and Stéphane Menozzi. A forward - Backward stochastic algorithm for quasilinear PDES. *The Annals of Applied Probability*, 2006.

[17] Christian Bender and Jianfeng Zhang. Time discretization and Markovian iteration for coupled FBSDEs. *The Annals of Applied Probability*, 18(1):143–177, February 2008.

[18] T. P. Huijskens, M. J. Ruijter, and C. W. Oosterlee. Efficient numerical Fourier methods for coupled forward–backward SDEs. *Journal of Computational and Applied Mathematics*, 296:593–612, April 2016.

[19] Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

[20] Weinan E, Jiequn Han, and Arnulf Jentzen. Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations. *Communications in Mathematics and Statistics*, 5(4):349–380, December 2017.

[21] Côme Huré, Huyên Pham, and Xavier Warin. Deep backward schemes for high-dimensional nonlinear PDEs. *Mathematics of Computation*, 89(324):1547–1579, 2020.

[22] Chengfan Gao, Siping Gao, Ruimeng Hu, and Zimu Zhu. Convergence of the Backward Deep BSDE Method with Applications to Optimal Stopping Problems. *SIAM Journal on Financial Mathematics*, 14(4):1290–1303, December 2023. Publisher: Society for Industrial and Applied Mathematics.

[23] 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*, page drad092, February 2024.

[24] Jiequn Han and Jihao Long. Convergence of the deep BSDE method for coupled FBSDEs. *Probability, Uncertainty and Quantitative Risk*, 5:1–33, 2020.

[25] 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.

[26] Jared Chessari, Reiichiro Kawai, Yuji Shinozaki, and Toshihiro Yamada. Numerical methods for backward stochastic differential equations: A survey. *Probability Surveys*, 20:486–567, 2023.

[27] Yifan Jiang and Jinfeng Li. Convergence of the Deep BSDE method for FBSDEs with non-Lipschitz coefficients. *Probability, Uncertainty and Quantitative Risk*, 6(4):391–408, 2021. Publisher: Probability, Uncertainty and Quantitative Risk.

[28] Zhipeng Huang, Balint Negyesi, and Cornelis W. Oosterlee. Convergence of the deep BSDE method for stochastic control problems formulated through the stochastic maximum principle, January 2024. arXiv:2401.17472 [cs, math, q-fin].

[29] Huyên Pham. *Continuous-time stochastic control and optimization with financial applications*, volume 61. Springer Science & Business Media, 2009.

[30] Shaolin Ji, Shige Peng, Ying Peng, and Xichuan Zhang. Solving stochastic optimal control problem via stochastic maximum principle with deep learning method. *Journal of Scientific Computing*, 93(1):30, 2022.

[31] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, January 1991.

[32] Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 1999.