Mini-Mémoire de Méthodes Probabilistes Avancées en Finance

Kim Sungjin et Keming ZHANG

April 2020

1 Introduction

In this project, we will be trying to solve numerically the system of decoupled forward and backward stochastic differential equations (FBSDEs) in light of the intriguing method of deep neural networks applied on the backward stochastic differential equations reformulation of the desired semi-linear parabolic PDEs (deep BSDE developed in [1]). Then we will discuss the generalized case in which the FSDE and BSDE are coupled as in [2].

2 Preliminaries: Deep BSDE method, Feymann-Kac Theorem and Assumptions

2.1 What is Deep BSDE method and Why? Motivation and Setup

> "In view of all that we have said in the foregoing sections, the many obstacles we appear to have surmounted, what casts the pall over our victory celebration? It is the curse of dimensionality, a malediction that has plagued the scientist from the earliest days."
>
> *Richard.E.Bellman*

The curse of dimensionality, first coined by Bellman in 1957, refers to the explosive nature of properties which are not present in spaces of lower dimensions (for example 3 dimensional real world). The most well-known catastrophic consequence would be the exponential growth time for computing and a considerable waste of memory space. One may consider the abundant information offered by the high dimensionality, but the complexity it brought to the calculation overwhelms the marginal benefit. To understand intuitively the cause, we could refer to the classic stable and consistent implicit numerical schemes. These kind of schemes solving numerically PDEs usually need to proceed matrix inversion on every time discretization interval, which leads to super-polynomial increments of the need of computing time.

The curse of dimensionality has always been an obstacle preventing numerical methods for physical (or dynamic programming) PDEs and data engineering to increase their magnitude in dimension. The classic approaches that we have encountered in path dealing with high-dimensional non-linearity, such as decomposition of non-linear function of by using polynomials or wavelets, are destined to be tackled down by the curse.

To overcome the curse, a promising and uprising technique is somewhat naturally the deep neural network. Deep neural networks with several layers show marvelous performance (yet not really exhaustively and theoretically comprehensible) when facing representation problems of non-linear functions. The well-studied universal approximation theorem, combined with the fact that simple functions would be able to form a composition of more complex ones, will endow us with satisfying approximation. In [1], J.Han, A.Jentzen and W.E 2017 has introduced the deep BSDE taking benefit from the not-so-modern idea of deep neural network to enlarge the horizon (especially w.r.t the dimensionality) of application with the existing numerical scheme for SDEs (forward and backward). A deep learning framework that has been intuitively embedded into the numerical scheme has also been designed in [1] in parallel.

On the contrary, it is not obvious to verify the efficiency and correctness of the method proposed. A limited number of cases where high-dimensional algorithms in classic sense are available are investigated in order to justify the newly proposed method.

- Inviscid Hamilton-Jacobi-Bellman equations for which Darbon & Osher have proposed an algorithm for high-dimensional case. In our case, a classic linear-quadratic-Gaussian control problem is studied.
- Non-linear parabolic PDEs for which a general algorithm has been recently developed based on Picard iteration. The method has successfully avoid the curse of dimensionality yet still suffer from the instability in finite time horizon.
- Non-linear Black-Scholes PDE with default risk modelled by a Poisson process.
- Other reaction-diffusion PDEs of mathematical physics such as Allen-Cahn equation describing the phase transition in multi-component alloy as well as the order-disorder transition in it.

We set a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_t, \mathbb{P})$ and a d-dimensional standard Brownian motion $\{W_t\}_{t\in[0,T]}$ that generates the filtration. For the temporal and spatial variables $(t,x) \in [0,T) \times \mathcal{O} \subset \mathbb{R}^+ \times \mathbb{R}^d$ and some terminal condition g, a fairly general form of semi-linear parabolic PDEs are considered:

$$\begin{cases} \frac{\partial u}{\partial t}(t,x) + \frac{1}{2} tr\Big(\sigma\sigma^T(t,x)(D_x^2 u)(t,x)\Big) + \nabla u(t,x) \cdot \mu(t,x) + f\big(t,x,u(t,x),(\sigma^T\nabla)u(t,x)\big) = 0 \\ u(T,x) = g(x) \end{cases} \tag{1}$$

where μ is vector-valued function and σ a matrix-valued one. f denotes a specified non-linear function. As in the standard financial mathematical setup, we are interested in the initial value, with the prices of a pack of assets being $x = \xi \in \mathbb{R}^d$.

We further denote $\{X_t\}_{t\in[0,T]}$ a d-dimensional Itô satisfying:

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

By the fact that Markovian BSDEs will give a non-linear Feymann-Kac representation of certain non-linear parabolic PDEs, we can rewrite the solution of the previous PDE by taking the following system of coupled FBSDEs into account.

$$\begin{cases} \mathbb{R}^d \ni X_t = \xi + \int_0^t \mu(s,X_s)ds + \int_0^t \sigma(s,X_s)dW_s \\ \mathbb{R} \ni Y_t = g(X_T) + \int_t^T f(s,X_s,Y_s,Z_s)ds - \int_t^T (Z_s)^T dW_s \end{cases} \tag{2}$$

with the solution process $\{(X_t, Y_t, Z_t)\}_t$ being adapted. Note also that he evolution of $\{X_t\}_t$ does not depend on $\{Y_t\}_t$, and here demonstrates the decoupledness of our model. Thanks to the well-developed theory of the existence and uniqueness of solutions (up to indistinguishability, cf.,e.g.[3,4]) , it suffices only to endow $\mu, \sigma \text{and} f$ with appropriate regularity to rewrite the equation:

$$Y_t = u(t, X_t) \quad \text{and} \quad Z_t = (\sigma^T \nabla u)(t, X_t)$$

By plugging this relation into the second equation of (2) and by reverting the time variable $t \leftarrow T - t$ we obtain:

$$u(t, X_t) = g(X_0) + \int_0^t f(s, X_s, u(s, X_s), (\sigma^T \nabla u)(s, X_s))ds - \int_0^t (\nabla u^T \sigma)(s, X_s)dW_s \tag{3}$$

and by the famous result in [4,5], we are aware that the solution of (1) will also satisfy this SDE.

The theoretical background being completed, we conduct a standard discretization to achieve a numerical algorithm to compute our goal quantity $u(0, X_0)$. It is crucial to take a different point of view (a deep learning one) of the initial values to make use of the deep neural network. We regard these quantities as parameters to be optimized: $u(0, X_0) \approx \theta_{u_0}$ and $\nabla u(0, X_0) \approx \theta_{\nabla u_0}$ and we we compute the terminal value, which is previously determined, according to the specific dynamic of (3). We set a N-step partition of $[0, T]$ which is $0 = t_0 < t_1 < ... < t_N = T$ and the following scheme:

$$X_{t_{n+1}} - X_{t_n} \approx \mu(t_n, X_{t_n})\Delta t_n + \sigma(t_n, X_{t_n})\Delta W_n, \text{ according to the dynamic of } X_t \tag{4}$$

and

$$\begin{aligned} u(t_{n+1}, X_{t_{n+1}}) - u(t_n, X_{t_n}) \approx & -f(t_n, X_{t_n}, u(t_n, X_{t_n}), (\sigma^T \nabla u)(t_n, X_{t_n}))\Delta t_n \\ & + [\nabla u(t_n, X_{t_n})]^T \sigma(t_n, X_n)\Delta W_n, \text{ according to (3)} \end{aligned} \tag{5}$$

where

$$\Delta t_n = t_{n+1} - t_n, \quad \Delta W_n = W_{t_{n+1}} - W_{t_n}$$

The path discretized paths $\{X_{t_n}\}_{t_n \in \{t_1, ..., t_N\}}$ being easy to approximate, it remains the mapping $x \mapsto \sigma^T \nabla u$ to compute to accomplish the scheme for $u(t_n, X_n)$. Without calculating it using a discretization numerical algorithm and do the matrix multiplication, we compute this term with an appropriate multi-layer feed forward neural network (FFNN), as in the spirit of our project. Suppose parameters $\theta_n, n = 1, ..., N$ are supposed to be at our knowledge when approximating our desired quantity $(\sigma^T \nabla u)(t, x)$ at the exact moment $t = t_n$, that is:

$$(\sigma \nabla u)^T(t_n, X_{t_n}) \approx (\sigma^T \nabla u)(t_n, X_{t_n} | \theta_n) \tag{6}$$

At this point, let us sort out our computing process. There are three main flows of sub-structure (let's suppose for the moment that $\forall x \in \mathbb{R}^d, \forall t \in [0, T], \sigma(t, x) = \text{Id}_{\mathbb{R}^d}$ just for illustration use) at moment $t = t_n$:

1. $X_{t_n} \to h_n^1 \to h_n^2 \to ... \to h_n^H \to \nabla u(t_n, X_{t_n})$. Here we deploy the multi-layer FFNN in order to approximate the spatial gradient, i.e. we will optimize the parameters θ_n. To this end, we apply a stochastic gradient descent-type algorithm developed and known as the Adam optimizer in [6]

2. $\left\{ u(t_n, X_n), \nabla u(t_n, X_n), \Delta W_n \right\} \to u(t_{n+1}, X_{t_{n+1}})$ is as straightforward as every standard scheme and will lead us to the final output $u(t_N = T, X_{t_N})$

3. $\left\{X_{t_n}, \Delta W_n\right\} \to X_{t_{n+1}}$ is the simple simulation of the process $\{X_t\}_t$

An illustrative figure which we took from [1] pictured the algorithm and show that once the previous step 3 is done for all t_k, the optimization process using the FFNN from $X_{t_k}, \forall k$ to $\nabla u(t_k, X_{t_k})$ can be practiced simultaneously.

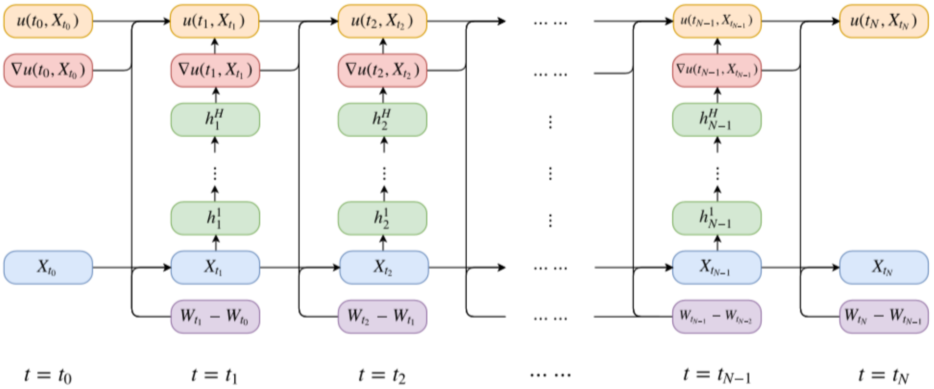


Figure 4: Illustration of the network architecture for solving semilinear parabolic PDEs with H hidden layers for each sub-network and N time intervals. The whole network has $(H+1)(N-1)$ layers in total that involve free parameters to be optimized simultaneously. Each column for $t = t_1, t_2, \ldots, t_{N-1}$ corresponds to a sub-network at time t. $h_n^1, \ldots, h_n^H$ are the intermediate neurons in the sub-network at time $t = t_n$ for $n = 1, 2, \ldots, N-1$.

Note that even though our figure shows the algorithm under an ordinary forward time flow for a specific initial position $x = \xi$, it is not complicated to generalize the approach to the case where we want to estimate the value of u at time $t = 0$ and starting in a region $D \subset \mathbb{R}^d$. Set $X_0 = \xi$ a non-degenerate D-valued r.v. and we add two more neural networks parameterized by $\{\theta_{u_0}, \theta_{\nabla u_0}\}$ to approximate the following mappings: $\begin{cases} x \mapsto u(0,x) \in \mathbb{R} \\ x \mapsto \nabla u(0,x) \in \mathbb{R}^d \end{cases}$.

Specifically, the complete structure will take the paths $\{X_{t_n}\}_{0\leq n\leq N}$ nad $\{W_{t_n}\}_{0\leq n\leq N}$ as the input and gives $\hat{u}(\{X_{t_n}\}_{0\leq n\leq N}, \{W_{t_n}\}_{0\leq n\leq N}) \approx u(t_N, X_{t_N})$. A loss function is given for the use of measuring the difference for the set of parameters $\theta = \left\{\theta_{u_0}, \theta_{\nabla u_0}, \theta_1, ..., \theta_{N-1}\right\}$:

$$l(\theta) = \mathbb{E}\left[\left|g(X_{t_N}) - \hat{u}\big(\{X_{t_n}\}_{0\leq n\leq N}, \{W_{t_n}\}_{0\leq n\leq N}\big)\right|^2\right] \tag{7}$$

2.2 Numerical Efficiency of deep BSDE method

[1] have demonstrated the outstanding efficiency and accuracy of the algorithm with several classic PDEs in relatively high dimensional setups such as non-linear Black-Scholes PDE for a portfolio of 100 underlying assets, Hamilton-Jacobi-Bellman equation of a stochastic continuous linear-quadratic problem in dimension 100 and the numerical resolution of a simplified Allen-Cahn equation with specific choice of the derivative of the double-well potential. The following chart shows the highly coherent and robust results achieved in [1] under proper theoretical setups.

Equation	Expression	Learning rate	Deep BSDE Error Rate	Computing Time (s)
Non-linear B.-S. equation with default risk	$\partial_t u+\bar{\mu}x\cdot\nabla u+\frac{\bar{\sigma}^2}{2}\sum\limits_{i=1}^{d}\lvert x_i\rvert^2\partial_{x_i}^2 u-(1-\delta)Q(u)u-Ru=0$	0.008	0.46%	1607
HJB equation in linear-quadratic-Gaussian control problem	$\begin{cases}\partial_t u+\Delta u-\lambda\lVert\nabla u\rVert^2=0\\ u(t,x)=-\frac{1}{\lambda}ln\Big(\mathbb{E}\big[-\lambda g(x+\sqrt{2}W_{T-t})\big]\Big)\\ \quad\text{is the explicit solution}\\ g(x)\text{ being the terminal condition}\end{cases}$	0.010	0.17%	330
Allen-Cahn equation	$\begin{cases}\partial_t u-\Delta u+u-u^3=0\\ u(0,x)=g(x)=\frac{1}{2+0.4\lVert x\rVert^2}\end{cases}$	0.0005	0.30%	647
An oscillating HJB-type PDE	$\begin{cases}\partial_t u+\frac{1}{2}\Delta u+min\Big\{1,\big(u-u^*\big)^2\Big\}=0\\ u^*(t,x)=Cte+sin(\lambda\sum\limits_{i=1}^{d}x_i)exp(\frac{\lambda^2 d(t-T)}{2})\end{cases}$	0.01/0.001	0.53% to 2.29%†	N.A

†The mean of relative error decreases from 2.29% with 29 layers to 0.53% when we have 145 layers. The comparison was designed to find out the most efficient number of hidden layers

We shall note that the equipment on which all the results are collected is not a powerful machine but an ordinary Macbook Pro with 2.9GHz Intel Core i5 processor. This shows that with appropriate optimization of distribution of memory and video cards with stronger computing power, one can control the running of the algorithm within a very reasonable time scale.

We have chosen the two cases that concerns us, students in mathematical finance the most: the non-linear B.-S.equation and the HJB equation to test the validity on our own. We have use the PyTorch library to proceed our algorithm. Detailed comments can be found on our python script.

2.3 Formulation of the problem

In the general system of decoupled BSDEs (2), the numerical simulation of the process $\{Z_t\}_t$ has been replaced by the usage of FFNNs. To comprehend the decoupledness of the system, we make the full use of the Markovian property of $\{X_t\}_t$ in this case (which will not hold for the coupled case: the combined process $\{X_t, Y_t\}_t$ will be Markovian instead). We will consider the following four-time-step scheme :

$$\begin{cases} X_0^\pi \;=\; \xi\;,\;\; Y_0^\pi \;=\; \mu_0^\pi(\xi), \\ X_{t_{i+1}}^\pi \;=\; X_{t_i}^\pi \;+\; b(t_i, X_{t_i}^\pi)\; h \;+\; \sigma(t_i, X_{t_i}^\pi)\; \Delta W_i, \\ Z_{t_i}^\pi \;=\; \phi_i^\pi(X_{t_i}^\pi), \\ Y_{t_{i+1}}^\pi \;=\; Y_{t_i}^\pi \;-\; f(t_i, X_{t_i}^\pi, Y_{t_i}^\pi, Z_{t_i}^\pi)\; h \;+\; (Z_{t_i}^\pi)^T\; \Delta W_i. \end{cases} \tag{8}$$

where we may consider that $\{X_t\}_t \in \mathbb{R}^m$ and $\{Z_t\}_t \in \mathbb{R}^d$ are of different dimensions. Moreover, the process indexed by π are those who are simulated numerically. Equipped with all the notions above, we will be considering the following stochastic optimization problem with the expected squared difference of the final values:

$$\inf_{\mu_0^\pi \in \mathcal{N}_0', \phi_i^\pi \in \mathcal{N}_i} F(\mu_0^\pi, \phi_0^\pi, ..., \phi_{N-1}^\pi) := \mathbb{E}\Big[\big|g(X_T^\pi) - Y_T^\pi\big|^2\Big] \tag{9}$$

where $\mathcal{N}_0'$ and $\mathcal{N}_i$ are the spaces of parametric functions generated by the FFNN (whose element are the parameters θ_{u0}, $\theta_{\nabla u_i}$ and θ_i, $0 \leq i \leq N-1$ mentioned before). We inherit from [2] the assumptions on these two spaces that $\mathcal{N}_0'$ is the subset of measurable functions from $\mathbb{R}^m$ to $\mathbb{R}$ and $\mathcal{N}_i$ are subsets of measurable functions from $\mathbb{R}^m$ to $\mathbb{R}^d$ Where (9) comes from can be captured by the following variational problem:

$$\inf_{Y_0, \{Z_t\}_{0 \leq t \leq T}} \mathbb{E}\big[|g(X_T) - Y_T|^2\big] \text{ with the forward dynamical system} \tag{10}$$

$$X_t = \xi + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s, \tag{11}$$

$$Y_t = Y_0 - \int_0^t f(s, X_s, Y_s, Y_s)ds + \int_0^t (Z_s)^T dW_s \tag{12}$$

Then we observe that the solution of the FBSDEs (2) is a minimizer of this problem because the loss function will vanish by the backward construction. By solving the optimization problem (9), we obtain in the minimizer $\mu_0^\pi(\xi)$ as an optimal approximation of $u(0, \xi)$.

2.4 Non-linear Feymann-Kac Formulae, Assumptions and Useful Estimations

In order to prove that the previously mentioned deep BSDE solver has an a posteriori error estimation that vanishes if the universal approximation capability of our neural networks is provided. As we have proposed, the deep BSDE was originally crafted for decoupled FBSDEs already. Before we enter statement of our main theorems and the proofs of them, it is easily notable that the bridge between the parabolic PDEs and its reformulation of a system of FBSDEs, which is the non-linear Feymann-Kac formulae, should be needed and declared. These formulae construct the connection between the FBSDEs and the semi-linear parabolic PDEs. With this knowledge, we regard the term $\mathbb{E}\big[|Y_0 - Y_0^\pi|^2\big]$ as $\mathbb{E}\big[|u(0, \xi = X_0) - \mu_0^\pi(\xi)|^2\big]$

Theorem 1. Assume

1. *$m = d$ and b, σ, f are smooth functions with bounded first order derivatives w.r.t all the spatial variables*

2. *(Uniform Ellipticity) $\exists$ a constant μ such that $\forall t, x, z$*

$$0 \leq \sigma\sigma^T(t, x) \leq \mu\mathbf{Id}$$
$$|b(t, x)| + |f(t, x, 0, z)| \leq \mu$$

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

Then the following quasi-linear PDEs has a unique strong solution $u(t, x)$ which is bounded with bounded temporal derivative, gradient and Hessian matrix (in the sense of matrix norm $\|M\| = \sqrt{tr(M^T M)}$)

$$\begin{cases} \partial_t u + \frac{1}{2} tr(\sigma\sigma^T(t, x) D_x^2 u) + b^T(t, x)\Delta_x u + f(t, x, u, \sigma^T(t, x)\Delta_x u = 0 \\ u(T, x) = g(X) \end{cases} \tag{13}$$

The associated FBSDEs (2) has a unique solution process $\{(X_t, Y_t, Z_t)\}_t$ with $Y_t = u(t, X_t)$ and $Z_t = \big(\sigma^T \nabla_x u\big)(t, X_t)$ and $\{X_t\}_t$ is the solution of the SDE:

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

Remark: We first notice that if the drift function b also depends on z, one can put the associated term $b^T \nabla_x u$ into the non-linear term f and apply the theorem itself to achieve a new system in which the drift function is independent of z. The result was obtained via a PDE-based argument that requires the assumptions (1) and (2). An somewhat probabilistic parallel analog of this theorem is the following.
Remark: In this case, we only used the Lipschitz assumption with supplementary conditions such as the weak coupling or monotonicity which will be mentioned in the next subsection. A fairly counter-intuitive note is given in [2], stating that the Lipshitz property could not guarantee the existence of a solution to the FBSDEs if they are coupled even if b, σ and f are all linear.

Now we will introduce some of the Lipschitz assumptions that allow us to apply an probabilistic alternative of **Theorem 1** and more importantly, a discret-time system of equations that has nice convergence property only depends on the initial position ξ and length of time discretization interval h. In the original paper, the assumptions are involving the variable y for that the FBSDEs are coupled. Supposing momentarily that the FBSDEs are coupled, we will still follow the original setup and will simply remove the dependence on y by setting the coefficients to 0, which will not cause significant influence on our future proofs. Notice also, as mentioned in [2], there are two ways of decoupling the system: (1): The forward process $\{X_t\}_t$ does not depend on the backward one $\{Y_t\}_t$, which is surely the case of our interest. (2): The backward equation is independent from the forward one. In this case, $Z_t = 0, \forall t$ and thus the BSDE for Y_t degenerates to an ODE.

Assumption 1. *With $x_1 \geq x_2$ (w.l.o.g.), $\Delta x = x_1 - x_2, \Delta y = y_1 - y_2$ and $\Delta z = z_1 - z_2$:*

1. There exists (possibly negative) constants k_b, k_f such that:

$$\begin{aligned}\left[b(t,x_1,y) - b(t,x_2,y)\right]^T \Delta x &\leq k_b |\Delta x|^2, \\ [f(t,x,y_1,z) - f(t,x,y_2,z)]\Delta y &\leq k_f |\Delta y|^2\end{aligned}$$

2. b, σ, f, g are uniformly Lipschitz w.r.t (x,y,z), i.e. $\exists K$ (being the upper bound of the set $\mathcal{L}$ containing all the constants mentioned), $b_y, \sigma_x, \sigma_y, f_x, f_z > 0$ such that:

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

3. $b(t,0,0), f(t,0,0,0)$ and $\sigma(t,0,0)$ are bounded. What is more, we assume that the increments of these functions are up to quadratic speed:

$$\begin{aligned}|b(t,x,y)|^2 &\leq b_0 + K|x|^2 + b_y|y|^2 \\ \|\sigma(t,x,y)\|^2 &\leq \sigma_0 + \sigma_x|x|^2 + \sigma_y|y|^2 \\ |f(t,x,y,z)|^2 &\leq f_0 + f_x|x|^2 + K|y|^2 + f_z|z|^2 \\ |g(x)|^2 &\leq g_0 + g_x|x|^2\end{aligned}$$

Assumptions 2. b, σ, f, g are uniformly $\frac{1}{2}$-Hölder continuous w.r.t t. The constant K is again supposed to be the square of the Hölder constants.

Assumptions 3. We supposed at least one of the following five cases is true:

1. Small time duration: T is small.
2. Weak coupling of Y into the process $\{X_t\}_t$, i.e. _y and σ_y are small. (In our case, both will equal zero.)
3. Weak coupling of X into $\{Y_t\}_t$., i.e. b_x and σ_x are small.
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.

This is the so-called weak coupling and monotonicity conditions that are used quite often in the literature such as [4,7].
A more compact equivalence of the **Assumption 3** can be deduced with mechanical yet elementary computations. Thanks to the result and summary in [7], we define the following constants to simplify the notation and assumptions.

$$
\begin{aligned}
&L_0 = [b_y + \sigma_y][g_x + f_x T] T e^{[b_y+\sigma_y][g_x+f_xT]T+[2k_b+2k_f+2+\sigma_x+f_z]T} \\
&L_1 = [g_x + f_x T][e^{[b_y+\sigma_y][g_x+f_xT]T+[2k_b+2k_f+2+\sigma_x+f_x]T+1} \vee 1] \\
&\Gamma_0(x) = \frac{e^x - 1}{x}, \text{ for} x > 0 \\
&\Gamma_1(x, y) = \sup_{0<\theta<1} \theta e^{\theta x} \Gamma_0(y) \\
&c = \inf_{\lambda_1>0} \bigg\{ \big[e^{[2k_b+1+\sigma_x+[b_y+\sigma_y]L_1]T} \vee 1\big] \big(1+\lambda_1^{-1}\big) \big[b_y+\sigma_y\big] T \\
&\times \bigg[g_x \Gamma_1 \bigg(\big[2k_f + 1 + f_z\big] T, \big[2k_b + 1 + \sigma_x + (1+\lambda_1)[b_y+\sigma_y] L_1\big] T \bigg) \\
&+ f_x T \Gamma_0 \bigg([2k_f + 1 + f_z] T \bigg) \times \Gamma_0 \bigg(2k_b + 1 + \sigma_x + (1+\lambda_1)[b_y+\sigma_y] L_1 T \bigg) \bigg] \bigg\}
\end{aligned}
$$

The expressions of these constants are surely complex but the minimizer of c is nothing else but the solution of the first order condition (computational details could be checked in [7]). These setups give a quantitative qualification of the **Assumption 3**:

$$
L_0 < e^{-1}, \quad \text{and} \quad c < 1 \tag{14}
$$

We will further refer the 3^{rd} assumption as the the condition (14). The pack of assumptions are crucial for us to make use out of the theoretical results from [7]. Unlike the the **Theorem 1**, these results are stronger in the sense that a coupled system of FBSDEs as well as a wider range of parabolic PDEs. Furthermore, we take benefit from a theorem giving the convergence order of an implicit scheme for coupled FBSDEs. Compared to the original paper [7] emphasizing on the scheme itself, [2] assumed more strict conditions in order to investigate the convergence of deep BSDE solver and managed to integrate the collection of new conditions, as well as parameters and coefficients, into the desired explicit error estimation with respect to the neural network. The convergence will naturally stand if such inequality is obtained. Our project aims at a degenerate case and is obviously comprehended in this framework. We will also be able to observe the righteousness of the previous assumptions as mentioned in [2]

At the end of the section, we re-announce the results from [7]: a stronger version of non-linear Feymann-Kac formula and a useful error estimate. Both theorems are applicable for the coupled FBSDEs and are naturally true for our originally decoupled case.

Theorem 2. *(A stronger Feymann-Kac formula) Under* ***Assumptions 1,2 and 3****, there exists a mapping* $(t,x) \mapsto u(t,x)$ *satisfying:*

1. $$|u(t,x_1) - u(t,x_2)|^2 \leq L_1|x_1 - x_2|^2$$
2. $$|u(s,x) - u(t,x)|^2 \leq C(1+|x|^2)|s-t| \text{ with } C = C_{\mathcal{L},T}$$
3. *u is a viscocity solution of the PDEs (13)*
4. *The FBSDEs (2) has a unique solution process* $\{(X_t, Y_t, Z_t)\}_t$ *and* $Y_t = u(t, X_t)$*, which satisfies the following decoupled FBSDEs :*

$$\begin{cases} X_t = \xi + \int_0^t b(s, X_s, u(s, X_s))ds + \int_0^t \sigma(s, X_s, u(s, X_s))dW_s \\ Y_t = g(X_T) + \int_t^T f(s, X_s, Y_s, Z_s)ds - \int_t^T (Z_s)^T dW_s \end{cases} \tag{15}$$

Further more, the path regularity of the solution of the FBSDEs satisfies the following inequality with some C:

$$\sup_{t\in[0,T]} (\mathbb{E}[|X_t - \widetilde{X}_t|^2] + \mathbb{E}[|Y_t - \widetilde{Y}_t|^2]) + \int_0^T \mathbb{E}[|Z_t - \widetilde{Z}_t|^2]dt \leq C(1 + \mathbb{E}[|\xi|^2])h \tag{16}$$

where $\widetilde{X}_t = X_{t_i}, \widetilde{Y}_t = Y_{t_i}$ *and* $\widetilde{Z}_t = h^{-1}\mathbb{E}[\int_{t_i}^{t_{i+1}} Z_t dt|\mathcal{F}_{t_i}]$ *for* $t \in [t_i, t_{i+1})$*. Besides,* $\widetilde{Z}_t$ *can be replaced by* $Z_{t_i}i$ *if* Z_t *is* càdlà*. (To guarantee the existence of up to a* càdlà *modification for* $\{Z_t\}_t$*,* $m = d$ *or the existence of lower bound for* $\sigma\sigma^T$ *will suffice.)*

The **Theorem 2** gives the error estimate with respect to the values of the processes on fixed temporal discretization points. The following theorem will demonstrate a very similar result but with respect to the solution of a truly discrete system.

Theorem 3. *Under* **Assumptions 1,2 and 3***, for sufficiently small h, the following discrete-time equation (for* $0 \leq i \leq N_1$*)*

$$\begin{cases} \overline{X}_0^\pi = \xi \\ \overline{X}_{t_{i+1}}^\pi = \overline{X}_{t_i}^\pi + b(t_i, \overline{X}_{t_i}^\pi, \overline{Y}_{t_i}^\pi)h + (t_i, \overline{X}_{t_i}^\pi, \overline{Y}_{t_i}^\pi)\Delta W_i \\ \overline{Y}_T^\pi = g(\overline{X}_T^\pi) \\ \overline{Z}_{t_i}^\pi = \frac{1}{h}\mathbb{E}[\overline{Y}_{t_{i+1}}^\pi \Delta W_i|\mathcal{F}_{t_i}] \\ \overline{Y}_{t_i}^\pi = \mathbb{E}[\overline{Y}_{t_{i+1}}^\pi + f(t_i, \overline{X}_{t_i}^\pi, \overline{Y}_{t_i}^\pi, \overline{Z}_{t_i}^\pi)h|\mathcal{F}_{t_i}] \end{cases} \tag{17}$$

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

$$\sup_{t\in[0,T]} (\mathbb{E}[|X_t - \overline{X}_t|^2] + \mathbb{E}[|Y_t - \overline{Y}_t|^2]) + \int_0^T \mathbb{E}[|Z_t - \overline{Z}_t|^2]dt \leq C(1 + \mathbb{E}[|\xi|^2])h \tag{18}$$

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

Remark. As mentioned in [2,14], the original result was obtained by applying the technique to an explicit scheme by replacing the term $\overline{Y}_t = X_{t_i}$ in the non-linearity f of the last equation. The same approach can be applied to the implicit scheme to demonstrate this theorem.

2.5 Statement of the Main Theorems

Hereby we state the theorems under reasonable assumptions (basically strong monotonicity) that show the validity of our intuitively correct deep BSDE method when the FBSDEs are decoupled.

Theorem 4: *Under reasonable assumptions, there exists a constant C independent of h, d and m such that for sufficiently small h:*

$$\sup_{t\in[0,T]} (\mathbb{E}[|X_t - \hat{X}_t|^2] + \mathbb{E}[|Y_t - \hat{Y}_t|^2]) + \int_0^T \mathbb{E}[|Z_t - \hat{Z}_t|^2]dt \leq C(h + \mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2]) \tag{19}$$

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

Theorem 5: *Under reasonable assumptions, there exists a constant C, independent of h, d and m such that for sufficiently small h:*

$$\inf_{\mu_0^\pi \in \mathcal{N}_0', \phi_i^\pi \in \mathcal{N}_i} \mathbb{E}\left[\left|g(X_T^\pi) - Y_T^\pi\right|^2\right] \tag{20}$$

$$\leq C\left\{h + \inf_{\mu_0^\pi \in \mathcal{N}_0', \phi_i^\pi \in \mathcal{N}_i} \left[\mathbb{E}\big[|Y_0 - \mu_0^\pi(\xi)|^2\big] + \sum_{i=0}^{N-1} \mathbb{E}\big[\mathbb{E}[\widetilde{Z}_{t_i}|X_{t_i}^\pi, Y_{t_i}^\pi] - \phi_i^\pi(X_{t_i}^\pi, Y_{t_i}^\pi)|^2\big]\right]\right\} \tag{21}$$

where the term $\mathbb{E}[\widetilde{Z}_{t_i}|X_{t_i}^\pi, Y_{t_i}^\pi]$ can be replaced with $\mathbb{E}[\widetilde{Z}_{t_i}|X_{t_i}^\pi]$ when the FBSDEs are decoupled.

Theorem 4 states that the sum of expected squared simulation error will be bounded by the value of the objective function (9), concerning both time discretization error and terminal distance. **Theorem 5**, on the other hand, shows that the optimal value of the objective function can be bounded and pushed to enoughly small value if the approximation capability of the parametric function space is high, i.e. $\mathcal{N}_0'$ and $\mathcal{N}_i$ are sufficiently large. In high-dimensional cases, deep neural networks perform with persuasive results when dealing with such sort of problems. A significant observation has been made in [2] and here we quote: " **Theorem 2** implies that if the involved conditional expectations can be approximated by the deep neural networks whose numbers of parameters to be optimized are 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 suffering from the curse of dimensionality."

3 Convergence of the Deep BSDE Method for decoupled FBSDEs: Proofs

In the following subsections, we will be focusing on the case where the FBSDEs are decoupled and discuss the difference with the generalized coupled case. From here forth, let's suppose that $b_y = \sigma_y = 0$.

3.1 Proof of Theorem 4: A Posteriori Estimation of the Simulation Error

To prove the theorem, we follow principally the same techniques and ideas that are practiced in [2]. Compare the already existing result **Theorem 3** and our goal **Theorem 4** and recall the dynamics of the four-time-step system of discrete equations (8):

$$\left\{ \begin{array}{lll} X_{t_{i+1}}^\pi &=& X_{t_i}^\pi \,+\, b(t_i, X_{t_i}^\pi)\, h \,+\, \sigma(t_i, X_{t_i}^\pi)\, \Delta W_i, \\ Y_{t_{i+1}}^\pi &=& Y_{t_i}^\pi \,-\, f(t_i, X_{t_i}^\pi, Y_{t_i}^\pi, Z_{t_i}^\pi)\, h \,+\, (Z_{t_i}^\pi)^T\, \Delta W_i. \end{array} \right. \tag{22}$$

Take the expectation w.r.t $\mathcal{F}_{t_i}$ on the second equation:

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

Note that this line is also exactly the same as the last equation of (17) in **T**hen we multiply by $(\Delta W_i)^T$ to the same line and take the same conditional expectation, by the fact that $\Delta W_t \cdot \Delta W_t \approx \Delta t = h$, one should get:

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

Notice this is again the mentioned in (17). These observations together will drive us to the following discrete version of (17) which takes the specific form that is very similar to (8).

$$\begin{cases} X_0^{\pi} \ = \ \xi \\ X_{t_{i+1}}^{\pi} \ = \ X_{t_i}^{\pi} \ + \ b(t_i, X_{t_i}^{\pi}) \, h \ + \ \sigma(t_i, X_{t_i}^{\pi}) \, \Delta W_i, \\ Z_{t_i}^{\pi} \ = \frac{1}{h}[Y_{t_{i+1}}^{\pi} \Delta W_i | \mathcal{F}_{t_i}] \\ Y_{t_{i+1}}^{\pi} \ = \mathbb{E}[Y_{t_{i+1}}^{\pi} + f(t_i, X_{t_i}^{\pi}, Y_{t_i}^{\pi}, Z_{t_i}^{\pi})|\mathcal{F}_{t_i}]. \end{cases} \tag{23}$$

We should recall the fact that in systems such as (8) and (23), the terminal conditions of Y_T are not specified. This allow an infinite number of solutions and we shall need a measure of difference between two solutions.

Lemma 1: *For j = 1,2, suppose* $(\{X_{t_i}^{\pi,j}\}_{0 \le i \le N}, \{Y_{t_i}^{\pi,j}\}_{0 \le i \le N}, \{Z_{t_i}^{\pi,j}\}_{0 \le i \le N-1})$ *are two distinc solutions of (23), with* $X_{t_i}^{\pi,j}, Y_{t_i}^{\pi,j} \in L^2(\Omega, \mathcal{F}_t, \mathbb{P}), \forall i$. *For any* $\lambda_1 > 0, \lambda_2 \ge f_z$ *and sufficiently small h, denote:*

$$\begin{aligned} A_1 &:= 2k_b + \lambda_1 + \sigma_x + Kh \\ A_2 &:= (\lambda_1^{-1} + h) b_y + \sigma_y \\ A_3 &:= -\frac{ln\big(1 - (2k_f + \lambda_2)h\big)}{h} \\ A_4 &:= \frac{f_x}{\big(1 - (2k_f + \lambda_2)h\big)\lambda_2} \end{aligned}$$

$\cdot Notethatthequantity A_2$ *will not be used in our error estimation since we have already assumed that* b_y *and* σ_y *equal zero. We keep this quantity for further discussion use.*
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 every* n:

$$\begin{aligned} &\mathbb{E}\big[|\delta X_n|^2\big] \le A_2 \sum_{i=0}^{n-1} e^{A_1(n-i-1)h} \mathbb{E}[|\delta Y_i|^2] \quad \text{if coupled} \\ &\mathbb{E}\big[|\delta X_n|^2\big] \le 0 \quad \text{if decoupled} \\ &\mathbb{E}\big[|\delta Y_n|^2\big] \le e^{A_3(N-n)h} \mathbb{E}\big[|\delta Y_N|^2\big] + A_4 \sum_{i=n}^{N-1} e^{A_3(i-n)h} \mathbb{E}\big[|\delta X_i|^2\big] h \end{aligned}$$

To show the **Lemma 1**, we need another simple but not trivial supplementary lemma.

Lemma 2. *Let* $0 \le s_1 < s_2$, *given a square-integrable r.v* $Q \in L^2(\Omega, \mathcal{F}_{s_2}, \mathbb{P})$, *by the martingale representation theorem, there exists a* $\mathcal{F}_t$*-adapted process* $\{H_s\}_{s_1 \le s \le s_2}$ *such that* $H \in \mathbb{M}^2$ *and* $Q = \mathbb{E}[Q|\mathcal{F}_{s_1}] + \int_{s_1}^{s_2} H_s dW_s$. *Then we have:*

$$\mathbb{E}[Q(W_{s_2} - W_{s_1})|\mathcal{F}_{s_1}] = \mathbb{E}[\int_{s_1}^{s_2} H_s ds | \mathcal{F}_{s_1}] \tag{24}$$

Proof: To see the result, we only have to use the auxiliary process $Q_s = (\mathbb{E}[Q|\mathcal{F}_{s_1}] + \int_{s_1}^{s} H_s dW_s)(W_s - W_{s_1})$. By writing its dynamic, thanks to the Itö lemma, on can obtain the result by:

$$\begin{aligned}
&\mathbb{E}[Q(W_{s_2} - W_{s_1})|\mathcal{F}_{s_1}] = \mathbb{E}[Q_{s_2}|\mathbb{F}_{s_1}] \\
&\quad = \mathbb{E}\bigg[Q_{s_1} + \int_{s_1}^{s_2} (W_s - W_{s_1}) H_s dW_s + \int_{s_1}^{s_2} (\mathbb{E}[Q|\mathcal{F}_{s_1}] + \int_{s_1}^{s} H_t dW_t) dW_s + \int_{s_1}^{s_2} H_s ds|\mathcal{F}_{s_1}\bigg] \\
&\quad = \mathbb{E}\bigg[0 + 0 + \int_{s_1}^{s_2} H_s ds|\mathcal{F}_{s_1}\bigg]
\end{aligned}$$

Q.E.D

*Proof of **Lemma 1**:* Let

$$\begin{aligned}
\delta Z_i &= Z_{t_i}^{\pi,1} - Z_{t_i}^{\pi,2} \\
\delta b_i &= b(t_i, X_t^{\pi,1}) - b(t_i, X_{t_2}^{\pi,2}) \\
\delta \sigma_i &= \sigma(t_i, X_t^{\pi,1}) - \sigma(t_i, X_{t_2}^{\pi,2}) \\
\delta f_i &= f(t_i, X_t^{\pi,1}, Y_t^{\pi,1}, Z_t^{\pi,1}) - f(t_i, X_{t_2}^{\pi,2}, Y_t^{\pi,2}, Z_t^{\pi,2})
\end{aligned}$$

which lead to

$$\delta X_{i+1} = \delta X_i + \delta b_i h + \delta \sigma_i \Delta W_i \tag{25}$$

$$\delta Z_i = \frac{1}{h} \mathbb{E}[\delta Y_{i+1} \Delta W_i | \mathcal{F}_{t_i}] \tag{26}$$

$$\delta Y_i = \mathbb{E}[\delta Y_{i+1} + \delta f_i h | \mathcal{F}_{t_i}] \tag{27}$$

By the martingale representation theorem, there exists a $\mathcal{F}_t$-adapted square-integrable process $\{\delta Z_t\}_{t_i \leq t \leq t_{i+1}}$. (Note that this process is different from the quantity δZ_i we defined above.) And we have

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

in particular,

$$\delta Y_{i+1} = \delta Y_i - \delta f_i h + \int_{t_i}^{t_{i+1}} (\delta Z_t)^T dW_t \tag{28}$$

We notice that $\delta X_i, \delta Y_i, \delta b_i$, and δf_i are all $\mathcal{F}_t$-measurable. Furthermore, by the nice conditions we set on Y and the construction of conditional expectation of Z, the stochastic integral $\int_{t_i}^{t_{i+1}} (\delta Z_t)^T dW_t$ is well-defined and $\mathbb{E}[\int_{t_i}^{t_{i+1}} (\delta Z_t)^T dW_t | \mathcal{F}_{t_i}] = 0$. We shall have

$$\mathbb{E}[|\delta X_{i+1}|^2] = \mathbb{E}[|\delta X_i + \delta b_i h|^2] + \mathbb{E}[(\Delta W_i)^T (\delta \sigma_i)^T \delta \sigma_i \Delta W_i] \tag{29}$$

$$= \mathbb{E}[|\delta X_i + \delta b_i h|^2] + h\mathbb{E}[\|\delta \sigma_i\|^2]\mathbb{E}[|\delta Y_{i+1}|^2] \tag{30}$$

$$= \mathbb{E}[|\delta Y_i - \delta f_i h|^2] + \int_{t_i}^{t_{i+1}} \mathbb{E}[|\delta Z_t|^2] dt \tag{31}$$

The last equation is a straightforward development using (28). From (29), by the **Assumptions 1,2**, we have for every $\lambda_1 > 0$:

$$\mathbb{E}[|\delta X_{i+1}|^2] \tag{32}$$

$$=\mathbb{E}[|\delta X_i|^2] + \mathbb{E}[|\delta b_i|^2 h^2] + h\mathbb{E}[\|\delta \sigma_i\|^2] \tag{33}$$

$$+ 2h\mathbb{E}\left[\left(b(t_i, X^{\pi,1}_{t_t}) - b(t_i, X^{\pi,2}_{t_t})\right)^T \delta X_i\right] \tag{34}$$

$$\leq \mathbb{E}[|\delta X_i|^2] + h^2 K\mathbb{E}[|\delta X_i|^2] + 2k_b hE[|\delta X_i|^2] + \lambda_1 h\mathbb{E}[|\delta X_i|^2] + \sigma_x h\mathbb{E}[|\delta X_i|^2] \tag{35}$$

$$= \left[1 + (2k_b + \lambda_1 + \sigma_x + Kh)h\right]\mathbb{E}[|\delta X_t|^2] \tag{36}$$

By the definition of δX_i, we know that $\delta X_0 = 0$ and an easy induction, we can obtain for $0 \leq n \leq N$:

$$\mathbb{E}[|\delta X_n|^2] \leq 0 \tag{37}$$

This result should not be surprising since if the FBSDEs are decoupled, plus the initial condition of the forward stochastic process X is determined to be ξ, we would get almost surely an identical distribution for every moment t, which also yields this consequence.

Similarly, by using the elementary root-mean square and geometric mean inequality, we could get for any $\lambda_2 > 0$

$$\begin{aligned}
&\mathbb{E}[|\delta Y_{i+1}|^2] \\
\geq &\mathbb{E}[|\delta Y_i|^2] + \int_{t_i}^{t_{i+1}} \mathbb{E}[|\delta Z_t|^2]dt \\
&- 2h\mathbb{E}\left[\left(f(t_i, X^{\pi,1}_{t_t}, Y^{\pi,1}_{t_t}, Z^{\pi,1}_{t_t}) - f(t_i, X^{\pi,1}_{t_t}, Y^{\pi,2}_{t_t}, Z^{\pi,1}_{t_t})\right)^T \delta Y_i\right] \\
&- 2h\mathbb{E}\left[\left(f(t_i, X^{\pi,1}_{t_t}, Y^{\pi,2}_{t_t}, Z^{\pi,1}_{t_t}) - f(t_i, X^{\pi,2}_{t_t}, Y^{\pi,2}_{t_t}, Z^{\pi,2}_{t_t})\right)^T \delta Y_i\right] \\
\geq &\mathbb{E}|[\delta Y_i|^2] + \int_{t_i}^{t_{i+1}} \mathbb{E}[|\delta Z_t|^2]dt - 2k_f h\mathbb{E}[|\delta Y_t|^2] \\
&- \left[\lambda_2\mathbb{E}\big[|\delta Y_i|^2\big] + \lambda_2^{-1}(f_x\mathbb{E}[|\delta X_i|^2] + f_z\mathbb{E}[|\delta Z_i|^2])\right]
\end{aligned}$$

we then apply the **Lemma 2** to lines (26) and (28) and get

$$\delta Z_i = \frac{1}{h}\mathbb{E}[\int_{t_i}^{t_{i+1}} |\delta Z_t|^2 dt|\mathcal{F}_{t_i}]$$

and by Cauchy-Schwarz Inequality,

$$\mathbb{E}[|\delta Z_i|^2] = \frac{1}{h}\mathbb{E}\left[\left|\mathbb{E}[\int_{t_i}^{t_{i+1}} |\delta Z_t|^2 dt|\mathcal{F}_{t_i}]\right|^2\right] \leq \int_{t_i}^{t_{i+1}} \mathbb{E}[|\delta Z_i|^2]dt \tag{38}$$

Finally, we plug it to the previous computation

$$\mathbb{E}[|\delta Y_{i+1}|^2] \geq [1 - (2k_f + \lambda_2)h]\mathbb{E}[|\delta Y_i|^2] + (1 - f_z\lambda_2^{-1})\mathbb{E}[|\delta Z_i|^2]h - f_x\lambda_2^{-1}\mathbb{E}[|\delta X_i|^2]h \tag{39}$$

This show that it suffice to have $\lambda_2 \geq f_z$ and sufficiently small h such that $h(2k_f + \lambda_2^{-1}) < 1$, we could achieve the one-step estimation:

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

Recall the definition of A_3 and A_4, we will obtian the desired result:

$$\mathbb{E}\big[|\delta Y_n|^2\big] \leq e^{A_3(N-n)h}\mathbb{E}\big[|\delta Y_N|^2\big] + A_4 \sum\limits_{i=n}^{N-1} e^{A_3(i-n)h}\mathbb{E}\big[|\delta X_i|^2\big]h \tag{40}$$

Equipped with the right lemmas, we prove the full version of **Theorem 4** with clear statements of the conditons. Here we state the complete version theorem where the FBSDEs can be coupled. It is particularly interesting to decouple the system through specific values of parameters and observe the influence.

Theorem 4': *Under* ***Assumptions 1, 2, 3*** *and there exists* $\lambda_1 > 0, \lambda_2 \geq f_z$ *such that* $\overline{A_0} < 1$, where

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

Then there exists a constant C *independent of* h, d *and* m*, dependent on* $\mathbb{E}[|\xi|^2], \mathcal{L}, T, \lambda_1$ *and* λ_2 *such that for sufficiently small* h*:*

$$\sup_{t\in[0,T]}(\mathbb{E}[|X_t - \hat{X}_t|^2] + \mathbb{E}[|Y_t - \hat{Y}_t|^2]) + \int_0^T \mathbb{E}[|Z_t - \hat{Z}_t|^2]dt \leq C(h + \mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2]) \tag{41}$$

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

Remark: The coercity of the objective function can be deduced directly from **Theorem 4'**. On the other hand, under the framework of our project, $\overline{A_2}$ will vanish automatically and so does $\overline{A_0}$. This yields directly that $\overline{A_0} < 1$ and unconditionally, we will achieve the desired spatial regularity if the FBSDEs are not coupled. For the moment, we will prove the theorem under the assumption that the system of FBSDEs are decoupled.

Proof: From this proof and forth, we will set C to be a generic constant that depends on $\mathbb{E}[|\xi|^2], \mathcal{L}andT$. We also inherit all the notations and we re-define the processes of different solutions:

$$\begin{cases}
X_{t_i}^{1,\pi} \to X_{t_i}^\pi, Y_{t_i}^{1,\pi} \to Y_{t_i}^\pi, Z_{t_i}^{1,\pi} \to Z_{t_i}^\pi \\
X_{t_i}^{2,\pi} \to \overline{X}_{t_i}^\pi, Y_{t_i}^{2,\pi} \to \overline{Y}_{t_i}^\pi, Z_{t_i}^{2,\pi} \to \overline{Z}_{t_i}^\pi
\end{cases}$$

Again, along the proof we will find that the decoupling FBSDEs strongly reduces the complexity of the demonstration. To begin with, we apply once again the RMS-GM inequality on $\mathbb{E}[|\delta Y_N|^2]$:

$$\mathbb{E}[|\delta Y_N|^2] = \mathbb{E}[|g(\overline{X}_T^\pi) - Y_T^\pi|^2] \leq (1+\lambda_3^{-1})\mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2] + \underbrace{g_x(1+\lambda_3)\mathbb{E}[|\delta X_N|^2]}_{=0\ by\ the\ decoupling}$$

Follow the approach from [2], let us set

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

By **Lemma 1**, we get

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

and

$$\begin{aligned} e^{A_3 nh}\mathbb{E}[|\delta Y_n|^2] &\le e^{A_3 T}\mathbb{E}[|\delta Y_N|^2] + 0 \\ &\le e^{A_3 T}(1+\lambda_3^{-1})\mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2] \end{aligned}$$

Therefore, by construction,

$$\begin{aligned} &P = 0 \le A_2 h e^{-A_1 h}\frac{e^{-(A_1+A_3)T} - 1}{e^{-(A_1+A_3)h} - 1}, \text{ which is trivial by the monotonicity and } \textbf{Assumption 3} \\ &S \le e^{A_3 T}(1+\lambda_3^{-1})\mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2] \end{aligned}$$

We denote

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

and

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

Then trivially, for any $\epsilon > 0, \exists\lambda_3 > 0$ and sufficiently small h satisfying

$$\overline{P} = 0 \le (1+\epsilon)\times 0 \tag{42}$$

$$\overline{S} \le (1+\epsilon)e^{\overline{A_3}T}(1+\lambda_3^{-1})\mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2] \tag{43}$$

By fixing $\epsilon = 1$ and choose suitable λ_3, one would get error estimations of $\mathbb{E}[|\delta X_n|^2]$ (in our case, the estimation is trivially 0) and $\mathbb{E}[|\delta Y_n|^2]$

$$\max_{0\le n\le N}\mathbb{E}[|\delta X_n|^2] \le 0 \tag{44}$$

$$\max_{0\le n\le N}\mathbb{E}[|\delta Y_n|^2] \le e^{-\overline{A_3}\vee 0}\overline{S} \le C\mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2] \tag{45}$$

Finally to estimate $\mathbb{E}[|\delta Z_n|^2]$, we will be using the estimate we have done in (38) and we can take any value for $\lambda_2 > f_z$. Here we take $\lambda_2 = 2f_z$ and we have

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

We sum this estimate from 0 to $N-1$:

$$\sum_{i=0}^{N-1}\mathbb{E}[|\delta Z_i|^2]h \le \frac{f_x T}{f_z}\max_{0\le n\le N}\mathbb{E}[|\delta X_n|^2] + [4(k_f + f_z)T\vee 0 + 2]\max_{0\le n\le N}\mathbb{E}[|\delta Y_n|^2] \tag{46}$$

$$\le C\mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2] \tag{47}$$

At the end, we combine the estimates (44),(45) and (47) and we apply the **Theorem 3**, we will prove the **Theorem 4'**.

3.2 Proof of Theorem 5: An Upper Bound for the Minimized Objective Function

To prove this theorem, we will be needing three other lemmas.

- **Lemma 3** will give an estimate on the final distance $\mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2]$ in term of the value of approximated variables $Y_0^\pi, Z_{t_i}^\pi$ and the real solutions.
- **Lemma 4** is a general result to serve as an estimate of difference between two backward processes driven by different forward processes.
- **Lemma 5** demonstrates that the discrete stochastic process defined in (8) can be linked to some deterministic functions.

Here, we will state the complete version of all the three lemmas. Nevertheless, when proving the **Theorem 5**, we will make use of them in the case where $b_y = \sigma_y = 0$.

Lemma 3: *Suppose* ***Assumptions 1,2,3****. Let us consider a system as (8) and a simulation process* $\widetilde{Z}_{t_i} = h^{-1}\mathbb{E}[\int_{t_i}^{t_{i+1}} Z_t dt|\mathcal{F}_{t_i}]$*. Given* $\lambda_4 > 0$*, there exists a constant* $C > 0$ *such that for sufficiently small* h*:*

$$\mathbb{E}[|g(X_T^\pi) - Y_T^\pi|^2] \leq (1+\lambda_4)H_{min}\sum_{i=0}^{N-1}\mathbb{E}[|\delta\widetilde{Z}_{t_i}|^2]h + C\Big[h + \mathbb{E}[|Y_0 - Y_0^\pi]|^2\Big]$$

where $\delta\widetilde{Z}_{t_i} = \widetilde{Z}_{t_i} - Z_{t_i}^\pi$ *and* $H_{min} = \underset{x\in\mathbb{R}_+}{min}\, H(x) = \underset{x\in\mathbb{R}_+}{min}\Big\{(1+\sqrt{g_x})^2 e^{(2K+2Kx^{-1}+x)T}(1+f_z x^{-1})\Big\}$
Sketch of the proof: The inequalities applied in the proof of this lemma are all standard techniques in the probability theory such as Grönwall inequality. The main progress are made by the repetitive use of RMS-RG inequality the and the path regularity given by **Theorem 2.**

We first construct continuous processes as in the standard Monte-Carlo approaches. For $t \in [t_i, t_{i+1})$:

$$\begin{aligned} X_t^\pi &= X_{t_i}^\pi + b(t_i, X_{t_i}^\pi)(t-t_i) + \sigma(t_i, X_{t_i}^\pi)(W_t - W_{t_i}) \\ Y_t^\pi &= Y_{t_i}^\pi - f(t_i, X_{t_i}^\pi, Y_{t_i}^\pi, Z_{t_i}^\pi)(t-t_i) + (Z_{t_i}^\pi)^T(W_t - W_{t_i}) \end{aligned}$$

By straightforward computations (including developments via Itö lemma), we can achieve the the following quantities:

$$\begin{matrix} d(\delta X_t) & d(\delta Y_t) \\ d|\delta X_t|^2 & d|\delta Y_t|^2 \end{matrix}$$

By a double usage of RMS-GM inequality and algbraic manipulations, we will achieve an estimation of $\mathbb{E}[|\delta X_t|^2]$ involving itself. We then apply the Grönwall lemma and the similar approach on $\mathbb{E}[|\delta Y_t|^2]$.

Combining the results with the definition of the objective function, we can achieve the desired result by manipulating the parameters λ.

Lemma 4: *Let* $X_{t_i}^{\pi,j} \in L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P})$ *for* $0 \leq i \leq N, j = 1, 2$*. Suppose* $Y^{\pi,j}$ *and* $Z^{\pi,j}$ *satisfy:*

$$\begin{cases} Y_T^{\pi,j} &= g(X_T^{\pi,j}) \\ Z_{t_i}^\pi &= \frac{1}{h}[Y_{t_{i+1}}^\pi \Delta W_i|\mathcal{F}_{t_i}] \\ Y_{t_{i+1}}^\pi &= \mathbb{E}[Y_{t_{i+1}}^{\pi,j} + f(t_i, X_{t_i}^{\pi,j}, Y_{t_i}^{\pi,j}, Z_{t_i}^{\pi,j})h|\mathcal{F}_{t_i}]. \end{cases} \tag{48}$$

Then for any $\lambda_7 > f_z$ and h is sufficiently small, we have:

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

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

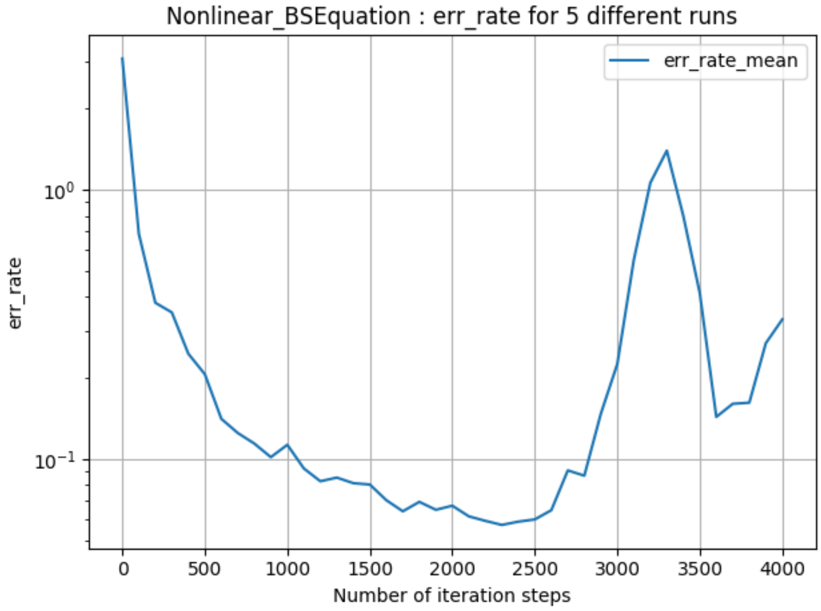
Lemma 5: *Consider again a system (8) and processes that satisfies it. When $h < \frac{1}{\sqrt{K}}$, there exists deterministic functions U_i^π and V_i^π such that $Y_{t_i}^{\pi,'} = Y_i^\pi(X_{t_i}^\pi, Y_{t_i}^\pi)$ and $Z_{t_i}^{\pi,'} = V_i^\pi(X_{t_i}^\pi, Y_{t_i}^\pi)$ satisfy:*

$$\begin{cases} Y_T^{\pi,'} &= g(X_T^\pi) \\ Z_{t_i}^{\pi,'} &= \frac{1}{h}[Y_{t_{i+1}}^{\pi,'} \Delta W_i | \mathcal{F}_{t_i}] \\ Y_{t_{i+1}}^{\pi,'} &= \mathbb{E}[Y_{t_{i+1}}^{\pi,'} + f(t_i, X_{t_i}^\pi, Y_{t_i}^{\pi,'}, Z_{t_i}^{\pi,'})h | \mathcal{F}_{t_i}]. \end{cases} \tag{49}$$

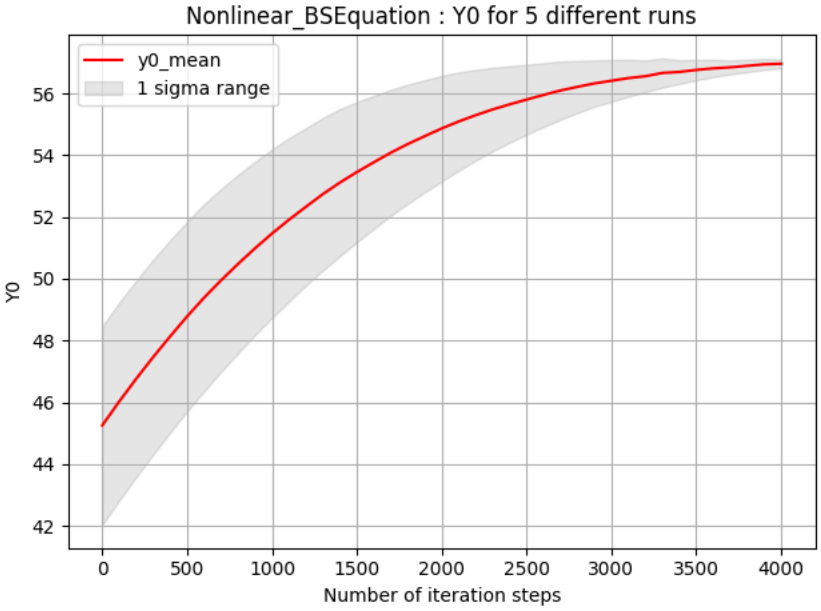
4 Numerical Experiments

We have chosen the non-linear Black-Scholes PDE ans the HJB equation to simulate using the deep BSDE method. We have used the same setups and parameters as in [1], trying to reproduce the same results (or somewhat better). Basically we were able to achieve results of same order as in the articles, which are fairly inspiring and satisfying. The actual differences between our experiments and those in the articls are that we had a slightly better equipped Macbook and learning rate with tiny modifications. Resolver codes, graphics and output logs can be found in the ZIP file. Our material was a Macbook with a 2,6 GHz Intel Core i7 processor, RAM memory of 16 GB 2133 MHz LPDDR3 and a video card of Radeon Pro 450 2 GB. For a collection of information concerning these simulations, please refer to the chart at the end of the section. Also note that our error rate was calculated and pictured with the unit of percentage. Therefore our results are clearly coherent with the article [1].

4.1 Non-linear Black-Scholes Equation with Default Risk

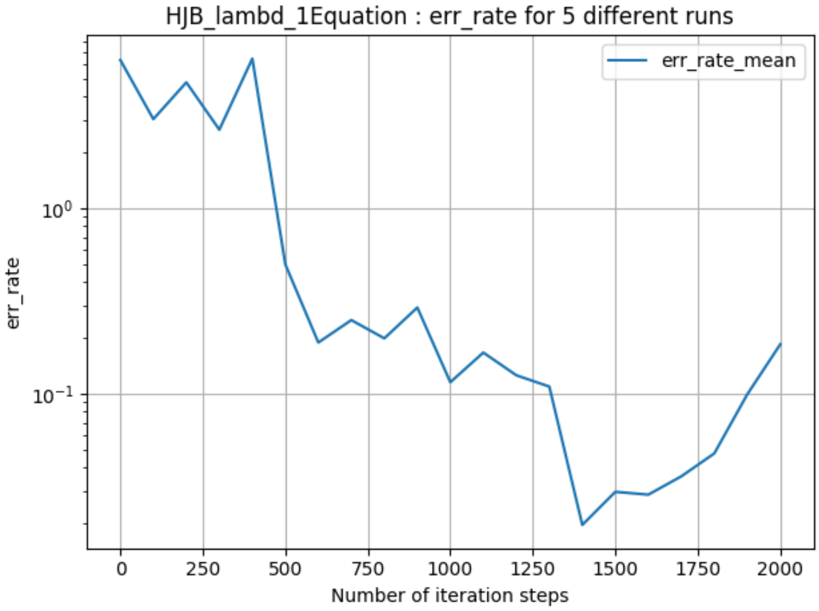


We have made use of a classical finite difference method to approach the classic numerical solution of the non-linear BS equation and we complete the error rate curve with respect to the number of iteration steps. According to our analysis, the relatively apparent bump in the middle of increasing number of iteration may be caused by the residual error from the part of the classic numerical methods. After a peak of error around 3200 iterations, the error decrease considerably after then.

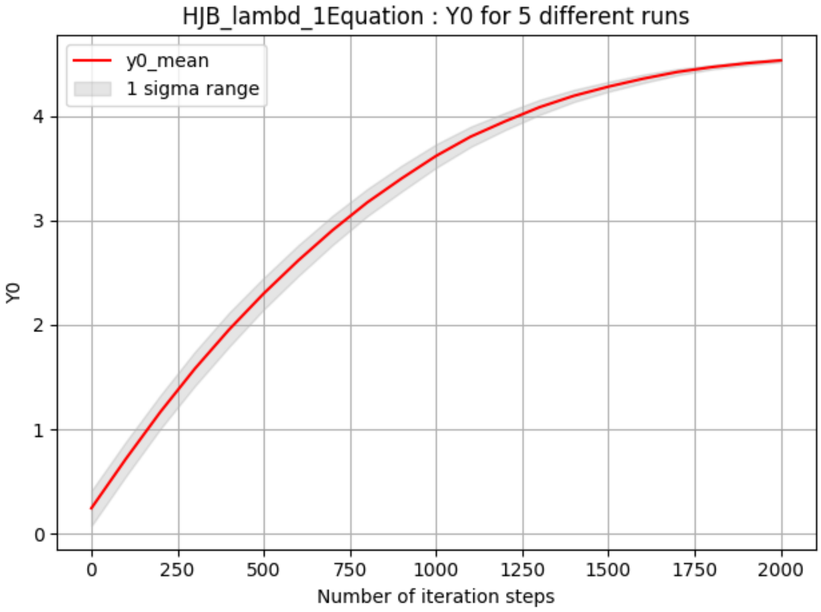


We see clearly that the convergence pattern is the same as what was shown in [1] and the limit is almost the same value 57.3.

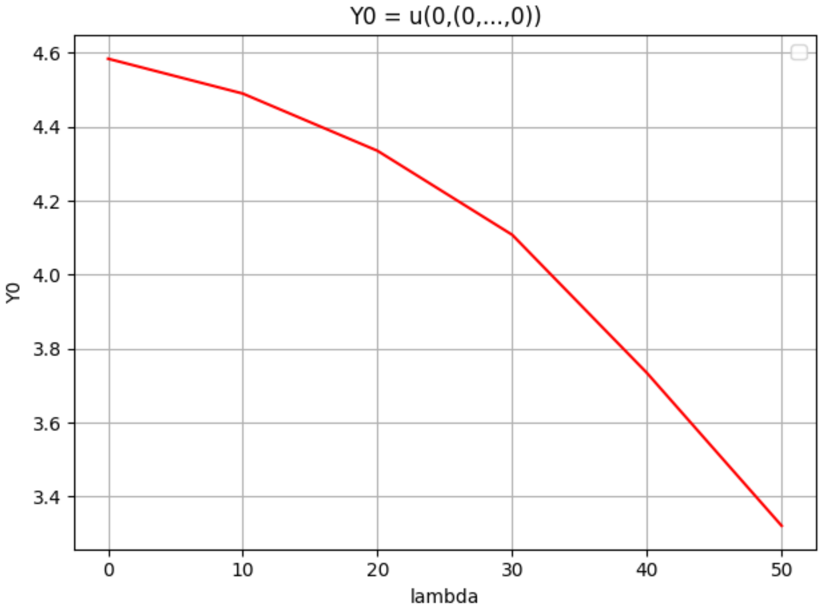
4.2 HJB Equation



Our experiments demonstrate the same tendency as in [1]. The error rate decreases drastically at the beginning phase and then it bounced back to some extent, eventually it convergences at the level of 0.17



From Fig.2 in P7 of [1], we can read that when $\lambda = 1$, the value Y_0 should be slightly less than 4.6, which justifies the convergence of our simulations.



Here we practiced the experiments for several times and the decreasing curve of the value $u(0, 0, ..., 0) = Y_0$ is obvious. Nevertheless, these initial values are different with those from [1] when λ is big. This maybe caused by the probability fluctuation of the deep neural network and the different learning rates chosen by us and the authors.

equationn	learning Rate	Error Rate	Computing Time(s)
Non-linear Black-Scholes equation	0.008	0.33%	855
HJB equation	0.005	0.18%	194.5

Our results are very consistent with that of [1] and the computing time is about a half of what has been used in the article. This is generally because our equipment has certain additional computing power.

5 Conclusion

In this project, we have investigate the newly proposed deep BSDE method in order to avoid the curse of dimensionality that appears when the number of agents in some PDEs are large. We have studied the algorithmic and mathematical structure of the methods. Then we perform the BSDE reformulation on the classical semi-linear parabolic PDEs and proved the spatial regularity and convergence of the method for decoupled FBSDEs. Numerical tests are also carried out with desirable outputs. Further tests and studies regarding the weak-coupling and monotonicity assumptions could be interesting: To which condition can we make less restricted assumptions? What are the impacts of specific parameters on the performance of the algorithms? Why do the error rates somehow oscillate? Can we improve the algorithms to achieve more stable and consistent error behavior? We believe that the asymptotic behaviors of simulating processes can be the next key subject in this field.

6 References

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

[2] 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, 2017.

[3] El Karoui, N., Peng, S. Quenez, M.-C. Backward stochastic differential equations in finance, *Mathematical Finance* **7**, 1-71 (1997).

[4] Pardoux, É & Peng, S. Backward stochastic differential equations and quasilinear parabolic partial differential equations. In *Stochastic Partial Differential Equations and their Applications (Charlotte, NC, 1991)*, vol. 178 of *Lecture Notes in Control and Inform. Sci.,* 200-217 (Springer, Berlin, 1992).

[5] Pardoux, É. & Peng, S. Forward-backward stochastic differential equations and quasilinear parabolic PDEs. *Probability Theory and Related Fields* **114**, 123-150 (1999).

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

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