Deep learning methods for high dimensional PDE's

An application to N-agent games

by

Carlos Daniel Contreras Quiroz

Advisor: Mauricio Junca

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Master in
Mathematics

at the
Universidad de los Andes
2023

Deep learning methods for high dimensional PDE's

An application to N-agent games

Carlos Daniel Contreras Quiroz

Abstract

Nada

Acknowledgements

A mi lulú y mi pancita.Nadita.

Contents

List of Figures

Chapter 1

Introduction

Las ecuaciones en derivadas parciales aparecen comúnmente como herramientas útiles para la modelación en múltiples disciplinas. Se encuentran frecuentemente aplicaciones en ciencias naturales como la física y biología, en diseño en ingeniería , y también en áreas como la economía y finanzas. Sin embargo, las propiedades matemáticas de las ecuaciones que aparecen son tan diversas como las áreas en que se aplican, y aunque se pueden clasificar parcialmente según algunas de sus características, no podría existir una teoría completa que describa nuestro conocimiento sobre estas.

Por otro lado, las soluciones analíticas a estos modelos generalmente no están a nuestro alcance, por lo que es necesario recurrir a métodos numéricos para obtener aproximaciones. Para esto, usualmente se recurre a métodos clásicos como diferencias finitas, elementos finitos, volúmenes finitos o métodos espectrales, para los cuales existe una amplia teoría que soporta y justifica rigurosamente su funcionamiento.

No obstante, la aplicación de estos métodos a problemas particulares a veces se restringe por propiedades especificas de la ecuación que se resuelve. Por ejemplo, los métodos mencionados sufren de la maldición de la dimensionalidad (*"the curse of dimentionality"*), esto es, su complejidad computacional escala exponencialmente en la dimensión del problema, por lo que su uso se restringe a problemas de dimensión baja ($n = 1, 2, 3, 4$). Lo anterior dificulta su implementación en aplicaciones como valoración en matemática financiera, donde la dimensión del problema está determinada por el número de activos considerados . También, su eficiencia computacional se reduce considerablemente conforme se aumenta la complejidad de los dominios en que se resuelven, o por las no-linealidades que aparecen, como es el caso de la ecuación de Navier-Stokes modelando flujos turbulentos.

Otra área en donde estos inconvenientes aparecen es en el análisis de datos y aprendizaje de maquinas. Por ejemplo, la complejidad de algunos modelos de regresión no lineal crece exponencialmente con el tamaño de los datos subyacentes. Para este tipo de problemas se han desarrollado herramientas poderosas que permiten modelar problemas en altas dimensiones y con posibles no linealidades. Entre estas, las redes neuronales han demostrado ser de gran utilidad como modelo para representar funciones con estas complejidades[1].

En consecuencia, intentando replicar el éxito obtenido con estas herramientas en aprendizaje de máquinas, recientemente han surgido nuevas perspectivas para aproximar soluciones de ecuaciones en derivadas parciales usando estas mismas herramientas. Entre estas se encuentran las PINNs (Physics Informed Neural Networks)[**PINNs**, **PINNS2**], FNO (Fourier Neural Operators)[2], y DGM (Deep Garlekin Method)[3]. La evidencia práctica muestra que estos métodos pueden proporcionar soluciones en casos donde los clásicos no [4, 5], a pesar de usualmente no competir con su eficiencia en las situaciones donde los últimos sí aplican. Además, se ha venido desarrollando un marco teórico riguroso que permite justificar su aplicación en situaciones específicas.

Chapter 2

Backward stochastic differential equations and PDEs

When addressing deterministic optimal control problems of dynamical systems, there are two approaches, one involving Bellman's dynamic programming principle, and the other relying on the Pontryagin's maximum principle. The former approach leads to a partial differential equation, the Hamilton-Jacobi-Bellman equation, to be solved for the value function and the optimal control of the process. The latter leads to a system of ordinary differential equations, one equation forward in time for the state and one backward in time for its adjoint.

The stochastic version of these problems is solved by methods analogous to those of the deterministic case. However, there are issues with desirable mathematical properties of solutions when we state them extending directly the ones proposed by deterministic methods. That is the case of the stochastic version of the Pontryagin's maximum principle, in which the backward differential equation cannot be stated directly as an SDE with terminal condition, as the solution is not guaranteed to be adapted to the filtration generated by the brownian motion.

The theory of backward stochastic differential equations (BSDEs) emerged in Bismut's [6] early work, and later generalized by Pardoux and Peng [7], as an attempt to formalize the application of the stochastic maximum principle. Here we give an introduction and compilation of results about them based on [8, 9, 10, 11], including its relation with a certain class of nonlinear parabolic partial differential equations, which will be the main tool for the method explained in the following chapters.

2.1 Backward stochastic differential equations

2.1.1 Motivation

Let's introduce the necessity for a different formulation of stochastic differential equations through an example [10]. In the usual setting for a stochastic differential equation (SDE), we specify the evolution of a $\mathbb{R}^n$-valued stochastic process X_t through its dynamics and an initial value $x_0 \in \mathbb{R}^n$(possibly random), in the form

$$X_t = x_0 + \int_0^t \mu(t, X_t)dt + \int_0^t \sigma(t, X_t)dW_t, \tag{2.1}$$

or equivalently,

$$\begin{aligned} dX_t &= \mu(t, X_t)dt + \sigma(t, X_t)dW_t \\ X_0 &= x_0, \end{aligned} \tag{2.2}$$

where W_t is a d-dimensional Brownian motion process and the stochastic integral is defined in the Ito sense.

We know that, under some Lipschitz and boundedness conditions for the drift μ and the volatility σ, the equation with initial condition (2.2) has a unique solution which is adapted with respect to the filtration $\mathbb{F} = (\mathcal{F}_t)_t$ generated by W_t.

Now, what happens if we consider the problem (2.2) with a terminal condition at time $T > 0$? Consider, for instance, the particular case with $\mu(t, X_t) = \sigma(t, X_t) = 0$, and a square-integrable $\mathcal{F}_T$-measurable random variable $\xi \in L^2(0, T)$ for which we try to solve the problem of finding a process Y_t such that

$$\begin{aligned} dY_t &= 0 \\ Y_t(T) &= \xi. \end{aligned} \tag{2.3}$$

This equation has a unique solution given by $Y(t) = \xi$, which is not necessarily $\mathcal{F}_t-$measurable for every $0 \leq t \leq T$, and therefore (2.3) may not have solution in the usual SDE sense.

Despite this, we can try to solve this problem reinterpreting the solution to (2.3) based on the following representation theorem.

Theorem 2.1.1 (Martingale representation theorem [12])**.** *Let $(M_t)_{0 \leq t \leq T}$ be a continuous $\mathbb{R}^n$-valued square-integrable martingale with respect to $\mathcal{F}_t$, the augmented filtration generated by an d-dimensional Brownian motion $(W_t)_t$. Then, there is a unique $\mathbb{R}^{n \times d}$-valued $\mathcal{F}_t$-adapted stochastic process $f(s)$, with $\mathbb{E}[\int_0^T |f|^2 dt] < \infty$, such that*

$$M_t = M_0 + \int_0^t f(s)dW_s \quad \textit{for} \quad t \in [0, T], \tag{2.4}$$

where the uniqueness is interpreted in the mean squared norm.

We can intend to enforce the solution Y_t to be $\mathcal{F}_t-$measurable for every $0 \leq t \leq T$ by taking its conditional expectation with respect to the evolving σ-algebra

$$Y(t) := \mathbb{E}[\xi | \mathcal{F}_t], \tag{2.5}$$

which satisfies the terminal condition $Y(T) = \xi$, since ξ is $\mathbb{F}_T$-measurable. Thus, as a consequence of the Martingale representation theorem 2.1.1, we conclude that there exist a square-integrable $\mathcal{F}_t$-measurable process Z_t such that

$$Y(t) = Y(0) + \int_0^t Z_s dW_s \quad \text{for} \quad t \in [0, T], \tag{2.6}$$

which can be written as

$$\begin{aligned} dY_t &= Z_t dW_t \\ Y(T) &= \xi \end{aligned} \tag{2.7}$$

Therefore, problem (2.3) can be reinterpreted as in problem (2.7), that we will denote as a bacward stochastic differential equation (BSDE) , in which we seek a pair of processes (Y_t, Z_t) that will provide an adapted solution to our original problem. Indeed, the process Z_t will "steer" the system so that the process Y_t remains adapted, and is thus called a control process. It is not possible to revert time as $t \to T - t$ as the filtration goes only in one direction [13].

Finally, we can write this equation in another form. Note that (2.7) is a forward SDE problem, hence we can solve for $Y(0)$ in the integral form , and so we have

$$Y(0) = \xi - \int_0^T Z_s dW_s, \tag{2.8}$$

that is inserted in (2.6) to obtain

$$Y(t) = \xi - \int_0^T Z_s dW_s + \int_0^t Z_s dW_s = \xi - \int_t^T Z_s dW_s \quad \forall t \in [0, T], \tag{2.9}$$

which is the standard way to write the BSDE in integral form.

2.1.2 Some useful theorems

Now that we have motivated the use of BSDEs, we follow [14] to provide a formal definition and prove that under certain regularity conditions, we can ensure the existence of a solution for that kind of equations.

Let be $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space and $T > 0$ a fixed horizon time. We consider a d-dimensional Brownian motion $W = (W_t)_{t\in[0,T]}$ and let $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ be the corresponding natural augmented filtration (i.e with the completeness and right continuity conditions).

Denote by $\mathbb{S}^2(0, T)$ the set of $\mathbb{R}$-valued progressively measurable processes Y_t such that

$$\mathbb{E}\left[\sup_{0\le t\le T} |Y_t|^2\right] < \infty, \tag{2.10}$$

and by $\mathbb{H}^2(0, T)^d$ the set of $\mathbb{R}^d$-valued progressively measurable processes Z_t such that

$$\mathbb{E}\left[\int_0^T |Z_t|^2 dt\right] < \infty. \tag{2.11}$$

Here we consider the backward stochastic differential equation

$$\begin{aligned} dY_t &= -f(t, Y_t, Z_t)dt + Z_t \cdot dW_t \\ Y(T) &= \xi \end{aligned} \tag{2.12}$$

Definition 2.1.1. *A solution to the BSDE* (2.12) *is a pair* $(Y, Z) \in \mathbb{S}^2(0,T) \times \mathbb{H}^2(0,T)^d$ *such that*

$$Y_t = \xi + \int_t^T f(s, Y_s, Z_s)\, ds - \int_t^T Z_s \cdot dW_s, \quad 0 \le t \le T \tag{2.13}$$

Now we establish an existence and uniqueness theorem for $\mathbb{R}$-valued process, which can be extended to $\mathbb{R}^m$-valued processes.

Assumptions 2.1.2. *Let* (ξ, f) *satisfy*

I. $\xi \in L^2(\Omega, \mathcal{F}_T, \mathbb{P}; \mathbb{R})$

II. $f : \Omega \times [0,T] \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ *such that*

a) $f(\cdot, t, y, z)$*, written* $f(t, y, z)$ *for simplicity, is progressively measurable for all* y, z

b) $f(t, 0, 0) \in \mathbb{H}^2[0,T]$

c) f *is uniformly Lipschitz in* (y, z)*, i.e ,there exist a constant* C_f *such that for all* $y_1, y_2 \in \mathbb{R} \times \mathbb{R}$ *and* $z_1, z_2 \in \mathbb{R}^d \times \mathbb{R}^d$ *we have*

$$|f(t, y_1, z_1) - f(t, y_2, z_2)| \le C_f \left(|y_1 - y_2| + |z_1 - z_2|\right) \quad a.s \tag{2.14}$$

Theorem 2.1.2 (Existence and uniqueness of solutions to BSDEs [14])**.** *Given a pair* (ξ, f)*, called the terminal condition and the driver of the BSDE, that satisfy the assumptions 2.1.2 , there exist a unique solution* (Y, Z) *to the backward stochastic differential equation* (2.12)*.*

To give a demonstration we will need the following inequalities about SDEs, whose proofs will be omitted.

Theorem 2.1.3 (Doob's martingale inequality [12])**.** *Let* $\{M_t\}_t \ge 0$ *be a* $\mathbb{R}^m$*-valued martingale in* $L^p(\Omega; \mathbb{R}^m)$*. Let* $[0,T]$ *be a bounded interval with* $T > 0$ *and let* $p > 1$*. Then*

$$\mathbb{E}\left[\sup_{0 \le t \le T} |M_t|^p\right] \le \left(\frac{p}{p-1}\right)^p \mathbb{E}[|M_T|^p], \tag{2.15}$$

in particular, if $p = 2$*,*

$$\mathbb{E}\left[\sup_{0 \le t \le T} |M_t|^2\right] \le 4\mathbb{E}[|M_T|^2]. \tag{2.16}$$

Theorem 2.1.4 (Burkholder-Davis-Gundy inequality [12])**.** *Let* $g \in L^2(\mathbb{R}^+; \mathbb{R}^{m \times d})$*. Define for* $t \ge 0$

$$x(t) = \int_0^t g(s) dW_s \quad \text{and} \quad A(t) = \int_0^t |g(s)|^2 ds$$

then, for every $p > 0$ *there exist universal positive constants* c_p, C_p*, depending only on* p*, such that the following inequalities hold,*

$$c_p \mathbb{E}[|A(t)|^{\frac{p}{2}}] \le \mathbb{E}\left[\sup_{0 \le s \le t} |x(s)|^p\right] \le C_p \mathbb{E}[|A(t)|^{\frac{p}{2}}], \tag{2.17}$$

in particular, if $p = 1$*, we can take* $c_p = \frac{1}{2}$ *and* $C_p = 4\sqrt{2}$.

Proof of theorem 2.1.2 . Here we give a fixed point argument. To do it, lets consider a pair of process $(U, V) \in \mathbb{S}^2(0,T) \times \mathbb{H}^2(0,T)^d$ and, as in the motivation example, consider the martingale

$$M_t = \mathbb{E}\left[\xi + \int_0^T f(s, U_s, V_s)ds \middle| \mathcal{F}_t\right], \tag{2.18}$$

which is square-integrable under the hypothesis on (ξ, f). Using to the martingale representation theorem 2.1.1, we deduce the existence and uniqueness of a process $Z_s \in \mathbb{H}^2(0,T)^d$ such that

$$M_t = M_0 + \int_0^t Z_s \cdot dW_s. \tag{2.19}$$

Now, define the process Y_t for $0 \leq t \leq T$ as

$$\begin{aligned} Y_t &= \mathbb{E}\left[\xi + \int_t^T f(s, U_s, V_s)ds \middle| \mathcal{F}_t\right] = \mathbb{E}\left[\xi + \int_0^T f(s, U_s, V_s)ds - \int_0^t f(s, U_s, V_s)ds \middle| \mathcal{F}_t\right] \\ &= M_t - \int_0^t f(s, U_s, V_s)ds \end{aligned} \tag{2.20}$$

and note that from this and using (2.19), Y_t satisfies

$$\begin{aligned} Y_t &= M_0 + \int_0^t Z_s \cdot dW_s - \int_0^t f(s, U_s, V_s)ds \\ &= \xi + \int_t^T f(s, U_s, V_s)ds - \int_t^T Z_s \cdot dW_s. \end{aligned} \tag{2.21}$$

Thus, consider the function $\Phi : \mathbb{S}^2(0,T) \times \mathbb{H}^2(0,T)^d \to \mathbb{S}^2(0,T) \times \mathbb{H}^2(0,T)^d$ that maps the pair (U, V) to the pair (Y, Z) constructed as above, $\Phi(U, V) = (Y, Z)$. Note that it is well-defined as the Z process is unique, and by Doob's martingale inequality 2.1.3 we have

$$\mathbb{E}\left[\sup_{0 \leq t \leq T} \left|\int_t^T Z_s \cdot dW_s\right|^2\right] \leq 4\mathbb{E}\left[\int_0^T |Z_s|^2 ds\right] < \infty, \tag{2.22}$$

and therefore, by assumptions I, IIa) and IIb), Y_t lies in $\mathbb{S}^2(0,T)$. Also note that a solution to the BSDE (2.12) is a fixed point of Φ. We will show that such fixed point exist by showing it is a contraction if we endow the $\mathbb{S}^2(0,T) \times \mathbb{H}^2(0,T)^d$ space with the metric

$$\|(Y, Z)\|_\beta = \left(\mathbb{E}\left[\int_0^T e^{\beta s}(|Y_s|^2 + |Z_s|^2)ds\right]\right)^{\frac{1}{2}}, \tag{2.23}$$

where $\beta > 0$ is a parameter to be chosen later.

To show that Φ is a contraction, let $(U,V),(U',V') \in \mathbb{S}^2(0,T) \times \mathbb{H}^2(0,T)^d$ and $(Y,Z) = \Phi(U,V)$, $(Y',Z') = \Phi(U',V')$. We denote $(\bar{U},\bar{V}) = (U-U',V-V')$, $(\bar{Y},\bar{Z}) = (Y-Y',Z-Z')$ and $\bar{f}_t = f(t,U_t,V_t) - f(t,U'_t,V'_t)$.

Using equation (2.21), we know that $\bar{Y}_s$ satisfies

$$\bar{Y}_s = -\int_0^t \bar{f}_s ds + \int_0^t \bar{Z}_s \cdot dW_s \tag{2.24}$$

So let's apply Ito's formula to the process $e^{\beta s}|\bar{Y}_s|^2$ between 0 and T to obtain

$$\begin{aligned} e^{\beta T}|\bar{Y}_T|^2 = |\bar{Y}_0|^2 &+ \int_0^T (\beta e^{\beta s}|\bar{Y}_s|^2 - 2e^{\beta s}\bar{Y}_s \cdot \bar{f}_s + e^{\beta s}|\bar{Z}_s|^2) ds \\ &+ \int_0^T 2e^{\beta s}\bar{Y}_s\bar{Z}_s \cdot dW_s. \end{aligned} \tag{2.25}$$

Observe that we can apply the Burkholder-Davis-Gundy inequality 2.1.4 with $p = 1$ to the following expectation of the supremum associated with the last term

$$\begin{aligned} \mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t 2e^{\beta s}\bar{Y}_s\bar{Z}_s \cdot dW_s\right|\right] &\le 4\sqrt{2}\,\mathbb{E}\left[\left(\int_0^T 4e^{2\beta s}|\bar{Y}_s|^2|\bar{Z}_s|^2 ds\right)^{\frac{1}{2}}\right] \\ &\le 4\sqrt{2}e^{\beta T}\,\mathbb{E}\left[\sup_{0\le t\le T}|Y_t|^2 + \int_0^T |\bar{Z}_s|^2 ds\right] \\ &< \infty, \end{aligned} \tag{2.26}$$

which shows that the local martingale $\int_0^t 2e^{\beta s}\bar{Y}_s\bar{Z}_s \cdot dW_s$ is actually a uniformly integrable martingale and therefore its expected value remains constant zero. Also, note that $\bar{Y}_T = Y_T - Y'_T = \xi - \xi = 0$.

Using these facts, take the expected value to (2.25) and reorder terms to obtain

$$\begin{aligned} \mathbb{E}\left|\bar{Y}_0\right|^2 + \mathbb{E}\left[\int_0^T e^{\beta s}\left(\beta\left|\bar{Y}_s\right|^2 + \left|\bar{Z}_s\right|^2\right) ds\right] &= 2\mathbb{E}\left[\int_0^T e^{\beta s}\bar{Y}_s \cdot \bar{f}_s ds\right] \\ \le 2C_f\mathbb{E}\left[\int_0^T e^{\beta s}\left|\bar{Y}_s\right|\left(\left|\bar{U}_s\right| + \left|\bar{V}_s\right|\right) ds\right] &\quad \text{(by condition } IIc)) \\ \le 4C_f^2\mathbb{E}\left[\int_0^T e^{\beta s}\left|\bar{Y}_s\right|^2 ds\right] + \frac{1}{2}\mathbb{E}\left[\int_0^T e^{\beta s}\left(\left|\bar{U}_s\right|^2 + \left|\bar{V}_s\right|^2\right) ds\right], \end{aligned} \tag{2.27}$$

so if we choose $\beta = 1 + 4C_f^2$ and ignore the $\mathbb{E}\left|\bar{Y}_0\right|^2$ term, we obtain

$$\mathbb{E}\left[\int_0^T e^{\beta s}\left(\left|\bar{Y}_s\right|^2 + \left|\bar{Z}_s\right|^2\right) ds\right] \le \frac{1}{2}\mathbb{E}\left[\int_0^T e^{\beta s}\left(\left|\bar{U}_s\right|^2 + \left|\bar{V}_s\right|^2\right) ds\right], \tag{2.28}$$

which is $\|(\Phi(U,V))\|_\beta \le \frac{1}{2}\|(U,V)\|_\beta$, that means Φ is a contraction in a Banach space, as $\mathbb{S}^2(0,T) \times \mathbb{H}^2(0,T)^d$ is the product of Banach spaces, and therefore has a unique fixed point. ■

As in the every differential equation, there are cases where we can provide an explicit solution. The next theorem provides one for the BSDE with linear generator

Theorem 2.1.5 (Linear BSDEs [14]). *Let A_t,B_t be bounded progressively measurable processes with values in $\mathbb{R}$ and $\mathbb{R}^d$, C_t a process in $\mathbb{H}^2(0,T)$ and $\xi \in L^2(\Omega, \mathcal{F}_T, \mathbb{P}, \mathbb{R})$. Then, the linear backward stochastic differential equation*

$$\begin{aligned} dY_t &= -(A_tY_t + Z_t \cdot B_t + C_t)dt + Z_t \cdot dW_t \\ Y_T &= \xi \end{aligned} \tag{2.29}$$

has a unique solution, and is given by the formula

$$\Gamma_t Y_t = E\left[\Gamma_T \xi + \int_t^T \Gamma_s C_s ds \mid \mathcal{F}_t\right], \tag{2.30}$$

where Γ_t is the solution to the adjoint process

$$\begin{aligned} d\Gamma_t &= \Gamma_t(A_t dt + B_t \cdot dW_t) \\ \Gamma_0 &= 1 \end{aligned} \tag{2.31}$$

Proof. First apply Ito's formula to $\Gamma_t Y_t$ to obtain

$$\begin{aligned} d(\Gamma_t Y_t) &= Y_t d\Gamma_t + \Gamma_t dY_t + d\Gamma_t dY_t \\ &= Y_t(\Gamma_t A_t dt + \Gamma_t B_t \cdot dW_t) + \Gamma_t(-(A_tY_t + Z_t \cdot B_t + C_t)dt + Z_t \cdot dW_t) \\ &\quad + \Gamma_t Z_t \cdot B_t dt \\ &= -\Gamma_t C_t dt + \Gamma_t(Z_t + Y_t B_t) \cdot dW_t, \end{aligned} \tag{2.32}$$

that can be written in integral form as

$$\Gamma_t Y_t + \int_0^t \Gamma_s C_s ds = Y_0 + \int_0^t \Gamma_s \left(Z_s + Y_s B_s\right) \cdot dW_s. \tag{2.33}$$

We will show, as in the proof of theorem 2.1.2, that the stochastic integral in the last expression, which is a local martingale, is in fact a uniformly integrable martingale. We have $\mathbb{E}\left[\sup_{0\le t\le T} |\Gamma_t|^2\right] < \infty$,since A_t and B_t are bounded. Also, let's denote b_∞ the upper bound on B_t, then the following inequalities hold

$$\begin{aligned} \mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t \Gamma_s \left(Z_s + Y_s B_s\right) \cdot dW_s\right|\right] &\le 4\sqrt{2}\,\mathbb{E}\left[\left(\int_0^T |\Gamma_s|^2 |Z_s + Y_s B_s|^2 ds\right)^{\frac{1}{2}}\right] \\ &\quad \text{(By BDG inequality 2.1.4)} \\ &\le \frac{4\sqrt{2}}{2} E\left[\sup_{0\le t\le T} |\Gamma_t|^2 + 2\int_0^T |Z_t|^2\, dt + 2b_\infty^2 \int_0^T |Y_t|^2\, dt\right] \\ &< \infty. \end{aligned} \tag{2.34}$$

Consequently, the right-hand side of is a uniformly integrable martingale, and so, if we take expected values to the equality (2.33), we have

$$\begin{aligned}\Gamma_t Y_t + \int_0^t \Gamma_s C_s ds &= \mathbb{E}\left[\Gamma_T Y_T + \int_0^T \Gamma_s C_s ds \middle| \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\Gamma_T \xi + \int_0^T \Gamma_s C_s ds \middle| \mathcal{F}_t\right]\end{aligned} \tag{2.35}$$

and, as $\int_0^t \Gamma_s C_s ds$ is $\mathcal{F}_t$-measurable, we obtain

$$\Gamma_t Y_t = \mathbb{E}\left[\Gamma_T \xi + \int_t^T \Gamma_s C_s ds \middle| \mathcal{F}_t\right], \tag{2.36}$$

that is what we wanted to prove. The control solution Z_t can be obtained by the martingale representation theorem 2.1.1 applied to this process. ■

Finally, we state the next comparison principle for solution of BSDEs

Theorem 2.1.6 (Comparison principle for BSDEs [14])**.** *Let (ξ_1, f_1) and (ξ_2, f_2) two pairs of terminal conditions and generators satisfying assumptions 2.1.2,and let $(Y_{1,t}, Z_{1,t})$ and $(Y_{2,t}, Z_{2,t})$ the solutions to their corresponding BSDE. Suppose that*

1. $\xi_1 \leq \xi_2$ *a.s*
2. $f_1(t, Y_{1,t}, Z_{1,t}) \leq f_2(t, Y_{1,t}, Z_{1,t})$ $dt \times d\mathbb{P}$*-a.e*
3. $f_2(t, Y_{1,t}, Z_{1,t}) \in \mathbb{H}^2(0, T)$

Then $Y_{1,t} \leq Y_{2,t}$ for all $0 \leq t \leq T$, a.s. Furthermore, if $Y_{2,0} \leq Y_{1,0}$, then $Y_{1,t} = Y_{2,t}$ for $t \in [0, T]$. In particular, if $\mathbb{P}(\xi_1 < \xi_2) > 0$ or $f_1(t, \cdot, \cdot) < f_2(t, \cdot, \cdot)$ on a set with strictly positive measure $dt \times d\mathbb{P}$ then $Y_{1,0} < Y_{2,0}$.

Proof. To simplify notation, we give a proof with $d = 1$. We denote $\bar{Y_t} = Y_{2,t} - Y_{1,t}$ and $\bar{Z_t} = Z_{2,t} - Z_{1,t}$. Then $(\bar{Y_t}, \bar{Z_t})$ satisfy the BSDE

$$\begin{aligned}d\bar{Y_t} &= -\left(\Delta_t^y \bar{Y_t} + \Delta_t^z \bar{Z_t} + \bar{f_t}\right) dt + \bar{Z_t} dW_t \\ \bar{Y_T} &= \xi_2 - \xi_1,\end{aligned} \tag{2.37}$$

where

$$\begin{aligned}\Delta_t^y &= \frac{f_2\left(t, Y_{2,t}, Z_{2,t}\right) - f_2\left(t, Y_{1,t}, Z_{2,t}\right)}{Y_{2,t} - Y_{1,t}} 1_{Y_{2,t} - Y_{1,t} \neq 0} \\ \Delta_t^z &= \frac{f_2\left(t, Y_{1,t}, Z_{2,t}\right) - f_2\left(t, Y_{1,t}, Z_{1,t}\right)}{Z_{2,t} - Z_{1,t}} 1_{Z_{2,t} - Z_{1,t} \neq 0} \\ \bar{f_t} &= f_2\left(t, Y_{1,t}, Z_{1,t}\right) - f_1\left(t, Y_{1,t}, Z_{1,t}\right).\end{aligned} \tag{2.38}$$

By assumption, f_2 is Lipschitz in y, z, hence Δ_t^y and Δ_t^z are bounded. Moreover, $\bar{f_t} \in \mathbb{H}^2(0, T)$. Therefore, the solution to (2.37) is given by theorem 2.1.5 as

$$\Gamma_t \bar{Y_t} = \mathbb{E}\left[\Gamma_T\left(\xi_2 - \xi_1\right) + \int_t^T \Gamma_s \bar{f_s} ds \middle| \mathcal{F}_t\right], \tag{2.39}$$

where Γ_t satisfies

$$\begin{aligned} d\Gamma_t &= \Gamma_t(\Delta_t^y dt + \Delta_t^z dW_t) \\ \Gamma_0 &= 1. \end{aligned} \tag{2.40}$$

Note that Γ_t is strictly positive(Why?). We conclude the stated result using that $\xi_2 - \xi_2 \geq 0$ by assumption 1), and $\bar{f}_t \geq 0$ by assumption 2).

■

2.1.3 Forward-Backward stochastic differential equations

Now we consider a special case of backward stochastic differential equations in which the randomness of the drift enters through a process satisfying a forward stochastic differential equation. In its more general form, the problem is stated as find three processes $(X_t, Y_t, Z_t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{m\times d}$ such that

$$\begin{aligned} dX_s &= \mu(t, X_s, Y_s, Z_s)ds + \sigma(s, X_s, Y_s, Z_s)dW_s \\ X_t &= x \\ dY_s &= -f(s, X_s, Y_s, Z_s)ds + Z_s dW_s \\ Y_T &= g(X_T), \end{aligned} \tag{2.41}$$

for all $t \leq s \leq T$, where μ, σ and g are known functions, and x is the initial condition at starting time s. This coupled system is called a forward-backward stochastic differential equation (FBSDE).

This problem is rather difficult, as the coupling between the processes may forbid a solution to exist. There are conditions on μ, σ, g where we can establish the existence and uniqueness of solutions to the former system, but their detailed proof is very technical and thus is not presented here, see [8].

However, we can say something simpler about the decoupled case

$$\begin{aligned} dX_s &= \mu(s, X_s)ds + \sigma(s, X_s)dW_s \\ X_t &= x, \\ dY_s &= -f(s, X_s, Y_s, Z_s)ds + Z_s dW_s \\ Y_T &= g(X_T). \end{aligned} \tag{2.42}$$

for all $t \leq s \leq T$.

In this case, if μ and σ satisfy enough regularity conditions to ensure that a solution to the forward SDE in (2.42) exists, for example, if they are Lipschitz and bounded, then we can solve it for the process X_t and insert the solution into the backward equation in 2.42 and solve for the backward process. However, the main property of FBSDEs is that the solution process (Y, Z) of the BSDE can be written as a deterministic function of time and the state process, in this case the solution is said to be *markovian*.

Let's establish this assertions in the following theorem.

Assumptions 2.1.3. *Let* (μ, σ, f, g)*. There exist a constant* $C > 0$ *such that for all* x, y, t

I. $|\mu(t,x)-\mu(t,y)|+|\sigma(t,x)-\sigma(t,y)|\le C(1+|x-y|)$

II. $|f(t,x,y_1,z_1)-f(t,x,y_2,z_2)|\le C(|y_1-y_2|+|z_1-z_2|)$

III. $|\sigma(t,x)|+|\mu(t,x)|\le C(1+|x|)$

IV. $|f(t,x,y,z)|+|g(x)|\le C(1+|x|^p)$ *para* $p\ge\frac{1}{2}$

Theorem 2.1.7 (Existence and markovianity of solutions of FBSDEs [15])**.** *Under assumptions 2.1.3, the uncoupled forward-backward stochastic differential equation (2.42) has a unique solution $(X_s^{t,x},Y_s^{t,x},Z_s^{t,x})$ starting from x at time t. Moreover, $(Y_s^{t,x},Z_s^{t,x})$ is adapted to the future σ-algebra of W after t, i.e, it is $\mathcal{F}_s^t$-adapted where for each $s\in[t,T]$ we define $\mathcal{F}_s^t=\sigma(W_u-W_t,t\le u\le s)$. In particular, $Y_t^{t,x}$ is deterministic and for $0\le s\le t$ we have $Y_s^{t,x}=Y_t^{t,x}$ and $Z_s^{t,x}=0$.*

Proof. The first part about existence and uniqueness of solution to the FBSDE follows from the fact that in assumptions 2.1.3, I and III are the standard Lipschitz and linear growth conditions that guarantee the existence of a solution for the forward process [12], and that II and IV are sufficient conditions to ensure the existence of the solution to the backward process from theorem 2.1.2.

For the second part, consider the translated Brownian motion W' and its associated filtration given by $W'_s=W_{t+s}-W_t$ for $0\le s\le T-t$ and $\mathcal{F}'_s:=\mathcal{F}^t_{t+s}$ or $0\le s\le T-t$. Let $X_s'^{0,x}$ be the adapted solution to the SDE

$$\begin{aligned}dX'_s&=\mu(s,X'_s)dt+\sigma(s,X'_s)dW_s\\X'_0&=x.\end{aligned}\tag{2.43}$$

By the the uniqueness provided by the former theorems, we have $X_s^{t,x}=X_{s-t}'^{0,x}$ a.s for $s\in[0,T-t]$, hence $X_s^{t,x}$ is $\mathcal{F}_s^t$-adapted.

Now consider the associated $\mathcal{F}'$-adapted solution (Y'_s,Z'_s) with $s\in[0,T-t]$ to the BSDE

$$\begin{aligned}dY'_s&=-f(s+t,X'_s,Y'_s,Z'_s)ds+Z'_s\cdot dW_s\\Y'_{T-t}&=g(X'_{T-t}).\end{aligned}\tag{2.44}$$

We have that (Y'_{s-t},Z'_{s-t}) with $s\in[t,T]$ is also a solution of the backward equation in (2.42) in $[t,T]$. Hence, by the uniqueness provided before we have that $(Y'_{s-t},Z'_{s-t})=(Y_s^{t,x},Z_s^{t,x})$ for $s\in[t,T]$, therefore (Y'_{s-t},Z'_{s-t}) is $\mathcal{F}_s^t$-adapted.

∎

From now on, we will denote by

$$v(t,x):=Y_t^{t,x},\tag{2.45}$$

the deterministic function of t and x provided by the last theorem. We also notice that $Y_t=v(t,X_t)$ for $t\in[0,T]$.

2.2 The Feynman-Kac formulas

Now we shall establish the connection between stochastic differential equations with parabolic linear partial differential equations and its non-linear generalization based on backward stochastic differential equations.

2.2.1 The linear Feynman-Kac formula

We will start by the linear case to introduce the necessity for a non-linear generalization. Consider the $\mathbb{R}^n$-valued process $X_s^{t,x}$ defined to be the solution in $s \in [t, \infty)$ of the SDE

$$\begin{aligned} dX_s &= \mu(s, X_s)ds + \sigma(s, X_s)dW_s \\ X_t &= x, \end{aligned} \tag{2.46}$$

where, again, W_t is a d-dimensional Brownian motion, μ is $\mathbb{R}^n$-valued function, σ is a $(n \times d)$-valued matrix of functions and $x \in \mathbb{R}^n$ is the initial condition. We have the following estimate

Theorem 2.2.1 ([16]). *Let μ and σ satisfy conditions I and III of assumptions 2.1.3, then, there exist a constant $C > 0$ such that the solution to 2.46 satisfies*

$$\mathbb{E}\left[\sup_{t \le s \le T} |X_s|^2\right] \le C(1 + \mathbb{E}\big[|x|^2\big])e^{C(T-t)}. \tag{2.47}$$

Now, for a fixed $T > 0$, consider the following parabolic PDE with terminal condition for the function $v(t, x) : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}$,

$$\begin{aligned} &\frac{\partial v}{\partial t} + \mathcal{L}_t v - k(t, x)v + f(t, x) = 0 \\ &v(T, x) = g(x), \end{aligned} \tag{2.48}$$

where $f(x, t)$ and $g(x)$ are some $\mathbb{R}$-valued continuous functions, $k(x, t)$ is a non-negative $\mathbb{R}$-valued function, and $\mathcal{L}_t$ is the *generator* of the process X_s, defined as

$$\begin{aligned} \mathcal{L}_t v &= \mu(t, x) \cdot D_x v(t, x) + \frac{1}{2}\operatorname{Tr}(\sigma(t, x)\sigma^T(t, x)D^2_{xx}v(t, x)) \\ &= \sum_{i=1}^{n} \mu_i(t, x)\frac{\partial v}{\partial x_i}(t, x) + \sum_{1,k=1}^{n} a_{i,k}(t, x)\frac{\partial^2 v}{\partial x_i \partial x_k}(t, x), \end{aligned} \tag{2.49}$$

where we denote by $a_{i,k}$ the coefficients of the *diffusion matrix*, calculated as

$$a_{i,k} = \sum_{j=1}^{d} \sigma_{i,j}(t, x)\sigma_{k,j}(t, x). \tag{2.50}$$

The linear Feynman-Kac formula establishes a connection between the process satisfying (2.46) and the classical solution to equation (2.48) as follows

Assumptions 2.2.1. *Let μ and σ satisfy conditions I and III of assumptions 2.1.3, and assume that*

I. The functions $g(x) : \mathbb{R}^n \to \mathbb{R}$, $f(t,x) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ and $k(t,x) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^+$ are continuous.

II. The function $v(t,x) : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ is continuous in $[0,T] \times \mathbb{R}^n$, one time differentiable in t, and two times differentiable in x, and satisfies the PDE (2.48). Moreover, its first derivative in x is bounded, i.e , $|D_x v(t,x)| < M$ for some constant M and all $t \in [0,T]$ and $x \in \mathbb{R}^n$.

Theorem 2.2.2 (Linear Feynman-Kac formula [16])**.** *Under the assumptions 2.2.1, the solution $v(t,x)$ to the equation (2.48) admits the stochastic representation*

$$v(t,x) = \mathbb{E}\left[g(X_T^{t,x})e^{-\int_t^T k(s,X_s^{t,x})ds} + \int_t^T f(s,X_s^{t,x})e^{-\int_t^s k(u,X_u^{t,x})du}ds\right], \tag{2.51}$$

for all $t \in [0,T]$ and $x \in \mathbb{R}^n$. In particular, this solution is unique.

Proof. In order to simplify notation, we set $X_s = X_s^{x,t}$. Let's apply Ito's formula to the process $v(s,X_s)e^{-\int_t^s k(u,X_u)du}$ in $s \in [t,T]$ to obtain

$$\begin{aligned} e^{-\int_t^T k(u,X_u)du}v(T,X_T) &= e^{-\int_t^t k(u,X_u)du}v(t,X_t) \\ &+ \int_t^T e^{-\int_t^s k(u,X_u)du}\left(\frac{\partial v}{\partial t}(s,X_s) - k(s,X_s)v(s,K_s) + \mathcal{L}_s v(s,X_s)\right)ds \\ &+ \int_t^T e^{-\int_t^s k(u,X_u)du}D_x v(s,X_s)\cdot\sigma(s,X_s)dW_s \\ &= v(t,X_t) - \int_t^T e^{-\int_t^s k(u,X_u)du}f(s,X_s)ds \\ &+ \int_t^T e^{-\int_t^s k(u,X_u)du}D_x v(s,X_s)\cdot\sigma(s,X_s)dW_s, \end{aligned} \tag{2.52}$$

and therefore, using that $X_t = x$, $v(T,X_t) = g(X_T)$ and solving for $v(x,t)$, we have

$$\begin{aligned} v(t,x) = g(X_T)e^{-\int_t^T k(u,X_u)du} &+ \int_t^T f(s,X_s)e^{-\int_t^s k(u,X_u)du}ds \\ - \int_t^T e^{-\int_t^s k(u,X_u)du}&D_x v(s,X_s)\cdot\sigma(s,X_s)dW_s. \end{aligned} \tag{2.53}$$

To obtain the desired formula, we take expectation to this expression, and observe that the stochastic integral is a square integrable martingale by assumption 2.2.1 III on $D_x v$, the non-negativity of k, the linear growth condition on σ and the estimation 2.2.1, therefore it's expected value is constant 0. ∎

Note that we required that equation (2.48) has a classical smooth solution, for which we need some regularity conditions on μ, σ and growth conditions on f and g to ensure the uniform ellipticity of $\mathcal{L}_t$. Also note, that assumptions 2.2.1 are rather restrictive, especially the boundedness of D_x, but can be relaxed imposing some quadratic growth condition on v and additional growth condition on f and g. However, even if there is no smooth solution to this problem, the Feynman-Kac formula may provide a solution with other meaning, which will be named a viscosity solution and will be defined in what follows.

In addition, this formula is useful for approximating solutions $v(t,x)$ to PDEs of the form (2.48), even in high dimensions, where classical methods fails because of the *curse of dimensionality*. This expected value can be approximated by Monte-Carlo simulation using sample paths, whose rate of convergence is independent on the dimension n of the underlying process(Sure?). However, its use is limited to linear equations that may not be useful for certain problems.

2.2.2 The non-linear Feynman-Kac formula

Now we deal with a more general PDE than (2.48). We consider, for some fixed $T > 0$, the problem

$$\begin{aligned} &\frac{\partial v}{\partial t}(t,x) + \mathcal{L}_t v(t,x) + f(t,x,v(t,x),\sigma(t,x)' D_x v(t,x)) = 0 \\ &v(T,x) = g(x), \end{aligned} \tag{2.54}$$

for all $t \in [0,T]$ and $x \in \mathbb{R}^n$, and where $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is a non-linear function.

We will associate the solution to this problem with a forward-backward stochastic differential equation of the form (2.42). We have an easy association given by the following theorem

Theorem 2.2.3 (Verification theorem [14])**.** *Let $v(t,x)$ be a classical solution to 2.54, that is continuous on $[0,T] \times \mathbb{R}^n$, one time differentiable in t and two times differentiable in x and satisfy the linear growth condition $|v(t,x)| \leq L(1+|x|)$ for some $L > 0$ and all $x \in \mathbb{R}^n$ $t \in [0,T]$. Also, let its first space derivative satisfy the growth condition $|D_x v(t,x)| \leq C(1+|x|^q)$ for some $C > 0$, $q > 0$ and all $x \in \mathbb{R}^n$ $t \in [0,T]$.Then, the pair (Y,Z) defined by*

$$Y_t = v(t,X_t) \quad Z_t = \sigma(t,X_t)' D_x v(t,X_t) \tag{2.55}$$

is the solution to the backward stochastic differential equation in 2.42.

Proof. Apply Ito's formula to $Y_t = v(t,X_t)$ to obtain

$$\begin{aligned} dY_t &= \left(\frac{\partial v}{\partial t}(t,X_t) + \mathcal{L}_t v(t,X_t) \right) dt + \sigma(t,X_t)' D_x v(t,X_t) \cdot dW_t \\ &= -f(t,X_t,v(t,X_t,D_x v(t,X_t)))dt + \sigma(t,X_t)' D_x v(t,X_t) \cdot dW_t \\ &= -f(t,X_t,Y_t,Z_t)dt + Z_t dW_t \end{aligned} \tag{2.56}$$

and observing that $Y_T = v(T,X_T) = g(X_T)$, we have the first part, as this process is in $\mathbb{S}^2(0,T) \times \mathbb{H}^2(0,T)^d$ due to the growth condition of v and $D_x v$. ∎

Nevertheless, the reciprocal affirmation y somewhat more complicated. If we have a solution (Y_t, Z_t) to the FBSDE (2.42), not necessarily $v(t,x) = Y_t^{t,x}$ will be a classical smooth solution to (2.54) as it may not exist due to the non-linearity. Nevertheless, we can define a new weaker notion of solution as follows

Definition 2.2.2. *Let $v(t,x) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ be a locally bounded continuous function. Then*

- *$v(t,x)$ is called a viscosity sub-solution of (2.54) if $v(T,x) \leq g(x)$ for $x \in \mathbb{R}^n$, and for all $\phi \in C^{1,2}([0,T] \times \mathbb{R}^n)$ such that the map $v(t,x) - \phi(t,x)$ attains a local maximum at $(t,x) \in [0,T] \times \mathbb{R}^n$ it holds*

$$\frac{\partial \phi}{\partial t} + \mathcal{L}_t \phi + f(t, x, v(x,t), \sigma(t,x)' D_x v(t,x)) \geq 0 \tag{2.57}$$

- *$v(t,x)$ is called a viscosity super-solution of 2.54 if $v(T,x) \geq g(x)$ for $x \in \mathbb{R}^n$, and for all $\phi \in C^{1,2}([0,T] \times \mathbb{R}^n)$ such that the map $v(t,x) - \phi(t,x)$ attains a local minimum at $(t,x) \in [0,T] \times \mathbb{R}^n$ it holds*

$$\frac{\partial \phi}{\partial t} + \mathcal{L}_t \phi + f(t, x, v(x,t), \sigma(t,x)' D_x v(t,x)) \leq 0 \tag{2.58}$$

- *If $v(t,x)$ is a sub-solution and a super-solution it is called a viscosity solution of (2.54).*

In particular, note that this definition does not require the smoothness of $v(t,x)$.

With this new concept of solution, we can establish the reverse relation as follows

Theorem 2.2.4 (Representation theorem [14])**.** *(Poner explicitas las condiciones para la FBSDE) Let (X, Y, Z) be the solution to the uncoupled FBSDE (2.42) and set $v(t,x) = Y_t^{t,x}$. Then, v is a continuous function and is a viscosity solution to 2.54.*

Terminar prueba

Proof. Step 1: Continuity of $v(t,x)$:
Let $(t_1, x_1), (t_2, x_2) \in [0,T] \times \mathbb{R}^n$, with $t_1 \leq t_2$. For lighten the notation we write $X_s^i = X_s^{t_i, x_i}$, $i = 1, 2$.

$$\begin{aligned}
\left|Y_t^1 - Y_t^2\right|^2 = & \left|g\left(X_T^1\right) - g\left(X_T^2\right)\right|^2 - \int_t^T \left|Z_s^1 - Z_s^2\right|^2 ds \\
& + 2 \int_t^T \left(Y_s^1 - Y_s^2\right) \cdot \left(f\left(s, X_s^1, Y_s^1, Z_s^1\right) - f\left(s, X_s^2, Y_s^2, Z_s^2\right)\right) ds \\
& - 2 \int_t^T \left(Y_s^1 - Y_s^2\right)' \left(Z_s^1 - Z_s^2\right) dW_s
\end{aligned} \tag{2.59}$$

Then,

$$
\begin{aligned}
&E\left[\left|Y_t^1-Y_t^2\right|^2\right]+E\left[\int_t^T\left|Z_s^1-Z_s^2\right|^2ds\right]\\
=&E\left[\left|g\left(X_T^1\right)-g\left(X_T^2\right)\right|^2\right]\\
&+2E\left[\int_t^T\left(Y_s^1-Y_s^2\right)\cdot\left(f\left(s,X_s^1,Y_s^1,Z_s^1\right)-f\left(s,X_s^2,Y_s^2,Z_s^2\right)\right)ds\right]\\
\leq&E\left[\left|g\left(X_T^1\right)-g\left(X_T^2\right)\right|^2\right]\\
&+2E\left[\int_t^T\left|Y_s^1-Y_s^2\right|\left|f\left(s,X_s^1,Y_s^1,Z_s^1\right)-f\left(s,X_s^2,Y_s^1,Z_s^1\right)\right|ds\right]\\
&+2C_fE\left[\int_t^T\left|Y_s^1-Y_s^2\right|\left(\left|Y_s^1-Y_s^2\right|+\left|Z_s^1-Z_s^2\right|\right)ds\right]\\
\leq&E\left[\left|g\left(X_T^1\right)-g\left(X_T^2\right)\right|^2\right]\\
&+E\left[\int_t^T\left|f\left(s,X_s^1,Y_s^1,Z_s^1\right)-f\left(s,X_s^2,Y_s^1,Z_s^1\right)\right|^2ds\right]\\
&+\left(1+4C_f^2\right)E\left[\int_t^T\left|Y_s^1-Y_s^2\right|^2ds+\frac{1}{2}E\int_t^T\left|Z_s^1-Z_s^2\right|^2ds\right],
\end{aligned}
\tag{2.60}
$$

So,

$$
\begin{aligned}
E\left[\left|Y_t^1-Y_t^2\right|^2\right]\leq&E\left[\left|g\left(X_T^1\right)-g\left(X_T^2\right)\right|^2\right]+E\left[\int_t^T\left|f\left(s,X_s^1,Y_s^1,Z_s^1\right)-f\left(s,X_s^2,Y_s^1,Z_s^1\right)\right|^2ds\right]\\
&+\left(1+4C_f^2\right)E\left[\int_t^T\left|Y_s^1-Y_s^2\right|^2ds\right]
\end{aligned}
\tag{2.61}
$$

and, by Gronwall's lemma,

$$
\begin{aligned}
E\left[\left|Y_t^1-Y_t^2\right|^2\right]\leq C\Big\{&E\left[\left|g\left(X_T^1\right)-g\left(X_T^2\right)\right|^2\right]\\
&+E\left[\int_t^T\left|f\left(s,X_s^1,Y_s^1,Z_s^1\right)-f\left(s,X_s^2,Y_s^1,Z_s^1\right)\right|^2ds\right]\Big\}.
\end{aligned}
\tag{2.62}
$$

Step 2: $v(t,x)$ is a viscosity solution.
Let's prove it is a super-solution. Assume by contradiction that

$$
-\frac{\partial\varphi}{\partial t}(t,x)-\mathcal{L}\varphi(t,x)-f\left(t,x,v(t,x),\left(D_x\varphi\right)'(t,x)\sigma(x)\right)>0.
\tag{2.63}
$$

By continuity of f,φ and its derivatives, there exists $h,\varepsilon>0$ such that for all $t\leq s\leq t+h,|x-y|\leq\varepsilon$,

$$
\begin{aligned}
&v(s,y)\leq\varphi(s,y)\\
&-\frac{\partial\varphi}{\partial t}(s,y)-\mathcal{L}\varphi(s,y)-f\left(s,y,v(s,y),\left(D_x\varphi\right)'(s,y)\sigma(y)\right)>0.
\end{aligned}
$$

Let $\tau = \inf\left\{s \geq t : \left|X_s^{t,x} - x\right| \geq \varepsilon\right\} \wedge (t+h)$, and consider the pair

$$\left(Y_s^1, Z_s^1\right) = \left(Y_{s\wedge\tau}^{t,x}, 1_{[0,\tau]}(s) Z_s^{t,x}\right), \quad t \leq s \leq t+h$$

By construction, $\left(Y_s^1, Z_s^1\right)$ solves the BSDE

$$\begin{aligned} -dY_s^1 &= 1_{[0,\tau]}(s) f\left(s, X_s^{t,x}, u\left(s, X_s^{t,x}\right), Z_s^1\right) ds - Z_s^1 dW_s, \quad t \leq s \leq t+h, \\ Y_{t+h}^1 &= u\left(\tau, X_\tau^{t,x}\right) \end{aligned}$$

On the other hand, the pair

$$\left(Y_s^2, Z_s^2\right) = \left(\varphi\left(s, X_{s\wedge\tau}^{t,x}\right), 1_{[0,\tau]}(s) D_x\varphi\left(s, X_s^{t,x}\right)' \sigma\left(X_s^{t,x}\right)\right), \quad t \leq s \leq t+h$$

satisfies, by Itô's formula, the BSDE

$$\begin{aligned} -dY_s^2 &= -1_{[0,\tau]}(s)\left(\frac{\partial\varphi}{\partial t} + \mathcal{L}\varphi\right)\left(s, X_s^{t,x}\right) - Z_s^2 dW_s, \quad t \leq s \leq t+h, \\ Y_{t+h}^1 &= \varphi\left(\tau, X_\tau^{t,x}\right). \end{aligned}$$

From the inequalities (6.18)-(6.19), and the strict comparison principle in Theorem 6.2.2, we deduce $Y_0^1 < Y_0^2$, i.e. $u(t,x) < \varphi(t,x)$, a contradiction. ■

Finally, note that this representation can be restated in a more similar form to the linear Feynman-Kac formula as

Theorem 2.2.5. *Under assumptions the same assumptions as 2.2.4, the function defined by*

$$v(t,x) := Y_t^{t,x} = \mathbb{E}\left[g\left(X_T^{t,x}\right) + \int_t^T f\left(s, X_s^{t,x}, Y_s^{t,x}, Z_s^{t,x}\right) ds\right], \tag{2.64}$$

where (X_t, Y_T, Z_T) is the solution to the FBSE 2.42 restricted to $[t,T]$, is a viscosity solution to the parabolic PDE (2.54).

Chapter 3

Deep Learning Methods for PDEs

Partial differential equations (PDE's) are ubiquitous among the tools for modeling complex phenomena in all sciences. However, we almost never have explicit solutions for them, making it difficult to describe those phenomenons and make accurate predictions about them. Hence, we need numerical methods to provide approximate solutions to those equations, for example, classical methods are finite differences, finite elements and spectral methods. Those rely on different discretizations of the particular problem that we can use for calculating approximations in different forms and with varying levels of accuracy. Since the advent of fast computers and efficient tools for programming them, this process is effective for many kinds of problems.

Now, when we attempt to solve numerically some particular problem, we need to play with the trade-off between accuracy of the approximate solution and the computational cost needed to obtain it. Indeed, with those classical methods, a small approximation error requires a finer grid, which implies more computational resources to store and process the information required by the method. In consequence, for some problems, we may not be able to calculate an accurate enough solution in a feasible computational time.

This is the case for high dimensional PDE's, for which the size of discretization usually scales exponentially with the number of points used for each dimension. For example, if we try to use a finite difference scheme in a 100-dimensional unit square $[0,1]^{100}$ with N points in each dimension, we would need N^{100} points in total, making it impossible to even store them in a computer. In practice, high dimension can be considered as low as $d > 4$, for which traditional methods cannot be used as regularly. This problem is known as the *curse of dimensionality*, a term established by Bellman when considering problems in dynamic programming.

High dimensional PDE's appear in many contexts, such as asset pricing, image denoising, statistical physics, many-body quantum mechanics, optimal control and game theory. Therefore, there is a necessity for numerical methods that are able to overcome this difficulty. Early attempts to solve this kind of problems used the connection between stochastic diffusions and parabolic PDE's, as we seemed in the preceding chapter. In fact, if the

PDE is linear, the linear Feynman-Kac 2.2.2 formula can be used to provide an approximate solution by computing the expectation using simulated paths of the process through the Monte-Carlo approach. The convergence of this formulation is independent(Sure?) of the dimension of the underlying process, and therefore does not suffer from the curse dimensionality.

Nevertheless, if we try a similar approach using the non-linear Feynman-Kac formula 2.2.5 for more general non-linear equations, we would have to deal with solving numerically the associated BSDE. There are numerical methods to approximate the set of solution processes (X, Y, Z), but they are not as simple as an Euler-Maruyama discretization for a forward process. Generally, they require the computation of conditional expectations that almost never are computationally cheap and hence is not a straightforward generalization of the former linear approach. Despite this, some progress has been made under this formulation, see for example [17]. Other solutions methods are based in fixed point iterations and branching methods [**bibid**].

Representing functions in a high dimensional space is a problem encountered in many other areas of applied mathematics. Particularly, in recent times, the analysis and inference on big amounts of data has emerged as the fascinating research area of *machine learning*. Many methods have been proposed for this goal, for example, regression methods, support vector machines and tree methods. Nonetheless, the approach that has encountered more success when trying to approximate high dimensional functions using big amounts of data is deep learning. In this setting, we parametrize functions using structures that use composition of simpler function for approximate complex ones, these structures are called neural networks. We refer the reader to Appendix A for a brief introduction and to [1] for a deeper exposition of the topic.

The idea of using this neural network parametrization of functions to solve PDE's can be tracked to the 80's, when in [**empty citation**] a perceptron layer approximation was proposed to (Completar). However, due to the high computational cost of training a neural network, a successful attempt was not achieved until recently, with the works of [**bibid**](Blablabla).

This is a very new area of research, for which many open questions remain. Particularly, it is not well understood yet if the curse of dimensionality is solved, even if there is work for certain equations that ensures it [**bibid**]. Also, there is not yet a good understanding of why different classes of neural networks are useful to approximate certain classes of functions and how to tune adequately its parameters to do it efficiently. In consequence, even if it is possible to give a convergence proof for certain cases, most algorithms rely on empirical experimentation and heuristic arguments to provide reasonable approximate solutions.

In this chapter we review some of these methods, implement them for toy examples and perform a comparison of speed, accuracy and practical usefulness for solving PDE's.

3.1 Unbounded problems

Let's start with problems in free space. In the same setup as Theorem 2.2.3, we deal with the following equation with terminal condition

$$\begin{aligned} &\frac{\partial v}{\partial t}(t,x) + \mu(t,x) \cdot D_x v(t,x) + \frac{1}{2} \operatorname{Tr}(\sigma(t,x)\sigma^T(t,x) D^2_{xx} v(t,x)) \\ &+ f(t,x,v(t,x),\sigma(t,x)' D_x v(t,x)) = 0 \\ &v(T,x) = g(x). \end{aligned} \tag{3.1}$$

We have proven that we can construct a viscosity solution to this equation by setting $v(t,x) = Y_t^{t,x}$, where Y is the solution process to the FBSDE

$$\begin{aligned} &dX_s = \mu(s,X_s)ds + \sigma(s,X_s)dW_s \\ &X_t = x, \\ &dY_s = -f(s,X_s,Y_s,Z_s)ds + Z_s dW_s \\ &Y_T = g(X_T). \end{aligned} \tag{3.2}$$

Moreover, we have that $Y_t = v(t,X_t)$ and $Z_t = \sigma(t,X_t)' D_x v(t,X_t)$.

3.1.1 Deep BSDE method

The first deep learning algorithm that was successfully applied to solve equation (3.1) was proposed by Han, E and Jentzen [18, 19]. This algorithm aims to approximate $Y_0 = v(0,x)$ for some point $x \in \mathbb{R}^n$, and is similar in spirit to the stochastic shooting method for ODE's.

Here we discretize the time domain $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$ and the FBSDE system a forward equation using the Euler-Maruyama scheme for $n = 0,\ldots,N-1$,

$$X_{t_{n+1}} \approx X_{t_n} + \mu\left(t_n, X_{t_n}\right)\Delta t_n + \sigma\left(t_n, X_{t_n}\right)\Delta W_n \tag{3.3}$$

and

$$\begin{aligned} v\left(t_{n+1}, X_{t_{n+1}}\right) \approx\ & v\left(t_n, X_{t_n}\right) - f\left(t_n, X_{t_n}, v\left(t_n, X_{t_n}\right), \sigma'\left(t_n, X_{t_n}\right) D_x v\left(t_n, X_{t_n}\right)\right)\Delta t_n \\ &+ \sigma\left(t_n, X_{t_n}\right)' D_x u\left(t_n, X_{t_n}\right)\Delta W_n, \end{aligned} \tag{3.4}$$

where $\Delta t_n = t_{n+1} - t_n$ and $\Delta W_n \sim \mathcal{N}(0, \Delta t_n)$.

The main idea of this algorithm is to transform the problem in a learning one approximating the unknown product $\sigma\left(t_n, X_{t_n}\right)' D_x u\left(t_n, X_{t_n}\right)$ with a fully coupled neural network for each time step, i.e

$$\sigma\left(t_n, X_{t_n}\right)' D_x u\left(t_n, X_{t_n}\right) \approx \mathcal{Z}_n(X_{t_n}|\theta_n), \tag{3.5}$$

where θ_n denotes the parameters of the neural network at time t_n. Each one of these networks receives as inputs the simulated paths (3.3) and therefore its input layers have d neurons. Furthermore, the desired solution $v(0,x)$ and its derivative $D_x v(0,x)$ also will be parameters to be learned in the model, it means $v(0,x) \approx \theta_{v_0}$ and $D_x v(0,x) \approx \theta_{D_x v_0}$. Thus, the total set of parameters to be optimized is

$$\theta = \{\theta_{v_0}, \theta_{D_x v_0}, \theta_1, \theta_2, \cdots, \theta_n\}. \tag{3.6}$$

If we need the solution $v(0,x)$ for all x in some region Ω, we can choose to parametrize $v(0,x) \approx \theta_{v_0}$ with a neural network and simulate the process X_t with random initial conditions in Ω.

The set of parameters will be optimized such that the stacked solution $\hat{u}(\{X_{t_n}\}_0^N, \{W_{t_n}\}_0^N)$, constructed with (3.4), resembles the terminal condition $g(X_{t_N})$. This is achieved defining the loss function

$$\ell(\theta) = \mathbb{E}\big[|g(X_{t_N}) - \hat{u}(\{X_{t_n}\}_0^N, \{W_{t_n}\}_0^N)|^2\big], \tag{3.7}$$

which will be minimized using deep learning standard methods for training, for example, the ADAM optimizer.

The overall method can be thought as a constrained minimization problem of the form

$$\begin{aligned} &\inf_{\theta} \hat{u}(\{X_{t_n}\}_0^N, \{W_{t_n}\}_0^N) \\ s.t \quad & X_0 = \xi, \quad Y_0 = \theta_{v_0} \\ & X_{t_{n+1}} = X_{t_n} + \mu\left(t_n, X_{t_n}\right)\Delta t_n + \sigma\left(t_n, X_{t_n}\right)\Delta W_n \\ & Z_{t_n} = \mathcal{Z}_n(X_{t_n}) \\ & Y_{t_{n+1}} = Y_{t_n} - f(t_n, X_{t_n}, Y_{t_n}, Z_{t_n})\Delta t + Z_{t_n}' \Delta W_n \end{aligned} \tag{3.8}$$

where ξ is random variable uniformly distributed on Ω. And it can be summarized in the diagram shown in figure 3.1.

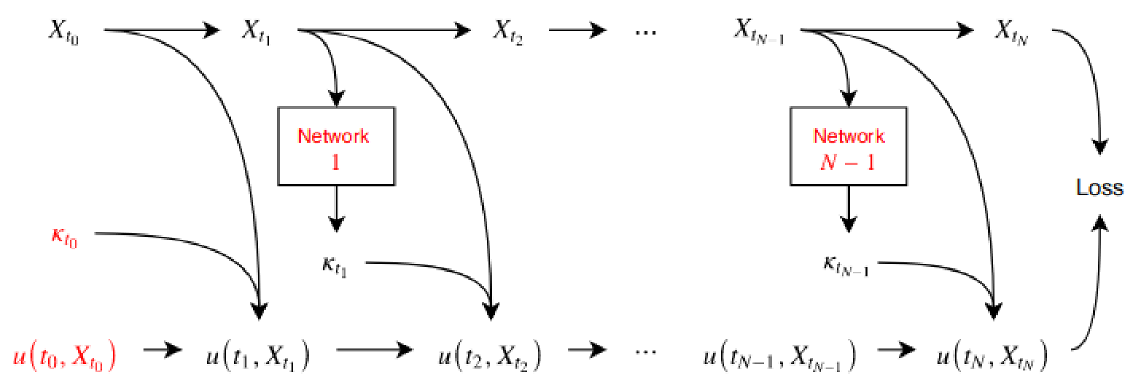

Figure 3.1: Deep BSDE diagram [20]. In red are the parameters to be optimized in the algorithm (Hay que cambiar κ_i)

The convergence of this algorithm has been proved in [21] for fully coupled FBSDEs. The assumptions needed are very general, and the proof is rather technical, so for the sake of brevity, we only state some imprecise results.

Denote by X_t, Y_t, Z_t the exact solution to the FBSDE (3.2), and by $X_{t_i}^{\pi}, Y_{t_i}^{\pi}, Y_{t_i}^{\pi}$ the discrete solution to the constrained optimization problem (3.8). Also, let's denote $h = \max_i \Delta t_i$. The first result states that the simulation error can be bounded through the value of the loss function (3.7).

Theorem 3.1.1 (Error of discretization is bounded by loss function [21]). *Under some assumptions, there exist a constant C, independent of h, d and n, such that for sufficiently small h*

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

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

The second result establishes that the optimal value of the loss function can be small if the approximation capability of the family of parametric functions (neural networks) is good enough. Denote by $\mathcal{N}_0'$ and $\{\mathcal{N}_i\}_0^{N-1}$ the parametric function spaces generated by neural networks, then we have

Theorem 3.1.2 (Optimal loss function is bounded by approximation error [21]). *Under some assumptions, there exists a constant C, independent of $\Delta t, d$ and n, such that for sufficiently small Δt,*

$$\begin{aligned}&\inf_{\mu_0^\pi\in\mathcal{N}_0',\phi_i^\pi\in\mathcal{N}_i}E\left|g\left(X_T^\pi\right)-Y_T^\pi\right|^2\\&\leq C\left\{h+\inf_{\mu_0^\pi\in\mathcal{N}_0',\phi_i^\pi\in\mathcal{N}_i}\left[E\left|Y_0-\mu_0^\pi(\xi)\right|^2\right.\right.\\&\left.\left.+\sum_{i=0}^{N-1}E\left|E\left[\tilde{Z}_{t_i}\mid X_{t_i}^\pi,Y_{t_i}^\pi\right]-\phi_i^\pi\left(X_{t_i}^\pi,Y_{t_i}^\pi\right)\right|^2h\right]\right\},\end{aligned}$$

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

Neural networks are a promising candidate for such approximation space of functions, as there are results in regard to the universal approximation and complexity of neural networks(Completar referencias).

Note that, in practice, we cannot minimize exactly the loss function in the space of parametric functions, as the methods we use are generally iterative. Also, it is not known how better is the approximation depending on the width, deep and connections of the neural network, so the capability of approximation is not well understood yet. Therefore, this results only establish the convergence of the method in a general setting and are not useful for estimate the real velocity of convergence, the achievable loss in the training stage, nor the real accuracy of the approximate solution.

Now, we highlight the major drawbacks of this method

1. The number of neural networks to train grows linearly with the number of time steps in the discretization, making it very computationally costly to use small time steps.

2. We only have a full solution at time $t=0$. At intermediate times we only have approximate solutions evaluated on sample paths $v(t,X_t)$. Therefore, we would need many of them to represent accurately the solution in the desired region.

3. Moreover, nothing guarantees that this intermediate steps resembles accurately the real solutions in between steps. This requirement is not well encoded in the loss function.

4. The time structure of the problem is not reflected in the separate approximations for each step, at least not directly.

5. It may be unstable or converge to saddle points, see [22].

Moreover, the structure of the neural networks is crucial for the practical convergence of the algorithm. The original work by [18] used fully coupled neural networks with 3 layers, the *relu* activation function and batch normalization, this structure is represented in figure 3.2. However, many changes can be made to accelerate the process of training or to achieve lower optimal losses. All of them are inspired by practical evidence and currently there is not enough understanding of how to choose theoretically the hyperparameters to reach the best convergence we can. Some of these modifications are explained below.

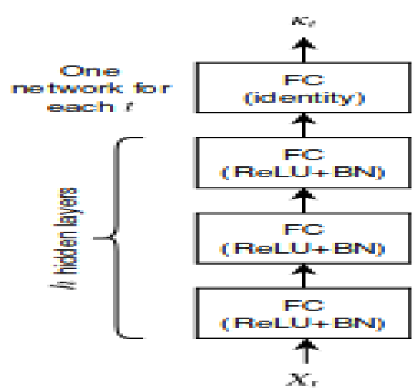

Figure 3.2: Neural network structure [20].(Hay que cambiar κ_i Mejor hago la mia :()

Merged Deep BSDE

The first such modification was proposed in [20]. In this work, the authors propose to use a single neural network to approximate the Z process for all time steps. Thus, this reduced model has less parameters to optimize, making this process faster, and also adds regularity to the computed gradients, which means that close in time parametrizations should be close for a given x.

Moreover, it was noted that using all information available at time t, like Y_t and $g(X_t)$, increases the performance and optimal loss. Thus, in order to merge all neural networks in a single one, we have to add new dimensions to the neural network to include the additional variables, hence we use a neural network $\mathcal{N}^\theta : (t, x) \in \mathbb{R}^{n+3} \to \mathbb{R}^n$ that uses $(t_n, Y_{t_n}, g(X_{t_n}), X_{t_n})$ as inputs to approximate Z_{t_n} for each t_n in the discretization. Note that Z_0 is also obtained through such network and therefore is not longer a parameter to be optimized. This process is summarized in the figure 3.3.

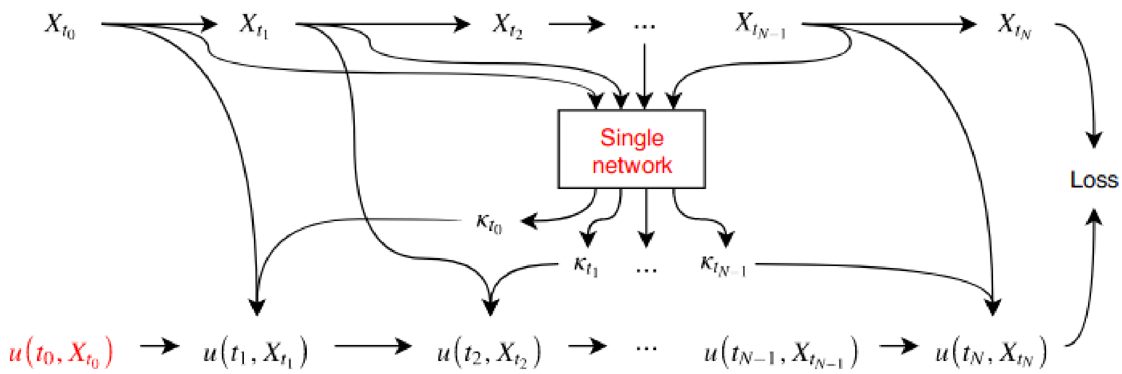

Figure 3.3: Merged Deep BSDE diagram [20]. In red are the parameters to be optimized in the algorithm (Hay que cambiar κ_i)

The structure of the neural network needs to be adapted to this new setup. In this case, the activation functions may be changed for the ELU function and batch normalization was not performed after each layer as the process in no longer stationary. This structure is represented in figure 3.4.

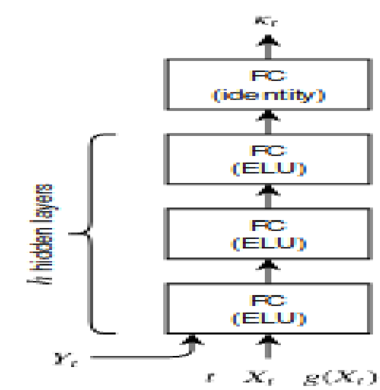

Figure 3.4: Neural network structure for merged BSDE [20].(Hay que cambiar κ_i Mejor hago la mia :()

Residual Merged Deep BSDE

Finally, another useful modification to the merged deep BSDE scheme was to modify the network structure adding shortcut connections between layers. This new configuration is represented in figure 3.5.

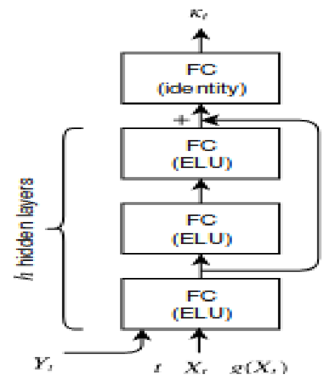

Figure 3.5: Neural network structure for merged BSDE [20].(Hay que cambiar κ_i Mejor hago la mia :()

Raissi's method

To circumvent some of the problems encountered with the deep BSDE method and last variants, Raissi [23] proposed a new scheme based on the same stochastic formulation. In this approach, the solution $v(t,x)$ is directly approximated with a neural network $\sqcup, \S$ that takes as inputs x and t instead of its gradient as before.

Now, the constraints on problem (3.8) are relaxed and they are enforced weakly through the loss function given by

$$\ell(\theta) := \mathbb{E}\left[\sum_{i=0}^{N-1} \Phi\left(t_i, X_{t_i}, Y_{t_i}, Y_{t_{i+1}}, \Delta W_{t_i}\right) + \left(g\left(X_{t_N}\right) - Y_{t_N}\right)^2\right] \tag{3.9}$$

where

$$\begin{aligned}\Phi\left(t_i, X_{t_i}, Y_{t_i}, Y_{t_{i+1}}, \Delta W_{t_i}\right) = \Big(Y_{t_{i+1}} - Y_{t_i} + f\left(t_i, X_{t_i}, Y_{t_i}, \sigma'\left(t_i, X_{t_i}\right)\widehat{Z}_{t_i}\right)(\Delta t_i) \\ -\widehat{Z}'_{t_i}\sigma\left(t_i, X_{t_i}\right)\left(\Delta W_{t_i}\right)\Big)^2,\end{aligned} \tag{3.10}$$

and $\widehat{Z}_{t_i}$ is calculated with automatic differentiation in the neural network $\widehat{Z}_{t_i} = \hat{D}\mathcal{N}(t_i, X_{t_i})$. With this approach we can compute the solution at all times t and point in space x.

3.1.2 An example

Let's test these algorithms with a toy problem from control theory using the Hamilton-Jacobi-Bellman equation. For completeness, we briefly review the standard formulation of these problems in Appendix B.

We will perform a comparison between the preceding algorithms by solving, just for fun, the Hamilton-Jacobi-Bellman equation associated to a linear-quadratic regulator in dimension 100.

3.2 Bounded problems

In this section, we consider PDE's with boundary conditions of the form

$$\frac{\partial v}{\partial t}(t,x) + \mathcal{L}v(t,x) + f(t,x,v(t,x),\sigma(t,x)' D_x v(t,x)) = 0 \tag{3.11a}$$

with terminal condition

$$v(T,x) = g(x) \quad \forall x \in \Omega, \tag{3.11b}$$

and Dirichlet and Neumman conditions

$$v(t,x) = h_d(x) \quad \forall x \in \Gamma_D \text{ and } \forall t \in [0,T] \tag{3.11c}$$

$$\frac{\partial v}{\partial n}(t,x) = h_n(x) \quad \forall x \in \Gamma_N \text{ and } \forall t \in [0,T], \tag{3.11d}$$

where Ω is a bounded domain, regular enough for a solution to exist, and $\Gamma_D \dot{\cup} \Gamma_N = \partial\Omega$.

If, in addition to the PDE, we impose Neumman/Dirichlet boundary conditions, the previous approaches must be modified because the Feynman-Kac formula does not account for them explicitly. We will try two different approaches.

3.2.1 Deep Galerkin method

The first approach is named the deep Galerkin method (DGM) and was proposed in [3]. This method is also called Physics Informed Neural Networks (PINN's). It does not rely on a probabilistic representation of the solution, but on the explicit form (3.11).

The DGM algorithm approximates the solution $v(t,x)$ with a deep neural network $\varphi(t,x|\theta)$, where θ are the network's parameters. We will choose this function aiming to minimize the cost function given by

$$\begin{aligned} J(\varphi) =& \alpha_{int} \left\| \frac{\partial \varphi}{\partial t} + \mathcal{L}\varphi + f(t,x,\varphi(t,x),\sigma(t,x)' D_x \varphi(t,x)) \right\|^2_{[0,T]\times\Omega,\nu_1} \\ &+ \alpha_T \left\| \varphi(T,x|\theta) - g(x) \right\|^2_{\Omega,\nu_2} \\ &+ \alpha_d \left\| \varphi(t,x) - h_d(x) \right\|^2_{[0,T]\times\partial\Gamma_D,\nu_3} \\ &+ \alpha_n \left\| \hat{n}\cdot D_x \varphi(t,x) - h_n(x) \right\|^2_{[0,T]\times\partial\Gamma_N,\nu_4} \end{aligned} , \tag{3.12}$$

where we use the notation for the norms $\|f(y)\|^2_{\mathcal{Y},\nu} = \int_{\mathcal{Y}} |f(y)|^2 d\nu$, given a positive probability density ν on $\mathcal{Y}$. Here, $J(\varphi)$ measure how well ϕ satisfies the PDE with its initial and boundary conditions weighted by the α coefficients. If $J(\varphi) = 0$, then φ is a solution to (3.11).

To obtain the minimizing parameters θ, we will perform a training procedure of the neural network with a stochastic gradient descent algorithm. We will sample points according to $\nu_1, \nu_2, \nu_3, \nu_4$ and calculate a Monte-Carlo approximation of the functional loss (3.12) using automatic differentiation to calculate the derivatives appearing in the PDE operator and boundary conditions. With this approximation, we will take a descent step in the gradient direction calculated again using automatic differentiation.

Due to the universal approximation theorem for neural networks and some more theoretical considerations, we can establish the convergence of this algorithm for some well-behaved partial differential equations, see [3]. However, the practical convergence presents difficulties as it depends strongly on a set of well-chosen hyper-parameters, such as the learning rate of the descent method, the batch sizes, the weights in the loss function, the network architecture, as well as the sampling densities. There is not a straightforward way to choose such parameters, thus the convergence is

Note also that there is a big operational cost involved in the computation of derivatives appearing in the loss function through automatic differentiation. For operators using fully coupled second order derivatives, this cost becomes prohibitive in high dimensions. Therefore, we need alternatives for problems with these restrictions.

3.2.2 Interpolating BSDEs with PINNs

To circumvent these issues, we may combine the powerful ideas of stochastic representation of solution of PDEs with the forcing technique used in the PINNs approach. This method was proposed by [24], where the authors define an interpolation parameter between both methods to

3.2.3 An example

Chapter 4

Crowd motion modeling

People run

4.1 N-agent games

4.2 Mean Field Games

4.3 Numerical methods

4.3.1 Finite differences

4.3.2 Deep Fictious Play

4.4 An example

Chapter 5

Conclusion

Appendix A

Neural Networks

-No se podia hasta recientemente por el poder computacional -Fully coupled -Residual -Batch normalization -Loss functions -Automatic differentiation -optimizacion -influencia parametros (profundidad y ancho) -densidad en L2

Appendix B

Stochastic Control

In this appendix we review, without proofs, the basics of stochastic optimal control leading to the Hamilton-Jacobi-Bellman equation used in this work, and give the linear-quadratic regulator as an example of this theory. We follow [14].

The Hamilton-Jacobi-Bellman Equation

Suppose that we want to control a process $X_t \in \mathbb{R}^n$ that satisfies a stochastic differential equation driven by d-dimensional Brownian motion of the form

$$\begin{aligned} dX_t &= \mu(X_t, \alpha_t)dt + \sigma(X_t, \alpha_t)dW_t \\ X_0 &= x, \end{aligned} \tag{B.1}$$

with a control function α_t taking values in some admissible space A. From now on we assume μ and σ satisfy the standard Lipschitz conditions required for a solution to this equation exist. We want to choose such control so that the total benefit functional given by

$$J(\alpha_t) = \mathbb{E}\left[\int_0^T f(s, X_s, \alpha_s)ds + g(X_T)\right]. \tag{B.2}$$

is maximum over all possible control functions. Here, the function f is called running cost and g is called terminal cost. At any time t, we can choose the controller using only information observed before t, as we are unable to foretell the future due to the system's randomness. Therefore, we require α_t to be $\mathcal{F}_t$-adapted and define the set of feasible controls as $\mathcal{A}([0,T]) = \{\alpha : [0,T] \times \Omega \to A\}$.arg2(Definir tipos de control morkoviano open,...)

A stochastic control problem consists on finding $\hat{\alpha} \in \mathcal{A}([0,T])$ such that

$$J(\hat{\alpha}) = \sup_{\alpha_t \in \mathcal{A}([0,T])} J(\alpha). \tag{B.3}$$

We need the following definitions. For all $(t,x) \in \mathbb{R}^+ \times \mathbb{R}^n$ and $\alpha \in \mathcal{A}([0,T])$, we denote by $X_s^{t,x,\alpha}$ the solution to the SDE

$$\begin{aligned} dX_s^{t,x,\alpha} &= \mu(X_s^{t,x,\alpha}, \alpha_s)ds + \sigma(X_s^{t,x,\alpha}, \alpha_s)dW_s \\ X_t^{t,x,\alpha} &= x. \end{aligned} \tag{B.4}$$

Now, we define the value functional starting at time t and position x as

$$J(t,x,\alpha) = \mathbb{E}\left[\int_t^T f(s, X_s^{t,x,\alpha}, \alpha_s)ds + g(X_T^{t,x,\alpha})\right]. \tag{B.5}$$

and the *value function* $V(t,x)$ as

$$V(t,x) = \sup_{\alpha \in \mathcal{A}[t,T]} J(t,x,\alpha), \tag{B.6}$$

which is the expected optimal reward starting the process at time t and point x.

To solve the stochastic control problem, we follow the approach based on the *dynamic programming principle*, which states informally that

> "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision" Richard Bellman

and can be translated in the following theorem

Theorem B.1 (Stochastic dynamic programming [14]). *For all $0 \le t \le s \le T$ and $x \in \mathbb{R}^n$ we have that the value function $V(t,x)$ satisfies (Revisar esto, la s no tiene sentido)*

$$V(t,x) = \sup_{\alpha \in \mathcal{A}[t,s]} \mathbb{E}\left[\int_t^s f(r, X_r^{t,x,\alpha}, \alpha_r)dr + V(s, V_s^{t,x,\alpha})\right], \tag{B.7}$$

from which a infinitesimal version can be derived, named the Hamilton-Jacobi-Bellman equation

$$\begin{aligned} &\frac{\partial V}{\partial t} + \sup_{a \in A}\{\mathcal{L}^a[V](t,x) + f(t,x,a)\} = 0 \\ &V(T,x) = g(x), \end{aligned} \tag{B.8}$$

where $\mathcal{L}$ is the infinitesimal generator of the controlled process X_t given by

$$\mathcal{L}^a[V](t,x) = \mu(x,a) \cdot D_x V(t,x) + \frac{1}{2}\operatorname{Tr}(\sigma(x,a)\sigma(x,a)' D_{xx}V(t,x)). \tag{B.9}$$

We can also write the Hamilton-Jacobi-Bellman equation as

$$\begin{aligned} &\frac{\partial V}{\partial t} + H(t,x,D_xV, D_{xx}V) = 0 \\ &V(T,x) = g(x), \end{aligned} \tag{B.10}$$

where the function $H(t,x,p,M)$ is the *hamiltonian* defined as

$$H(t,x,p,M) = \sup_{a\in A}\{\mu(x,a)\cdot p + \frac{1}{2}\operatorname{Tr}(\sigma\sigma'(x,a)M) + f(t,x,a)\}. \tag{B.11}$$

Note that we assume implicitly that the supremums appearing in these equations exists, but this condition is not necessary as pointed in [14].

Solving this equation the Hamilton-Jacobi-Bellman equation for the function $V(t,x)$ can be used to construct optimal controls for the original problem as will be shown below with the linear-quadratic regulator. However, we need a result stating that a solution to such equation is in fact the desired value function

Theorem B.2 (Verification theorem [14])**.** *(Revisar notacion) Let w be a function in $C^{1,2}([0,T)\times\mathbb{R}^n)\cap C^0([0,T]\times\mathbb{R}^n)$, and satisfying a quadratic growth condition, i.e. there exists a constant C such that*

$$|w(t,x)| \le C\left(1+|x|^2\right), \quad \forall (t,x)\in[0,T]\times\mathbb{R}^n$$

(i) Suppose that

$$-\frac{\partial w}{\partial t}(t,x) - \sup_{a\in A}\left[\mathcal{L}^a w(t,x) + f(t,x,a)\right] \ge 0, \quad (t,x)\in[0,T)\times\mathbb{R}^n,$$
$$w(T,x)\ge g(x), \quad x\in R^n.$$

Then $w\ge v$ on $[0,T]\times\mathbb{R}^n$.

(ii) Suppose further that $w(T,x)=g(x)$, and there exists a measurable function $\hat\alpha(t,x)$, $(t,x)\in[0,T)\times\mathbb{R}^n$, valued in A such that

$$\begin{aligned}-\frac{\partial w}{\partial t}(t,x) - \sup_{a\in A}\left[\mathcal{L}^a w(t,x) + f(t,x,a)\right] &= -\frac{\partial w}{\partial t}(t,x) - \mathcal{L}^{\hat\alpha(t,x)}w(t,x) - f(t,x,\hat\alpha(t,x))\\ &= 0\end{aligned}$$

the SDE

$$dX_s = \mu\left(X_s,\hat\alpha\left(s,X_s\right)\right)ds + \sigma\left(X_s,\hat\alpha\left(s,X_s\right)\right)dW_s$$

admits a unique solution, denoted by $\hat X_s^{t,x}$, given an initial condition $X_t=x$, and the process $\left\{\hat\alpha\left(s,\hat X_s^{t,x}\right) t\le s\le T\right\}$ lies in $\mathcal{A}(t,x)$. Then

$$w=v \quad \text{on } [0,T]\times\mathbb{R}^n,$$

and $\hat\alpha$ is an optimal Markovian control.

The linear-quadratic regulator (LQR)

To exemplify how to use the stochastic dynamic programming approach to solve optimal control problems we will solve the linear-quadratic regulator. This problem models a particle whose dynamics is described by the SDE

$$\begin{aligned} dX_t &= 2\sqrt{\lambda}\alpha_t dt + \sqrt{2}dW_t\\ X_0 &= x, \end{aligned} \tag{B.12}$$

where α_t is the control process and λ is a constant representing the strength of the control. We want to minimize the cost functional

$$J(\alpha_t) = \mathbb{E}\left[\int_0^T |\alpha_t|^2 dt + g(X_T)\right], \tag{B.13}$$

which models the cost of the particle to reach a desired state, whose distance is modeled by $g(x)$, using the least amount of fuel as possible, which is represented by α_t. For example, if we want the particle to reach the point z_0, we should choose $g(x) = |x - z_0|^2$.

To derive the Hamilton-Jacobi-Bellman satisfied by the value function associated with this problems, note that the generator of the X process is given by

$$\mathcal{L}^a V(t,x) = 2\sqrt{\lambda} a \cdot D_x V + \mathrm{Tr}(D^2_{xx} V) = 2\sqrt{\lambda} a \cdot D_x V + \Delta V, \tag{B.14}$$

hence the Hamiltonian for this problem is

$$H(t, x, D_x V, D^2_{xx} V) = \inf_{a \in A}\{2\sqrt{\lambda} a \cdot D_x V + \Delta V + |a|^2\}, \tag{B.15}$$

where we can calculate this inf analytically by taking the derivative with respect to a and equating to cero, as the function inside is convex in a. Therefore, the minimum is attained at $a = -\sqrt{\lambda} D_x V$ and the Hamiltonian is

$$H(t, x, D_x V, D^2_{xx} V) = -2\lambda |D_x V|^2 + \lambda |D_x V|^2 + \Delta V = \Delta V - \lambda |D_x V|^2. \tag{B.16}$$

Thus, the associated HJB equation is

$$\frac{\partial V}{\partial t} + \Delta V - \lambda |\nabla V|^2 = 0 \tag{B.17}$$

subject to the terminal condition

$$V(T, x) = g(x). \tag{B.18}$$

If we solve this equation for $V(t,x)$, then we can use the verification theorem to obtain the optimal control process as $\hat{\alpha} = D_x V(t,x)$.

Fortunately, we can solve this equation explicitly using the Hopf-Cole transformation, $u(t,x) = e^{-\lambda V(t,x)}$. For such $u(t,x)$ we have

$$\nabla u(t,x) = -\lambda e^{-\lambda V} \nabla V \tag{B.19}$$

and

$$\begin{aligned} \Delta u(t,x) &= \lambda^2 e^{-\lambda V} |\nabla V|^2 - \lambda e^{-\lambda V} \Delta V \\ &= -\lambda e^{-\lambda V}\left(\Delta V - \lambda |\nabla V|^2\right) \end{aligned} \tag{B.20}$$

and

$$\frac{\partial u}{\partial t}(t,x) = -\lambda e^{-\lambda V} \frac{\partial V}{\partial t}. \tag{B.21}$$

Therefore, the HJB equation for $u(t,x)$ is

$$\frac{\partial u}{\partial t} + \Delta u = 0 \tag{B.22}$$

subject to the final condition

$$u(T,x) = e^{-\lambda g(x)} \tag{B.23}$$

Hence, using the linear Feynman-Kac formula 2.2.2, we can give a probabilistic explicit representation of the solution as

$$u(t,x) = \mathbb{E}\Big[\exp(-\lambda g(x + \sqrt{2}W_{T-t}))\Big] \tag{B.24}$$

and solving for $V(t,x)$ we obtain

$$V(t,x) = -\frac{1}{\lambda}\ln\Big(\mathbb{E}\Big[\exp(-\lambda g(x + \sqrt{2}W_{T-t}))\Big]\Big)\,. \tag{B.25}$$

Bibliography

[1] Catherine F. Higham and Desmond J. Higham. "Deep Learning: An Introduction for Applied Mathematicians". In: *SIAM Review* 61.3 (Jan. 2019), pp. 860–891. ISSN: 0036-1445, 1095-7200. DOI: 10.1137/18M1165748. URL: https://epubs.siam.org/doi/10.1137/18M1165748 (visited on 09/26/2022).

[2] Zongyi Li et al. *Fourier Neural Operator for Parametric Partial Differential Equations*. May 16, 2021. arXiv: 2010.08895[cs,math]. URL: http://arxiv.org/abs/2010.08895 (visited on 10/10/2022).

[3] Justin Sirignano and Konstantinos Spiliopoulos. "DGM: A deep learning algorithm for solving partial differential equations". In: *Journal of Computational Physics* 375 (Dec. 2018), pp. 1339–1364. ISSN: 00219991. DOI: 10.1016/j.jcp.2018.08.029. URL: https://linkinghub.elsevier.com/retrieve/pii/S0021999118305527 (visited on 10/10/2022).

[4] Salvatore Cuomo et al. *Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What's next*. June 7, 2022. arXiv: 2201.05624[physics]. URL: http://arxiv.org/abs/2201.05624 (visited on 10/05/2022).

[5] Jan Blechschmidt and Oliver G. Ernst. "Three ways to solve partial differential equations with neural networks — A review". In: *GAMM-Mitteilungen* 44.2 (June 2021). ISSN: 0936-7195, 1522-2608. DOI: 10.1002/gamm.202100006. URL: https://onlinelibrary.wiley.com/doi/10.1002/gamm.202100006 (visited on 09/26/2022).

[6] Jean-Michel Bismut. "Conjugate convex functions in optimal stochastic control". In: *Journal of Mathematical Analysis and Applications* 44.2 (Nov. 1, 1973), pp. 384–404. ISSN: 0022-247X. DOI: 10.1016/0022-247X(73)90066-8. URL: https://www.sciencedirect.com/science/article/pii/0022247X73900668 (visited on 02/21/2023).

[7] E. Pardoux and S. G. Peng. "Adapted solution of a backward stochastic differential equation". In: *Systems & Control Letters* 14.1 (Jan. 1, 1990), pp. 55–61. ISSN: 0167-6911. DOI: 10.1016/0167-6911(90)90082-6. URL: https://www.sciencedirect.com/science/article/pii/0167691190900826 (visited on 02/21/2023).

[8] Jianfeng Zhang. *Backward Stochastic Differential Equations*. Vol. 86. Probability Theory and Stochastic Modelling. New York, NY: Springer New York, 2017. ISBN: 978-1-4939-7254-8 978-1-4939-7256-2. DOI: 10.1007/978-1-4939-7256-2. URL: http://link.springer.com/10.1007/978-1-4939-7256-2 (visited on 02/15/2023).

[9] Etienne Pardoux and Aurel R şcanu. *Stochastic Differential Equations, Backward SDEs, Partial Differential Equations*. Vol. 69. Stochastic Modelling and Applied Probability. Cham: Springer International Publishing, 2014. ISBN: 978-3-319-05713-2 978-3-319-05714-9. DOI: 10.1007/978-3-319-05714-9. URL: https://link.springer.com/10.1007/978-3-319-05714-9 (visited on 02/15/2023).

[10] Ricardo Romo Romero. "Maestro en ciencias con especialidad en probabilidad y estadística". In: ().

[11] Nizar Touzi. *Optimal Stochastic Control, Stochastic Target Problems, and Backward SDE*. Vol. 29. Fields Institute Monographs. New York, NY: Springer New York, 2013. ISBN: 978-1-4614-4285-1 978-1-4614-4286-8. DOI: 10.1007/978-1-4614-4286-8. URL: https://link.springer.com/10.1007/978-1-4614-4286-8 (visited on 02/15/2023).

[12] Xuerong Mao. *Stochastic differential equations and applications*. 2nd ed. OCLC: ocn176925635. Chichester: Horwood Pub, 2008. 422 pp. ISBN: 978-1-904275-34-3.

[13] Jared Chessari et al. *Numerical Methods for Backward Stochastic Differential Equations: A Survey*. Mar. 10, 2022. arXiv: 2101.08936[cs,math]. URL: http://arxiv.org/abs/2101.08936 (visited on 01/24/2023).

[14] Huyên Pham. *Continuous-time Stochastic Control and Optimization with Financial Applications*. Vol. 61. Stochastic Modelling and Applied Probability. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. ISBN: 978-3-540-89499-5 978-3-540-89500-8. DOI: 10.1007/978-3-540-89500-8. URL: https://link.springer.com/10.1007/978-3-540-89500-8 (visited on 02/28/2023).

[15] N. El Karoui, S. Peng, and M. C. Quenez. "Backward Stochastic Differential Equations in Finance". In: *Mathematical Finance* 7.1 (1997). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9965.00022, pp. 1–71. ISSN: 1467-9965. DOI: 10.1111/1467-9965.00022. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9965.00022 (visited on 02/27/2023).

[16] René Carmona. *Lectures on BSDEs, Stochastic Control, and Stochastic Differential Games with Financial Applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics, Feb. 18, 2016. ISBN: 978-1-61197-423-2 978-1-61197-424-9. DOI: 10.1137/1.9781611974249. URL: http://epubs.siam.org/doi/book/10.1137/1.9781611974249 (visited on 02/21/2023).

[17] Jared Chessari et al. "Numerical methods for backward stochastic differential equations: A survey". In: ().

[18] Jiequn Han, Arnulf Jentzen, and Weinan E. "Solving high-dimensional partial differential equations using deep learning". In: *Proceedings of the National Academy of Sciences* 115.34 (Aug. 21, 2018), pp. 8505–8510. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1718942115. URL: https://pnas.org/doi/full/10.1073/pnas.1718942115 (visited on 09/26/2022).

[19] Weinan E, Jiequn Han, and Arnulf Jentzen. "Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations". In: *Communications in Mathematics and Statistics* 5.4 (Dec. 1, 2017), pp. 349–380. ISSN: 2194-671X. DOI: 10.1007/s40304-017-0117-6. URL: https://doi.org/10.1007/s40304-017-0117-6 (visited on 04/10/2023).

[20] Quentin Chan-Wai-Nam, Joseph Mikael, and Xavier Warin. *Machine Learning for semi linear PDEs*. Dec. 10, 2018. arXiv: 1809.07609[cs,math,stat]. URL: http://arxiv.org/abs/1809.07609 (visited on 03/26/2023).

[21] Jiequn Han and Jihao Long. "Convergence of the Deep BSDE Method for Coupled FBSDEs". In: *Probability, Uncertainty and Quantitative Risk* 5.1 (Dec. 2020), p. 5. ISSN: 2367-0126. DOI: 10.1186/s41546-020-00047-w. arXiv: 1811.01165[cs,math]. URL: http://arxiv.org/abs/1811.01165 (visited on 04/10/2023).

[22] Côme Huré, Huyên Pham, and Xavier Warin. "Deep backward schemes for high-dimensional nonlinear PDEs". In: *Mathematics of Computation* 89.324 (Jan. 31, 2020), pp. 1547–1579. ISSN: 0025-5718, 1088-6842. DOI: 10.1090/mcom/3514. URL: https://www.ams.org/mcom/2020-89-324/S0025-5718-2020-03514-5/ (visited on 02/28/2023).

[23] Maziar Raissi. *Forward-Backward Stochastic Neural Networks: Deep Learning of High-dimensional Partial Differential Equations*. Apr. 19, 2018. arXiv: 1804.07010[cs,math,stat]. URL: http://arxiv.org/abs/1804.07010 (visited on 03/22/2023).

[24] Nikolas Nüsken and Lorenz Richter. *Interpolating between BSDEs and PINNs: deep learning for elliptic and parabolic boundary value problems*. Jan. 29, 2023. arXiv: 2112.03749[cs,math,stat]. URL: http://arxiv.org/abs/2112.03749 (visited on 03/26/2023).