UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

DESCRIPTION OF LOCAL AND NON-LOCAL EQUATIONS USING DEEP LEARNING TECHNIQUES

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MATEMÁTICO

JAVIER IGNACIO CASTRO MEDINA

PROFESOR GUÍA:
CLAUDIO MUÑOZ CERÓN

MIEMBROS DE LA COMISIÓN:
JOAQUIN FONTBONA TORRES
ARIS DANIILIDIS
MIRCEA PETRACHE
DUVAN HENAO MANRIQUE

Este trabajo ha sido parcialmente financiado por Fondecyt no. 1191412 and CMM Projects “Apoyo a Centros de Excelencia” ACE210010 and Fondo Basal FB210005.

SANTIAGO DE CHILE

2022

RESUMEN DE LA TESIS PARA OPTAR AL GRADO
DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA,
CON MENCIÓN MATEMÁTICAS APLICADAS
MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MATEMÁTICO
POR: JAVIER IGNACIO CASTRO MEDINA
FECHA: 2022
PROF. GUÍA: CLAUDIO MUÑOZ CERÓN

DESCRIPCIÓN DE ECUACIONES LOCALES Y NO-LOCALES USANDO TÉCNICAS DE DEEP LEARNING

En este trabajo se aborda la ecuación de Kolmororov mediante técnicas de aprendizaje profundo, en esencia se muestran dos resultados de interés independiente. En efecto, estudiamos la aplicación de las técnicas actuales de redes neuronales en la aproximación de soluciones EDPs no locales y en espacios de dimensión infinita. Con esto, generalizamos el trabajo de Hure, Pham y Warin en [HPW19] en dos direcciones particulares. Comenzamos introduciendo la ecuación de Kolmogorov lineal en $\mathbb{R}^d$ y su relación con las ecuaciones estocásticas. Destacamos la importancia de esta relación para el desarrollo de esquemas estocásticos para resolver ecuaciones diferenciales parciales (EDPs). Dado que nuestro marco es general, requerimos de las recientemente desarrolladas DeepOnets [LMK21] para describir en detalle el procedimiento de aproximación. Estos objetos actúan como una generalización de las Redes Neuronales a un contexto de dimensión infinita.

DESCRIPTION OF LOCAL AND NON-LOCAL EQUATIONS USING DEEP LEARNING TECHNIQUES

This work deals with the Kolmororov equation using deep learning techniques, in essentially two results of independent interest. Indeed, we study the application of current techniques of artificial neural networks to the approximation of solutions to the considered PDE in the non-local and the infinite dimensional settings. As a byproduct, we also generalize the work of Hure, Pham and Warin in [HPW19] in two particular directions. We start by introducing the linear Kolmogorov equation in $\mathbb{R}^d$ and its relation with random evolution equations. We remark the importance of this relation for the development of stochastic schemes to solve Partial Differential Equations (PDEs). Since our framework is general, we require the recently developed DeepOnets architectures [LMK21] to describe in detail the approximation procedure. These objects acts as a generalization of Neural Networks to an infinite-dimensional framework.

A mi madre y a mi padre.

Agradecimientos

Quiero agradecer a mi mamá y a mi papá por todos los sacrificios realizados para que yo, mi hermana y mi hermano pudieramos estudiar y desarrollarnos libremente como personas sin que nunca nos faltara nada. Fueron y son un ejemplo de perseverancia, siempre estaré en deuda por lo que han entregado para que yo pueda estar donde estoy. Quiero agradecer a mi hermana, la Berni, por siempre cuidarme, quererme y darme el apoyo necesario en todas las etapas de mi vida incluso estando a varios kilómetros de distancia. Quiero agradecer a mi hermano por ser siempre una figura ejemplar en mi vida, me doy por pagado si puedo ser la mitad del profesional y persona que eres. Esta tésis, que en cierto modo resumen 6 años y medio de carrera, no habría sido posible sin ustedes, soy el resultado de todos sus esfuerzos y enseñanzas. Les quiero mucho.

Quiero agradecer a mi otro hermano, el Waldo, por su voluntad a ayudarme en cualquier favor que necesitara. A la Paola, persona y profesional que admiro mucho. A mi amigo de la vida, Jano, gracias por estar cuando pensaba que el mundo se me acababa, gracias por quererme y quedarte en mi vida, sin tí este camino hubiera sido infinitamente más difícil.

A mis amigues del DIM. Benja, J y Caro, que me vienen siguiendo desde la inducción, les quiero mucho, sé que serán excelentes profesionales y nos seguiremos reencontrando en la vida. A la gente que conocí este año y se ganó un lugar muy importante en mi corazón; Manu, gracias por todas esas conversaciones sobre Matemáticas, me hiciste recuperar la motivación por la matraca y la investigación que había perdido durante la pandemia. Cynthia, fuiste un apoyo importante en esta parte final del proceso, gracias por el cariño y mostrarme esa vida universitaria que tanto quería experimentar. Willy, gracias por siempre apañar a un café y querer compartir conmigo, tu apoyo en los últimos momentos de la tésis fue importantísimo. Gonzalo, una persona de talla rápida, y Pedro mi otro compañero de zumba, gracias por las risas y la compañía, esos momentos de distención hacían más fáciles los días. Gracias también a la gente que me saludaba por los pasillos, sus saludos siempre me dejaban una sonrisa.

A mi profesor Guía, Claudio Muñoz, gracias por todas las oportunidades durante estos casi tres años trabajando juntos. Gracias por creer en mi y motivarme a postular a cuanto doctorado encontraba. A Jocelyn Dunstan y Joaquin Fontbona, muchas gracias por las infinitas cartas de recomendación. A los profesores Duvan, Mircea y Aris, gracias por su disposición y aceptar formar parte de la comisión.

A mis mascotas, Rocky, Morocha y gatito, gracias por ser muy suavecitos y dejarme hacerles cariño. Finalmente, a todas las personas que influyeron en este proceso, gracias por las experiencias vividas y lo que me entregaron.

Table of Content

Chapter 1

Introduction

1.1 The Kolmogorov equation

The main subject of study of this thesis is the representation via deep learning techniques of solutions of Kolmogorov Equations. Let $d \in \mathbb{N}$ with $d \geq 1$ and $T > 0$, a linear backward Kolmogorov equation can be written as

$$\begin{cases} \partial_t u(t,x) + \mathcal{L}[u](t,x) = 0, & (t,x) \in [0,T] \times \mathbb{R}^d, \\ u(T,x) = \phi(x), & x \in \mathbb{R}^d. \end{cases} \tag{1.1}$$

Here $u\colon [0,T] \times \mathbb{R}^d \to \mathbb{R}$ is the unknown of the problem, $\phi\colon \mathbb{R}^d \to \mathbb{R}$ is a terminal condition and ∇ represents the Fréchet derivative with respect to $x \in \mathbb{R}^d$. Finally, $\mathcal{L}$ is a second order parabolic operator defined for certain functions $f\colon [0,T] \times \mathbb{R}^d \to \mathbb{R}$. A. N. Kolmogorov introduced these equations in his foundational work [Kol38] by considering a stochastic model in which we do not know the exact state of a system but we rather know the probability that the system takes any of the possible states. This is, given a random evolution $(X_s^{t,x})_{s\in[t,T]}$ suitable to equation (1.1) and starting with $X_t^{t,x} = x$, it can be proved that the function $u(t,x) = \mathbb{E}\left(\phi(X_T^{t,x})\right)$ has the required regularity and is a solution of (1.1). The resulting theory allows to prove existence, uniqueness and properties of the solution of the parabolic equation by means of probabilistic ideas.

In this manuscript we are concerned with a slightly general framework which consist in adding a non-linear perturbation to equation (1.1); we consider a function $\psi\colon [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ and the target PDE transforms into

$$\begin{cases} \partial_t u(t,x) + \mathcal{L}[u](t,x) + \psi(t,x,u(t,x),\nabla u(t,x)) = 0, & (t,x) \in [0,T] \times \mathbb{R}^d, \\ u(T,x) = \phi(x), & x \in \mathbb{R}^d. \end{cases} \tag{1.2}$$

We shall also study the non-local version of this equation where the parabolic operator $\mathcal{L}$ additionally contains an integral operator as well as ψ having an extra integral dependence on u. The variable x will be taken in a Hilbert space in a second statement.

This thesis is mainly composed of two articles:

- J. C. *Deep Learning Schemes For Parabolic Nonlocal Integro-Differential Equations* , Preprint 2021, see https://arxiv.org/abs/2103.15008.
- J. C. *The Kolmogorov Infinite Dimensional Equation in a Hilbert space Via Deep Learning Methods*, Preprint 2022, see https://arxiv.org/abs/2206.06451.

These two articles contain the main details presented in this thesis; however, here we have chosen to expand and explain most of the standard ideas that are simply assumed in those papers. **Our main results are Theorems 4.7 and 5.9.** The first one deals with the nonlocal Euler scheme for the Kolmogorov model, the second one considers the Hilbert-valued case.

1.1.1 Brief historical review

Equation (1.2) can be recast as a nonlinear parabolic model, generalizing the classical Heat equation. The mathematical theory in this case is well-known, see e.g. [Eva10, Section 2.3]. Of great importance to the present work is the well known relation between probabilities and parabolic models, As we stated before, A. N. Kolmogorov was the first (of many) to notice these relations in his foundational work [Kol38], the resulting theory allows to prove existence, uniqueness and properties of solutions to parabolic models, known as Kolmogorov equations, by means of probabilistic ideas. These models, also known as diffusion equations, has many applications in Finance and other areas such as physics, biology, chemistry and economics. The success in applications came from the fact that these equations are describing the general phenomena of particles interacting under the influence of random forces (see e.g. [Cra16]).

However, if we replace $\mathbb{R}^d$ by a separable Hilbert space H, (1.2) becomes a highly complicated model that requires sophisticated treatment and generalizations for the classical existence and regularity theories. Infinite dimensional Kolmogorov equation was first investigated by Yu. Daleckij [Dal66] and L. Gross [Gro67]. In the context of PDEs it is common to define weaker notion of solutions. In this particular framework, *mild solutions* of (1.9) are treated in [FT02]. A function $u\colon [0,T]\times H \to \mathbb{R}$ is called a **mild solution** to (1.9) if it satisfies $u \in C^{0,1}([0,T]\times H)$, there exists $C>0$ and $p\in\mathbb{N}$ such that $|\langle \nabla u(t,x), h\rangle_H| \leq C\,\|h\|_H\,(1+\|x\|_H^p)$ for all $t\in[0,T]$ and $x,h\in H$ and the following weaker formulation of (1.9) is satisfied

$$u(t,x) = -\int_t^T \mathbb{E}\left(\psi(s, X_s^{t,x}, u(s,X_s^{t,x}), G(s,X_s^{t,x})^*\nabla u(s,X_s^{t,x}))\right) ds + \mathbb{E}\phi(X_s^{t,x}),$$

where $(X_s^{t,x})_{s\in[t,T]}$ is the solution to a stochastic evolution equation starting with $X_t^{t,x}=x$. In [FT02] the authors prove that there exists a unique mild solution to (1.2) which is related to the stochastic equations through $u(t,x)=Y_t^{t,x}$, where $Y^{t,x}$ is part of the solution to the stochastic equation in $[t,x]$ starting with $X_t^{t,x}=x$. As you may see in Section 5.1, for our framework we need a **strong solution** of (1.2) in order to be able to use Itô lemma. The existence of said solution can be seen as a strong assumption in our model, nevertheless, see [BHJ21, Lemma 2.2] for an existence result.

The mathematics presented here is strongly inspired by the article [HPW19] written by Hure, Pham and Warin, where they rely on the stochastic representation of (1.2) and the use of neural

networks to approximate a solution of the PDE and its spatial gradient. Due to the importance of this work to us, we provide a detailed enough description of the scheme presented in [HPW19] and certain ideas of generalization; Consider a partition π of $[0,T]$. By taking advantage of the relations $Y_t = u(t,X_t)$ and $Z_t = \sigma^T(t,X_t)\nabla u(t,X_t)$ showed in [PP90] and the Itô formula, mentioned authors proposed a pair of neural networks $\mathcal{U}_t(\cdot;\theta)$ and $\mathcal{Z}_t(\cdot;\theta)$ for every $t \in \pi$ such that

$$\mathcal{U}_t(X_t^\pi;\theta) \approx Y_t \text{ and } \mathcal{Z}_t(X_t^\pi;\theta) \approx Z_t,\ t \in \pi.$$

Where X^π is a suitable Euler approximation of the diffusion X and θ represents the neural network parameters (see [HPW19, Section 3]). Then, by imposing that the neural network representation satisfies the Ito formula with a cost incurred by the approximation, an iterative backward induction is produced such that at each time step a loss function representing the cost is minimized. This process generates optimal neural networks for every time step $t \in \pi$. The backward form of the algorithm emerges from the knowledge of the solution at the final time, also known as terminal condition. It is important to mention that Hure et al. extend this approach to treat variational inequalities. In Chaper 4 we study the non-local form of (1.2). This modification introduces complications such as the need of a general diffusion which admits discontinuities (see Chapter 2 for details). This type of processes are known in the literature as Lévy processes and are suitable to obtain the desire representation as in the local case (see [PP90]). Examples of non-local terms includes integrals with respect to a Levy measure λ, only finite Levy measures are taking under consideration in Chapter 4, this restriction leaves out interesting operators such as fractional laplacian. Other complication that we encounter setting is the need of a third neural network to approximate the non-local term. A different approach to treat the non-locality is considered by Lukas Gonon and Christoph Schwab in [GS21a, GS21b], they prove that NNs of a particular structure are able to approximate expectation of a certain type of functions defined on the space of stochastic processes with jumps, which can express certain PDEs solutions. In their proof, L. Gonon and C. Schwab provide dimension-explicit bounds evidencing that their scheme is free from the *curse of dimensionality* (mentioned below). These articles are a generalization of the method presented previously in [HJK$^+$20, HJKN20], for more details see the references there in.

Now we describe the two frameworks worked in this thesis.

1.1.2 Finite-dimensional non-local framework

Let $d \geq 1$ and $T > 0$. For the non-local case we consider the following target Partial Integral Differential Equation (PIDE),

$$\begin{cases} \mathcal{L}u(t,x) + f(t,x,u(t,x),\sigma(x)\nabla u(t,x),\mathcal{I}[u](t,x)) = 0, & (t,x) \in [0,T)\times\mathbb{R}^d, \\ u(T,x) = g(x), & x \in \mathbb{R}^d. \end{cases} \tag{1.3}$$

Here, $u = u(t,x)$ is the unknown of the problem. The operator $\mathcal{L}$ above is of parabolic nonlocal type, and is defined, for $u \in \mathcal{C}^{1,2}([0,T)\times\mathbb{R}^d)$, as follows:

$$\begin{aligned}\mathcal{L}u(t,x) = \partial_t u(t,x) + \nabla u(t,x)\cdot b(x) + \frac{1}{2}\text{Trace}(\sigma(x)\sigma(x)^T D^2 u(t,x)) \\ + \int_{\mathbb{R}^d}[u(t,x+\beta(x,y)) - u(t,x) - \nabla u(t,x)\cdot\beta(x,y)]\lambda(dy),\end{aligned} \tag{1.4}$$

where λ is a finite measure on $\mathbb{R}^d$, equipped with its Borel σ-algebra, and a Lévy measure as well which means that

$$\lambda(\{0\}) = 0 \qquad \text{and} \qquad \int_{\mathbb{R}^d} (1 \wedge |y|^2)\lambda(dy) < \infty.$$

On the other hand, the non-local, integro-differential operator $\mathcal{I}$ is defined as

$$\mathcal{I}[u](t,x) = \int_{\mathbb{R}^d} \big(u(t, x + \beta(x,y)) - u(t,x)\big)\lambda(dy). \tag{1.5}$$

Together with PIDE (1.3), consider the following stochastic system,

$$X_t = x + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s + \int_0^t \int_{\mathbb{R}^d} \beta(X_{s^-}, y)\overline{\mu}(ds, dy), \tag{1.6}$$

$$Y_t = g(X_T) + \int_t^T f(\Theta_s)ds - \int_t^T Z_s \cdot dW_s - \int_t^T \int_{\mathbb{R}^d} U_s(y)\overline{\mu}(ds, dy), \tag{1.7}$$

$$\Gamma_t = \int_{\mathbb{R}^d} U_t(y)\lambda(dy), \tag{1.8}$$

where $\Theta_s = (s, X_s, Y_s, Z_s, \Gamma_s)$ for $0 \le s \le T$ and $x \in \mathbb{R}^d$. Note that $(Z_t)_{0\le t\le T}$ is a vector valued process. Functions f, σ, β, b and g, satisfies the standard Lipschitz conditions in order to ensure the existence and uniqueness of solutions to the stochastic equations (see Assumptions 2.10 below).

1.1.3 Hilbert local framework

Let H, V be separable Hilbert spaces with inner products $\langle\cdot,\cdot\rangle_H$ and $\langle\cdot,\cdot\rangle_V$, and $T > 0$. We consider the infinite dimensional Kolmogorov model

$$\begin{cases} \partial_t u(t,x) + \mathcal{L}[u](t,x) + \psi\big(t, x, u(t,x), B^*(t,x)\nabla u(t,x)\big) = 0, & (t,x) \in [0,T] \times H, \\ u(T,x) = \phi(x), & x \in H. \end{cases} \tag{1.9}$$

Here $u\colon [0,T] \times H \to \mathbb{R}$ is the unknown of the problem, $B^*(t,\cdot)$ is the formal adjoint of a suitable mapping B, $\phi\colon H \to \mathbb{R}$ is a terminal condition and ψ represents the non-linear character of the problem. ∇ represents the spatial gradient in H. Finally, the operator $\mathcal{L}$ is defined for $f \in C^{1,2}([0,T] \times H)$ and $(t,x) \in [0,T] \times H$, as follows:

$$\mathcal{L}[f](t,x) = \langle \nabla f(t,x), Ax + F(t,x)\rangle_H + \frac{1}{2}\mathrm{tr}\left(\nabla^2 f(t,x)(B(t,x)Q^{1/2})(B(t,x)Q^{1/2})^*\right). \tag{1.10}$$

The precise details on these terms are fixed below in Assumptions 2.12. As well as in the finite dimensional case, we have to consider a stochastic system. In this setting we have a decoupled system of stochastic partial differential equations (SPDEs) for $(X_t, Y_t, Z_t)_{t\in[0,T]}$,

$$X_t = x + \int_0^t (AX_s + F(s, X_s))ds + \int_0^t B(s, X_s)dW_s, \tag{1.11}$$

$$Y_t = \phi(X_T) + \int_t^T \psi(s, X_s, Y_s, Z_s)ds - \int_t^T \langle Z_s, \cdot\rangle_0 dW_s, \tag{1.12}$$

where $\langle \cdot, \cdot \rangle_0$ is a suitable $\mathcal{L}$ based inner product to be defined below.

Forward Backward stochastic systems such as (1.6)-(1.7) were first studied by Pardoux and Peng in the finite dimensional local case [PP90], whereas Barles, Buckdahn and Pardoux [BBP97] generalized it to the case where also a non continuous process is considered. For the stochastic equation posed on infinite dimensional spaces (1.11)-(1.12), we mention [FT02] as an important article in the subject, see also the book [DPZ92] and [AGM+16].

1.2 Deep Learning

The huge amount of available data, due to social media, astronomical observatories and even Wikipedia, together with the progress of computational power, have allowed us to train more and more efficient Machine Learning (ML) algorithms and consider data that years ago were not possible to analyze. *Deep Learning* is a part of supervised ML algorithms and it concerns with the problem of approximating an unknown nonlinear function $f : X \to Y$, where X represents the set of possibles inputs and Y the outputs, for example, Y could be a finite set of classes and therefore f has a classification task. In order to perform a DL algorithm we need a set of observations $D = \{(x, f(x)) : x \in A\}$ of the phenomenon under consideration; in the literature this set is also known as training set. Here, A is a finite subset of X. The next step is to define a family of candidates $\left\{f^\theta : \theta \in \Xi\right\}$ where we can search for a good approximation of f, with $\Xi \subset \mathbb{R}^\kappa$ for some $\kappa \in \mathbb{N}$. Finally, how good the approximation is, will be measured by a cost function $L(\cdot; D) : \Xi \to \mathbb{R}$ and therefore, intuitively, we take f^{θ^*} as the chosen approximation where θ^* minimizes $L(\cdot; D)$ over Ξ.

1.2.1 Brief review of the literature

Neural Networks (NNs) are not recent. In [MP43] and [Ros58], published in 1943 and 1958 respectively, the authors introduce the concept of NN but far from the actual definition. Through the years, the use of NNs as a way to approximate functions, started to gain importance for its well performance in applications. A rigorous justification of this property was proven in [Hor91, LLPS93], using the Stone-Weierstrass theorem. These papers state that under suitable conditions on the approximated functions, measured in some mathematical terms, NNs have a remarkable performance. See [WR17, ATY+19] for a review on the origin and state of the art survey of DL, respectively.

The complexity and generality of the problem that DL is trying to solve, makes it useful to a large variety of disciplines in science. In astronomy, the large amount of data recollected by observatories makes it a suitable place to implement ML, see [Bar19] for a review of ML in astronomy and [AKF+17] for a concrete use of Convolutional Neural Networks (CNN) to classify light curves. See [Bou19] for a review of ML on experimental high energy physics and [TMC+18] for an application of NN on quantum state tomography. In [MBS17], the authors use DL to find patterns in fashion and style trends by space and time using data from Instagram. In [Alq19] the authors train a CNN to classify brain tumors into Glioma, Meningioma, and Pituitary Tumor reaching high levels of accuracy. See [LKB+17] for a survey on the use of DL in medical science where CNN are the most common type of DL structure.

In recent developments, finite dimensional Deep Learning (DL) has proven itself to be an efficient tool to solve nonlinear problems such as the approximation of PDEs solutions (see [ATY+19]). In particular, in high dimensions $d \gg 1$, typical methods such as finite difference or finite elements suffer from the fact that the complexity of the problem grows exponentially on d, problem known in the literature as *curse of dimensionality*. Without being exhaustive, we present some of the current developments in this direction. First of all, Monte Carlo algorithms are an important and widely used approach to the resolution of the dimension problem. This can be done by means of the classical Feynman-Kac representation that allows us to write the solution of a linear PDE as an expected value, and then approximate the high dimensional integrals with an average over simulations of random variables. On the other hand, Multilevel Picard method (MLP) is another approach and consists on interpreting the stochastic representation of the solution to a semilinear parabolic (or elliptic) PDE as a fixed point equation. Then, by using Picard iterations together with Monte Carlo methods for the computation of integrals, one is able to approximate the solution to the PDE, see [BHH+20, HJK+20] for fundamental advances in this direction. As another option, the so-called Deep Galerkin method (DGM) is another DL approach used to solve quasilinear parabolic PDEs of the form $\mathcal{L}(u) = 0$ plus boundary and initial conditions. The cost function in this framework is defined in an intuitive way, it consists of the differences between the approximated solution $\hat{u}$ evaluated at the initial time and spatial boundary, with the true initial and boundary conditions plus $\mathcal{L}(\hat{u})$. These quantities are captured by an L^2-type norm, which in high dimensions is minimized using Stochastic Gradient Descent (SGD) method. See [SS18] for the development of the DGM and [MNdH20] for an application. The article [EHJ17] by E, Han and Jentzen, is considered one of the first attempts to solve this issue by means of Deep Learning (DL) techniques. In said paper, the authors proposed an algorithm for solving parabolic PDEs by reformulating the problem as a stochastic control problem. This connection also come from the Feynman-Kac representation, proving once more that stochastic representations are a key tool in the area. More recent developments in this area can be found in Han-Jentzen-E [HJE18] and Beck-E-Jentzen [BEJ19].

The problem to generalize neural networks to an infinite dimensional framework has been investigated in dynamical systems and PDEs. In our case, following [HPW19, Cas21, Cas22] given the partition $\pi = \{t\}_{t\in\pi}$ of $[0,T]$, we want to approximate the solution $u(t,\cdot)$ to (1.9) and a fixed function of its gradient $\nabla u(t,\cdot)$ for $t \in \pi$, which in general are nonlinear operators from H to some other separable real Hilbert space $(W, \langle\cdot,\cdot\rangle_W, \|\cdot\|_W)$. Thus, we need a general Deep Learning framework which considers the approximation of operators $F\colon H \to W$ by a neural network $F^\theta\colon H \to W$, where θ is a finite dimensional parameter. Sandberg [San91] defined a set of infinite dimensional mappings parameterized by finite dimensional parameters, providing a universal approximation theorem for those mappings. Other important article in the development of infinite dimensional neural networks and an key reference for the theory presented here, is [CC95] by Chen and Chen. They deal with the approximation of mappings defined on a compact subset of $C(K)$ with values in $\mathbb{R}$ and $C(K)$, where K is a compact subset of a finite dimensional space. A key lemma ([CC95, Lemma 7]) presented in there says that, for a compact set V in $C(K)$, one can consider a transformation T and define $T(V) = \{Tu\colon u \in V\}$ such that every function in V is close to its transformation. One can compare these ideas with the approximation of measurable functions with simple functions in integral-type distance and continuous functions with polynomials in uniform norm. The transformed set is constituted by, in some sense, simpler functions that can be easily described by finite dimensional neural networks which allows them to create a proper architecture. Our Lemma 3.15 is the counterpart of [CC95, Lemma 7] for a compact set V in a Hilbert space. Here, the considered transformation is the projection into a finite-dimensional subspace.

Chen and Chen also demonstrate that their architectures approximate any continuous mapping in uniform norm. More recently Lu, Jin and Karniadakis, based on [CC95], introduced an architecture called **DeepONets** [LJP+21], these are mappings between Banach spaces of continuous functions. DeepONets rely on representing the input function as its evaluation on a fixed finite set of points. Then, via an activation function, one takes the finite dimensional information to an element of the set of continuous functions.

It is common in machine learning and, more generally, in some statistics frameworks, to consider mean square error due to its convexity properties. Here this framework emerges naturally because we make use of stochastic processes, which will be essentially square integrable random variables. The quantity used to measure the error incurred in our scheme will depend on how good our architectures are able to approximate elements of $L^2(H, \mu; W)$. Here, μ is the law of an H-valued random variable X (this random variable will be related to a stochastic process). Then, it is natural to consider the L^2-distance or mean square error

$$\mathbb{E}\left\|F(X) - F^\theta(X)\right\|_W^2 = \int_H \left\|F(x) - F^\theta(x)\right\|_W^2 \mu(dx),$$

where F is some mapping and F^θ the proposed architecture.

1.3 Thesis Layout

In Chapter 2 we provide the mathematical framework for the development of our results. A review on stochastic processes, Hilbert spaces and Linear operators is also presented.

Chapter 3 serves as a mathematical guide for Deep Learning techniques. We begin by defining Neural Networks as functions mapping the Euclidean spaces and parameterized by a set of finite dimensional parameters. This simple but key setting is useful when one generalizes to infinite dimensional problems. Indeed, our second PDE problem is posed on a Hilbert space, where we work with the so called DeepOnets, the generalized version of finite dimensional Neural Networks. These very recent objects were introduced by Lanthaler et al. in [LMK21], and consist on maps between Banach spaces of continuous functions. Taking advantage of the structure of Hilbert spaces, we define *Deep-H-Onets* and derive their key properties. We also prove density properties for Deep-H-Onets.

In Chapters 4 and 5 our main goal is to prove the main results for the two Kolmogorov models considered in this thesis. In the first one we extend Hure, Pham and Warin to the nonlocal case, where jump Lévy processes are needed to fully describe the stochastic setting. We prove consistence of suitable Euler schemes based on approximative neural networks. Finally, Chapter 5 is fully devoted to the Hilbert-posed case, and in this case Deep-H-Onets are used to prove consistence of the proposed Euler scheme. The difficulties here are in terms of the Hilbert valued stochastic setting, as well as well-posedness results in the considered framework.

Chapter 2

Mathematical Preliminaries

2.1 Notation

We cannot continue without introducing some notation needed along the manuscript.

Finite dimension. For any $m \in \mathbb{N}$, $\mathbb{R}^m$ represents the finite dimensional Euclidean space with elements $x = (x_1, ..., x_m)$ endowed with the usual norm $\|x\|^2_{\mathbb{R}^m} = \sum_{i=1}^m |x_i|^2$. We will simply write $\|x\|$ when no confusion can arise. Note that for scalars $a \in \mathbb{R}$ we also denote its norm as $|a| = \sqrt{a^2}$. For $x, y \in \mathbb{R}^m$ their scalar product is denoted as $x \cdot y = \sum_{i=1}^m x_i y_i$. Finally, along this paper we will use several times that for $x_1, ..., x_k \in \mathbb{R}$, the following bound holds,

$$(x_1 + \cdots + x_k)^2 \leq k(x_1^2 + \cdots + x_k^2). \tag{2.1}$$

Along this manuscript, $C > 0$ will denote a constant that may change from one line to another, specially in the proofs. Also, the notation $a \lesssim b$ means that there exists $C > 0$ such that $a \leq Cb$.

Banach spaces. Consider now two real Banach spaces E, F. Given a subset $A \subset E$ we denote as $\langle A \rangle$ the set containing all the finite linear combination of elements in A. Separable real Hilbert spaces will be denoted as $(H, \langle \cdot, \cdot \rangle_H, \|\cdot\|_H)$. We denote by $C^m(E; F)$ the set of all m times continuously differentiable functions from E to F and $C^m(E)$ when $F = \mathbb{R}$.

Measures. We also denote by $\mathcal{B}(E)$ the Borel σ-algebra on E. For a general measure space $(E, \mathcal{H}, \nu)$ and $p \geq 1$, $L^p(E, \mathcal{H}, \nu; F)$ represents the standard Lebesgue space of all p-integrable functions from E to F, with its Borel σ-algebra, and endowed with the norm

$$\|f\|^p_{L^p(E,\mathcal{H},\nu;F)} = \int_E \|f(x)\|^p_F \, \nu(dx).$$

We write $L^p(E, \mathcal{H}, \nu)$ when $F = \mathbb{R}$ and $L^p(E, \nu)$ when $F = \mathbb{R}$ and $\mathcal{H}$ is the Borel σ-algebra $\mathcal{B}(E)$. See the "Appendix A" section of [WL15] for a definition of the above Bochner integral and its properties. We also write

$$\int_E f(s)ds = \begin{pmatrix} \int_E f_1(s)ds \\ \vdots \\ \int_E f_m(s)ds \end{pmatrix},$$

whenever $f : E \to \mathbb{R}^m$ with $f = (f_1, ..., f_m)$.

Stochastic processes. We refer to [DPZ92] for a detailed development of Stochastic Calculus in infinite dimensions. Here we will need the following definitions.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a complete filtered probability space. Given a E-valued random variable $X : \Omega \to E$, we write $\mathbb{E}X = \mathbb{E}(X)$ and $\mathbb{E}_t X = \mathbb{E}_t(X) = \mathbb{E}(X|\mathcal{F}_t)$ for any $t \geq 0$. We denote by $\sigma(X)$ the σ-algebra generated by X and by $\mathcal{P}_s$ the predictable σ-algebra of $[0, s] \times \Omega$. Let us denote by $\mathscr{S}^2 = \mathscr{S}^2_T(E)$ the space of E-valued predictable processes $(X_t)_{t\in[0,T]}$ endowed with the norm $\|X\|_{\mathscr{S}^2} = \mathbb{E}\left(\sup_{t\in[0,T]} \|X_t\|^2_E\right)$. We denote $\mathscr{M}^2_T(E) \subset \mathscr{S}^2_T(E)$ the space of E-valued continuous, square integrable martingales $(M_t)_{t\in[0,T]}$ such that $M_0 = 0$ endowed with the norm $\|M\|_{\mathscr{M}^2} = \|M\|_{\mathscr{S}^2}$. Note that if $X \in L^2(\Omega, \mathcal{F}, \mathbb{P}; E)$, then $M_t = \mathbb{E}(X|\mathcal{F}_t)$ defines a martingale in $\mathscr{M}^2_T(E)$. We also have that if M is a continuous martingale, then Doob's inequality holds,

$$\mathbb{E}\left(\sup_{t\in[0,T]} \|M_t\|^2_E\right) \leq 4 \sup_{t\in[0,T]} \left(\mathbb{E}\|M_t\|^2_E\right).$$

If no confusion arises, we will drop the parentheses $(\cdot)$ in each $\mathbb{E}$.

2.2 Hilbert spaces and Linear Operators

Our main result of Chapter 5 is posed in a Hilbert space framework. Here we set up notation and terminology about these spaces, for a detailed presentation on these spaces we refer to [Bre11, Section 5]. We will also define a couple of spaces of bounded linear operators. In what follows consider a separable real Hilbert space H doted of a scalar product $\langle\cdot,\cdot\rangle_H$ and a corresponding norm $\|\cdot\|_H$, this will be denoted as $(H, \langle\cdot,\cdot\rangle_H, \|\cdot\|_H)$. Along this manuscript, the norm and scalar product of a Hilbert space H will always contains the subscript $\cdot_H$. It will be important for the proof of our main results the existence of an *orthonormal basis* in H defined below.

Definition 2.1 *A collection $(e_n)_{n\in\mathbb{N}}$ in H is said to be an orthonormal basis of H if it satisfies the following properties:*

- $\|e_n\|_H = 1$ *and* $\langle e_n, e_m\rangle_H = 0$ *for all* $n \neq m$,
- *the linear space spanned by* $\{e_n\}_{n\in\mathbb{N}}$ *is dense in* H.

The following results are well-known in the theory of Hilbert spaces, they will be used along this manuscript.

Theorem 2.2 *Let $(e_n)_{n\in\mathbb{N}}$ be a orthonormal basis in H. Then for every $x \in H$, we have*

$$x = \sum_{n=1}^{\infty} \langle x, e_n\rangle_H e_n$$

and

$$\|x\|_H^2 = \sum_{n=1}^{\infty} |\langle x, e_n \rangle_H|^2.$$

PROOF. See [Bre11, Corollary 5.10] □

Theorem 2.3 *Every separable Hilbert space has an orthonormal basis.*

PROOF. See [Bre11, Theorem 5.11] □

Let now $(V, \langle \cdot, \cdot \rangle_V, \|\cdot\|_V)$ be another separable Hilbert space with an orthonormal basis $(f_k)_{k\in\mathbb{N}}$. We denote by $L(V, H)$ the normed vector space of continuous linear operators $T\colon V \to H$ endowed with the usual operator norm $\|T\|_{L(V,H)} = \sup_{\|v\|_V=1} \|Tv\|_H$, we write $L(V) = L(V, V)$. Consider now the following particular linear operators spaces.

Definition 2.4 *An operator $Q \in L(V)$ is called a trace class operator if*

$$Tr(Q) = \sum_{k=1}^{\infty} \langle Q f_k, f_k \rangle_V < \infty \tag{2.2}$$

Remark 2.5 *Note that in finite dimension, expression (2.2) coincide with the well-known trace of a matrix $A \in \mathbb{R}^{d\times d}$. We have therefore that every linear operator in $L(\mathbb{R}^d)$ is of trace class.*

Definition 2.6 *Let $L_2(V, H)$ be the space of Hilbert-Schmidt operators $T \in L(V, H)$ such that*

$$\|T\|_{L_2(V,H)}^2 = \sum_{n=1}^{\infty} \|T f_n\|_H^2 < \infty.$$

The set $L_2(V, H)$ is endowed with the norm $\|T\|_{L_2(V,H)}$.

Setting 2.7 *From now on, we fix the following*

- *Separable real Hilbert spaces H, V. Recall the respective scalar products,*
- *$Q \in L(Q)$ a nonnegative trace class operator,*
- *$V_0 = Q^{1/2}V = \left\{Q^{1/2}v \;\middle|\; v \in V\right\}$ which is another Hilbert space endowed with $\langle \cdot, \cdot \rangle_0 = \langle Q^{-1/2}\cdot, Q^{-1/2}\cdot \rangle_V$ and the corresponding norm $\|\cdot\|_0$.*

Operator Q will appear in the definition of the operator $\mathcal{L}$ present in (1.2) when the equation is posed on a Hilbert space.

Remark 2.8 *Note that $L(V,H) \hookrightarrow L_2(V_0,H)$. Also, observe that if we replace H by $\mathbb{R}$, then for every $v \in L(V,\mathbb{R}) = V^*$ (up to isomorphism),*

$$\|v\|^2_{L_2(V,\mathbb{R})} = \sum_{j=1}^{\infty} |\langle v, f_j\rangle|^2 = \|v\|^2_V .$$

Therefore in this particular case $L(V,\mathbb{R}) = L_2(V,\mathbb{R})$.

We shall assume

Assumptions 2.9 *There exists a bounded sequence of nonnegative real numbers $(\lambda_k)_{k\in\mathbb{N}}$ such that $Qf_k = \lambda_k f_k$ for $k \in \mathbb{N}$.*

Due to Q been trace class, one can prove that $\mathrm{Tr}(Q) = \sum_{k=1}^{\infty} \lambda_k < \infty$, result known as Lidskii's theorem. We provide an example of trace class operator: consider the usual Hilbert space V (could be any Hilbert space) and $v, w \in V$, define the bounded linear operator $T_{v,w} \in L(V)$ such that $T_{v,w}z = \langle z, w\rangle_V v$ for any $z \in V$. Then $\mathrm{Tr}(T_{v,w}) = \langle v, w\rangle_V$. Furthermore, any bounded linear operator with finite-dimensional rank is trace class.

2.3 Assumptions

The existence and uniqueness of solution to both stochastic systems is strongly related with the structure of the non-linearities and the others functions involved. In this section we fix those structures in order to continue. Recall system (1.6)-(1.7), we assume the following:

Assumptions 2.10 *There exists a universal constant $K > 0$ such that*

- *(Regularity) $g : \mathbb{R}^d \to \mathbb{R}$, $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are K-Lipschitz real, vector and matrix valued functions, respectively.*
- *(Boundedness)$\beta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sup_{y\in\mathbb{R}^d} |\beta(0,y)| \le K$.*
- *(Uniformly Lipschitz) $\sup_{y\in\mathbb{R}^d} |\beta(x,y) - \beta(x',y)| \le K|x - x'|$, $\forall\, x, x' \in \mathbb{R}^d$.*
- *(Hölder continuity) For each $t, t' \in [0,T]$, $y, y', w, w' \in \mathbb{R}$ and $x, x', z, z' \in \mathbb{R}^d$, one has*

  $$|f(t,x,y,z,w) - f(t',x',y',z',w')| \le K\big(|t-t'|^{1/2} + |x-x'| + |y-y'| + |z-z'| + |w-w'|\big).$$

- *(Invertibility) For each $y \in \mathbb{R}^d$, the map $x \to \beta(x,y)$ admits a Jacobian matrix $\nabla\beta(x,y)$ such that the function $a(x,\xi;y) = \xi^T(\nabla\beta(x,y) + I)\xi$ satisfies, for all $x, \xi \in \mathbb{R}^d$, $a(x,\xi;y) \ge |\xi|^2K^{-1}$ or $a(x,\xi;y) \le -|\xi|^2K^{-1}$.*

Remark 2.11 *In the literature (see [BBP97, Del13]) a Lipschitz condition imposed on β is often written as*

$$|\beta(x,y) - \beta(x',y)| \le K_1|x - x'|(1 \wedge |y|), \quad \textit{for some } K_1 > 0 \textit{ and for all } x, x', y \in \mathbb{R}^d.$$

The reason to impose this requirement is to ensure that

$$\int_{\mathbb{R}^d} |\beta(x,y) - \beta(x',y)|^2 \lambda(dy) \leq K_2 |x - x'|^2,$$

for some constant $K_2 > 0$, this is another way of saying that β is Lipschitz with respect to its first variable in an integral sense. Our uniformly Lipschitz requirement on β and λ being a finite measure is enough to satisfy the said restriction.

In the other hand, the infinite dimensional framework in system (1.11)-(1.12) is subject to:

Assumptions 2.12 *There exists a constant $K > 0$ such that,*

1. **Structure of $\mathcal{L}$.** *The operator $\mathcal{L}$ is defined for $f \in C^{1,2}([0,T] \times H; \mathbb{R})$ and $(t,x) \in [0,T] \times H$ as in (1.10), where*

   - *$\nabla f \in H$ is the standard gradient, and $\nabla^2 f$ is the bilinear operator second derivative;*
   - *$A \colon \mathcal{D}(A) \subset H \to H$ is the infinitesimal generator of a C_0-semigroup $\{S(t), t \geq 0\}$ on H, with $\mathcal{D}(A)$ dense in H and $x \in \mathcal{D}(A)$.*
   - *F is a drift term and B is an diffusion operator satisfying*

     $$F \colon [0,T] \times H \to H, \qquad B \colon [0,T] \times H \to L_2(V_0, H),$$

     are $(\mathcal{B}([0,T]) \otimes \mathcal{B}(H))$-$\mathcal{B}(H)$ and $(\mathcal{B}([0,T]) \otimes \mathcal{B}(H))$-$\mathcal{B}(L_2(V_0,H))$ measurable mappings, respectively. Furthermore, they satisfy that for all $x, y \in H$ and $t \in [0,T]$,

     $$\|F(t,x) - F(t,y)\|_H + \|B(t,x) - B(t,y)\|_{L_2(V_0,H)} \leq K \|x - y\|_H,$$

     and

     $$\|F(t,x)\|_H^2 + \|B(t,x)\|_{L_2(V_0,H)}^2 \leq K^2 (1 + \|x\|_H^2).$$

     These mean that F and B are uniformly Lipschitz, with linear growth.

   - *For all $r, s \in [0,T]$ with $r < s$ and $y \in H$,*

     $$S(s-r)F(r,y) \in \mathcal{D}(A), \quad S(s-r)B(r,y) \in \mathcal{D}(A).$$

     And, there exists positive functions $g_1, g_2 \in L^1([0,T])$ such that

     $$\begin{aligned} \|AS(s-r)F(r,y)\|_H &\leq g_1(s-r)\left(1 + \|y\|_H\right), \\ \|AS(s-r)B(r,y)\|_{L_2(V_0,H)}^2 &\leq g_2(s-r)\left(1 + \|y\|_H^2\right). \end{aligned}$$

   Note that this tells us that F and B are uniformly bounded in $[0,T]$ for fixed $x \in H$. We also denote as B^* the adjoint operator of B.

2. **Structure of the nonlinearity.** *$\psi \colon [0,T] \times H \times \mathbb{R} \times V \to \mathbb{R}$ is the nonlinearity in (1.9), which satisfies that for $t, t' \in [0,T]$, $x, x' \in H$, $y, y' \in \mathbb{R}$ and $z, z' \in V$,*

   $$|\psi(t,x,y,z) - \psi(t',x',y',z')| \leq C(|t - t'|^{1/2} + \|x - x'\|_H + |y - y'| + \|z - z'\|_V). \tag{2.3}$$

These assumptions are standard in the literature, see e.g. [HPW19]. In particular, condition (2.3) on ψ is required to control our numerical scheme in a satisfactory way. As for the conditions on $\mathcal{L}$, these are also common in the infinite dimensional literature, as expressed for example in [FT02]. For any $u \in V$ we have that $\left\|Q^{1/2}u\right\|_V \leq \left\|Q^{1/2}\right\|_{L(V)} \|u\|_V = \left\|Q^{1/2}\right\|_{L(V)} \left\|Q^{1/2}u\right\|_0$, which will be implicitly used during the paper.

2.4 Stochastic Calculus

In this section we gather some necessary definitions and results involving the stochastic side of our problem. We consider a general probabilistic setting posed on a separable Hilbert space H, the finite dimensional case will follows naturally as a particular case by taking $H = \mathbb{R}^d$ for some $d \in \mathbb{N}$. The definitions presented here are detailed in [DPZ92].

2.4.1 A review on Stochastic processes

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0\leq t\leq T}, \mathbb{P})$ be a filtered probability space satisfying the usual conditions: $(\mathcal{F}_t)_{0\leq t\leq T}$ is right continuous, and $\mathcal{F}_0$ is complete (contains all zero measure sets). A H-valued random variable is a strongly $\mathcal{F}$-$\mathcal{B}(H)$ measurable function $X\colon \Omega \to H$, here $\mathcal{B}(E)$ denotes the Borel sigma algebra for any topological space E. Given a real interval $I \subset \mathbb{R}$, a family of random variables $(X_t)_{t\in I}$ is called a stochastic process. In this thesis there are two important stochastic process, Wiener and Poisson process, we introduce them both and provide theirs basis properties

Hilbert-valued Wiener processes

For this part, we refer to [DPZ92, Chapter I]. Recall the finite trace operator $Q \in L(V)$ from Setting 2.7. To introduce a V-valued Wiener process, first we need to talk about Gaussian probability measures defined on $(V, \mathcal{B}(V))$.

Definition 2.13 *A probability measure μ on $(V, \mathcal{B}(V))$ is called Gaussian if for an arbitrary $v \in V$ there exist $m \in \mathbb{R}$ and $q \geq 0$ such that,*

$$\mu\left(\{\langle v, \cdot\rangle_V \in A\}\right) = \mu\left(\{w \in V\colon \langle v, w\rangle_V \in A\}\right) = N(m, q)(A), \quad \forall A \in \mathcal{B}(\mathbb{R}).$$

Here,

$$N(m, q)(A) = \int_A \frac{1}{\sqrt{2\pi q}} e^{-\frac{(x-m)^2}{2q^2}},$$

is the Gaussian probability measure in $\mathbb{R}$.

From [DPZ92, Lemma 2.15], it can be proved that if μ is a Gaussian measure on V, then there

exists $m \in V$ and $Q \in L(V)$ such that

$$\int_H \langle v, w \rangle_V \mu(dw) = \langle m, v \rangle_V, \quad \forall v \in V, \tag{2.4}$$

$$\int_H \langle v_1, w - m \rangle_V \langle v_2, w - m \rangle_V \mu(dw) = \langle Q v_1, v_2 \rangle_V, \quad \forall v_1, v_2 \in V. \tag{2.5}$$

Moreover, μ is uniquely determined by m and Q. Vector m is called the mean and Q the covariance operator.

Definition 2.14 *A V-valued process $(W_t)_{t \geq 0}$ is called a Q-Wiener process if*

- (i) $W_0 = 0$,
- (ii) *W has continuous trajectories and independent increments and,*
- (iii) *The law of $(W_t - W_s)$ is a Gaussian measure on V with 0 mean and covariance operator $(t-s)Q$ for $t \geq s \geq 0$.*

An important observation here is that from the third point in Definition 2.14 and (2.4)-(2.5), for every $v \in V$ and $t \geq s \geq 0$, the random variable $\langle W_t - W_s, v \rangle_V : \Omega \to \mathbb{R}$ is a real-valued random variable with 0 mean and variance $\langle Qv, v \rangle_V$.

Proposition 2.15 *Let $(W_t)_{t \geq 0}$ be a Q-Wiener process. We have the following representation for $t \geq 0$*

$$W_t = \lim_{n \to \infty} W_t^n. \tag{2.6}$$

With

$$W_t^n = \sum_{k=1}^{n} \sqrt{\lambda_k} \beta_t^k f_k, \tag{2.7}$$

and

$$\beta_t^k = \frac{1}{\sqrt{\lambda_k}} \langle W_t, f_k \rangle_V.$$

Limit (2.6) is in $L^2(\Omega, \mathcal{F}, \mathbb{P}; V)$ and $(\beta^j)_{j \in \mathbb{N}}$ is a sequence of independent real valued Brownian motions on $(\Omega, \mathcal{F}, \mathbb{P})$.

PROOF. See [DPZ92, Proposition 4.3]. □

Definition 2.16 *For a Hilbert space K (usually $\mathbb{R}$ or H), we define the set $\mathcal{N}_W^2(0, T; L_2(V_0, K))$ of $L_2(V_0, K)$-valued predictable processes $\Phi : [0, T] \times \Omega \to L_2(V_0, K)$ such that*

$$\|\Phi\|^2_{\mathcal{N}_W^2(0,T;L_2(V_0,K))} = \mathbb{E} \int_0^T \|\Phi_s\|_0^2 ds < \infty,$$

endowed with the corresponding norm, i.e. $\|\cdot\|_{\mathcal{N}_W^2(0,T;L_2(V_0,K))}$ which we also denote as $\|\cdot\|_{\mathcal{N}_W^2}$ when no confusion arises.

Following the notation of Definition 2.16, such processes are suitable to integrate with respect to $(W_t)_{t\in\geq 0}$ obtaining another K-valued stochastic process

$$\int_0^t \Phi_s dW_s,\ t \geq 0, \tag{2.8}$$

which is a continuous square integrable martingale and the Itô isometry holds for every $t \in [0,T]$,

$$\mathbb{E}\left\|\int_0^t \Phi_s dW_s\right\|_K^2 = \mathbb{E}\int_0^t \|\Phi_s\|^2_{L_2(V_0,K)}\, ds.$$

See [DPZ92, Section 4.3] for key properties of this integral.

Remark 2.17 *In the case where $V = \mathbb{R}^d$, $Q = I_{\mathbb{R}^d}$ and $K = \mathbb{R}$, the corresponding space takes the form $\Phi \in \mathcal{N}_W^T(0,T;L_2(\mathbb{R}^d,\mathbb{R}))$ (recall Remark 2.8) and therefore such processes can be identified as vectors in $\mathbb{R}^d$. The following notation holds,*

$$\int_0^T \Phi_s dW_s = \int_0^T \Phi_s \cdot dW_s.$$

Lévy processes and Poisson random measures on $\mathbb{R}^d$

For this part of the section we work in a finite dimensional framework, the same definitions and results has their own version even in the Banach space case (see [AR05, Section 2] or [VM13, Chapter 2 and 3]). An important feature of Wiener processes is its continuity, in contrast, the canonical example of processes that admits "jumps" or discontinuities are called Poisson processes. Briefly, a Poisson process $(N(t))_{t\geq 0}$ is a stochastic process such that takes jumps of size 1 every exponentially distributed interval of time (see [Nor97, Section 2.4] for details). Those jumps are such that the trajectories, $t \to N(t)(\omega)$, are right continuous with left limits, property also known as cádlág from the french *continue à droite, limite à gauche*. A generalized version of a Poisson point process is given below. Let $d \in \mathbb{N}$, for the results in this subsection we refer to [App09, Chapter 1] or [Del13, Chapter 2].

Definition 2.18 *A $\mathbb{R}^d$-valued stochastic process $(X_t)_{t\in[0,T]}$ is said to be a $\mathcal{F}_t$-Lévy process on $(\Omega,\mathcal{F},\mathbb{P})$ if*

- *$(X_t)_{t\in[0,T]}$ is adapted to $(\mathcal{F}_t)_{t\in[0,T]}$,*
- *$X_0 = 0$ a.s.,*
- *$X_t - X_s$ is independent of $\mathcal{F}_s$ for $s \in [0,t)$,*
- *$X_t - X_s$ has the same distribution as X_{t-s} for $s \in [0,t)$,*
- *$(X_t)_{t\in[0,T]}$ is stochastically continuous or continuous in probability, i.e. for all $\varepsilon > 0$ and for all $s \geq 0$.*

$$\lim_{t\to s}\mathbb{P}\left(\|X_t - X_s\| > \varepsilon\right) = 0.$$

Remark 2.19 *Note that a Wiener process is also a Lévy process because it is, in particular, stochastically continuous.*

Let $(X_t)_{t\in[0,T]}$ be a $\mathbb{R}^d$-valued Lévy process, set $X_{t^-} = \lim_{s\nearrow t} X_s$ and $\Delta X_s = X_s - X_{s^-}$. In the following we introduce the Poisson random measure associated to a Lévy process. As usual, $\bar{A}$ and A^c denotes respectively the closure and complement of a given set A. Let now $A \subset \mathbb{R}^d$ be such that $A \in \mathcal{B}(\mathbb{R}^d)$ and $0 \notin \bar{A}$, this means that 0 is sufficiently "far" from A. We can count for any $t \in [0,T]$ the number of jumps of $(X_t)_{t\in[0,T]}$ of size bellowing to A by setting,

$$\mu(t,A) = \sum_{0<s\leq t} \mathbb{1}_A(\Delta X_s).$$

Note that $\mu(t,A)$ defines a random variable.

Theorem 2.20 *If $0 \notin \bar{A}$, $(\mu(t,A))_{t\in[0,T]}$ is a Poisson process.*

PROOF. See [App09, Thorem 2.3.5]. □

Theorem 2.21 *For any $\omega \in \Omega$ there is a unique σ-finite measure on $\mathcal{B}(\mathbb{R}^d \setminus \{0\})$ such that*

$$\begin{aligned} \mu(t,\cdot)(\omega)\colon \mathcal{B}(\mathbb{R}^d \setminus \{0\}) &\longrightarrow [0,+\infty) \\ A &\longmapsto \mu(t,A)(\omega). \end{aligned}$$

PROOF. See [AR05, Corollary 2.5]. □

Definition 2.22 *$\mu(t,\cdot)$ is called the Poisson random measure of the Lévy processes $(X_t)_{t\in[0,T]}$.*

Theorem 2.23 *There is a unique σ-finite measure on the σ-algebra $\mathcal{B}(\mathbb{R}^d \setminus \{0\})$ such that*

$$\begin{aligned} \lambda\colon \mathcal{B}(\mathbb{R}^d \setminus \{0\}) &\longrightarrow [0,+\infty) \\ A &\longmapsto \mathbb{E}\left(\mu(1,A)\right). \end{aligned}$$

Moreover, λ is a Lévy measure.

PROOF. See [AR05, Corollary 2.8] □

Definition 2.24 *A Lévy measure is a σ-finite measure on $(\mathbb{R}^d \setminus \{0\}, \mathcal{B}(\mathbb{R}^d \setminus \{0\}))$ such that*

$$\int_{\mathbb{R}^d\setminus\{0\}} (1 \wedge \|x\|_H^2)\lambda(dx) < \infty.$$

Definition 2.25 *The compensated Poisson random measure (cPrm) $\bar{\mu}$ is defined by*

$$\bar{\mu}(dt,dx) = \mu(dt,dx) - dt\lambda(dx).$$

We are now with a proper set up to introduce the stochastic integral with respect cPrm and state its principal properties.

Definition 2.26 *we define the set $\mathscr{N}_\mu^2(0,T;\mathbb{R})$ of predictable processes $U\colon [0,T]\times\mathbb{R}^d\setminus\{0\}:\Omega\to\mathbb{R}$ such that*

$$\|U\|^2_{\mathscr{N}_\mu^2(0,T;H)}=\mathbb{E}\int_0^T\int_{\mathbb{R}^d\setminus\{0\}}|U(s,x)|^2\lambda(dx)ds<\infty,$$

endowed with the corresponding norm, i.e. $\|\cdot\|_{\mathscr{N}_\mu^2(0,T;\mathbb{R})}$ which we also denote as $\|\cdot\|_{\mathscr{N}_\mu^2}$ when no confusion arises.

For those processes U, the stochastic integral

$$\int_0^t\int_{\mathbb{R}^d\setminus\{0\}}U(t,x)\bar{\mu}(dt,dx)\quad t\in[0,T],$$

is well-defined, is a càdlàg local martingale and the Itô isometry holds for every $t\in[0,T]$.

$$\mathbb{E}\left[\int_0^t\int_{\mathbb{R}^d\setminus\{0\}}U(t,x)\bar{\mu}(dt,dx)\right]^2=\mathbb{E}\int_0^t\int_{\mathbb{R}^d\setminus\{0\}}|U(t,x)|^2\nu(dx)dt.$$

See [Del13, Theorem 2.3.3].

2.4.2 Stochastic system in finite-dimensional non-local framework

Recall the stochastic system (1.6)-(1.7). Following lemmas provide well-known results concerning the existence and uniqueness of a solution to the required system. We only check that our hypotheses match those of [App09] and [BBP97], these results are the same as those given in [BE08] and [Del13, Section 4.1]. For this section, let $(W_t)_{t\in[0,T]}$ be a $\mathbb{R}^d$-valued Wiener process and set $\mathscr{B}^2=\mathscr{S}_T^2(\mathbb{R}^d)\times\mathscr{N}_W(0,T;\mathbb{R}^d)\times\mathscr{N}_\mu(0,T;\mathbb{R})$.

Lemma 2.27 *There exists a unique solution $X\in\mathscr{S}_T^2(\mathbb{R}^d)$ to (1.6) such that.*

$$\mathbb{E}\left(\sup_{s\le u\le t}\|X_u-X_s\|^2_{\mathbb{R}^d}\right)\le|t-s|\left(1+\mathbb{E}\|X_s\|^2_{\mathbb{R}^d}\right).\tag{2.9}$$

Proof. Recall Remark 2.11. Observe that Assumptions 2.10, particularly those imposed on β, implies that

$$\int_{\mathbb{R}^d}\|\beta(x,y)-\beta(x',y)\|^2_{\mathbb{R}^d}\lambda(dy)\le K^2\lambda(\mathbb{R}^d)\|x-x'\|^2_{\mathbb{R}^d}.$$

This, together with the rest of Assumptions 2.10 are enough to fulfill the Lipschitz and growth hypotheses needed on [App09, Section 6.2] to ensure the existence and uniqueness of a solution $X\in\mathscr{S}_T^2(\mathbb{R}^d)$ to the FSDEJ (1.6). Estimate (2.9) follows by considering the process $(X_u-X_s)_{u\in[s,t]}$ and using Doob's maximal inequality [Pro04, Theorem 20, Section 1] and Gronwall inequality. □

Lemma 2.28 *There exists a solution $(Y,Z,U)\in\mathscr{B}^2$ to (1.7).*

PROOF. We apply Theorem 2.1 of [BBP97] with $k = 1$, $Q = g(X_T)$ and a nonlinearity $\bar{f} : \Omega \times [0,T] \times \mathbb{R} \times \mathbb{R}^d \times L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda) \to \mathbb{R}$ defined as

$$\bar{f}(\omega, t, y, z, w) = f\left(t, X_t(\omega), y, z, \int_{\mathbb{R}^d} w(x)\lambda(dx)\right),$$
$$\text{for } (\omega, t, y, z, w) \in \Omega \times [0,T] \times \mathbb{R} \times \mathbb{R}^d \times L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda).$$

By the Lipschitz property on g and the bound given in Lemma 2.27 we can see that $Q \in L^2(\Omega, \mathcal{F}_T, \mathbb{P})$. The Lipschitz condition on f implies that for all $\omega \in \Omega, t \in [0,T], y, y' \in \mathbb{R}^d, z, z' \in \mathbb{R}^d$ and $w, w' \in L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$,

$$\begin{aligned}
&|\bar{f}(\omega, t, y, z, w) - \bar{f}(\omega, t, y', z', w')| \\
&= \left| f\left(t, X_t(\omega), y, z, \int_{\mathbb{R}^d} w(x)\lambda(dx)\right) - f\left(t, X_t(\omega), y', z', \int_{\mathbb{R}^d} w'(x)\lambda(dx)\right) \right| \\
&\leq K\left(|y - y'| + |x - x'| + \left| \int_{\mathbb{R}^d} (w - w')\lambda(dy) \right| \right) \\
&\leq K\left(|y - y'| + |x - x'| + \lambda(\mathbb{R}^d)^{1/2} \|w - w'\|_{L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda; \mathbb{R})}\right),
\end{aligned}$$

this proves the Lipschitz condition on $\bar{f}$. Using the previous bound is clear that,

$$\mathbb{E}\int_0^T |\bar{f}(\cdot, t, 0, 0, 0)|^2 dt < \infty.$$

These computations allow us to directly apply Theorem 2.1 of [BBP97] this finishes the proof. □

Combining previous lemmas we get that there exist a unique solution (X, Y, Z, U) to the system (1.6)-(1.7) in the space $\mathscr{S}_T^2(\mathbb{R}^d) \times \mathscr{B}^2$ this implies,

$$\mathbb{E}\left(\sup_{s\in[0,T]} \|X_s\|_{\mathbb{R}^d}^2\right) + \mathbb{E}\left(\sup_{s\in[0,T]} |Y_s|^2\right) + \mathbb{E}\int_0^T \|Z_s\|_{\mathbb{R}^d}^2 ds + \mathbb{E}\int_0^T \int_{\mathbb{R}^d} |U_s(y)|^2 \lambda(dy) ds < \infty. \tag{2.10}$$

Following lemma strongly depends on the filtration under consideration, recall that $(\mathcal{F}_t)_{t\in[0,T]}$ is generated by the two independent objects W and μ which allows us to state the representation property. See the end of Section 2.4 in [Del13] where it is stated that when the filtration is generated by a Brownian Motion and an independent jump process the required representation holds.

Lemma 2.29 (Martingale Representation Theorem) *For any square integrable martingale M there exists $(Z, U) \in \mathcal{N}_W^2(0, T; \mathbb{R}^d) \times \mathcal{N}_\mu^2(0, T; \mathbb{R})$ such that for $t \in [0, T]$*

$$M_t = M_0 + \int_0^t Z_s \cdot dW_s + \int_0^t \int_{\mathbb{R}^d} U(s, y)\overline{\mu}(ds, dy).$$

We will need the next property involving conditional expectation, Itô isommetry and that W is independent of $\overline{\mu}$.

Lemma 2.30 (Conditional Ito isometry) *For $V^1, V^2 \in \mathscr{N}_\mu^2(0,T;\mathbb{R})$ and $H, K \in \mathscr{N}_W^2(0,T;\mathbb{R}^d)$,*

$$\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} H_r \cdot dW_r \int_{t_i}^{t_{i+1}} K_r \cdot dW_r\right) = \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} H_r \cdot K_r dr\right), \tag{2.11}$$

$$\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} V^1(s,z)\overline{\mu}(ds,dz) \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} V^2(s,z)\overline{\mu}(ds,dz)\right) = \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} V^1(s,z)V^2(s,z)\lambda(dz)ds\right),$$

$$\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} V^1(r,y)\overline{\mu}(dy,dr) \int_{t_i}^{t_{i+1}} H_r \cdot dW_r\right) = 0.$$

PROOF. Follows from the classical Ito isommetry and using the definition of conditional expectation. □

Lemma 2.31 (Conditional Fubini) *Let $H \in \mathscr{N}_\mu^2(0,T;\mathbb{R})$ and $t > 0$, then*

$$\mathbb{E}\left(\int_{\mathbb{R}^d} \int_{t_i}^{t_{i+1}} H(s,y)ds\lambda(dy)\middle|\mathcal{F}_{t_i}\right) = \int_{\mathbb{R}^d} \mathbb{E}\left(\int_{t_i}^{t_{i+1}} H(s,y)ds\middle|\mathcal{F}_{t_i}\right)\lambda(dy).$$

PROOF. The proof is standard, but we included it by the sake of completeness. Let $A \in \mathcal{F}_{t_i}$, we have to prove that

$$\int_A \left(\int_{\mathbb{R}^d} \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} H(s,y)ds\right)\lambda(dy)\right)d\mathbb{P}(\omega) = \int_A \left(\int_{\mathbb{R}^d} \int_{t_i}^{t_{i+1}} H(s,y)(\omega)ds\lambda(dy)\right)d\mathbb{P}(\omega).$$

Note that because of $H \in \mathscr{N}_W^2(0,T;\mathbb{R})$,

$$\int_\Omega \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} |H(s,y)(\omega)|^2 \lambda(dy)ds d\mathbb{P}(\omega) < \infty;$$

this means that H can be seen as an element of $\in L^2(\Omega \times [t_i, t_{i+1}] \times \mathbb{R}^d) \subset L^1(\Omega \times [t_i, t_{i+1}] \times \mathbb{R}^d)$, both spaces endowed with the correspondent finite product measure. Then we can use classical Fubini theorem:

$$\begin{aligned}
\int_A \left(\int_{\mathbb{R}^d} \int_{t_i}^{t_{i+1}} H(s,y)(\omega)ds\lambda(dy)\right)d\mathbb{P}(\omega) &= \int_{\mathbb{R}^d} \left(\int_A \int_{t_i}^{t_{i+1}} H(s,y)(\omega)ds d\mathbb{P}(\omega)\right)\lambda(dy) \\
&= \int_{\mathbb{R}^d} \left(\int_A \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} H(s,y)(\omega)ds\right)d\mathbb{P}(\omega)\right)\lambda(dy) \\
&= \int_A \left(\int_{\mathbb{R}^d} \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} H(s,y)(\omega)ds\right)\lambda(dy)\right)d\mathbb{P}(\omega).
\end{aligned}$$

This finishes the proof. □

2.4.3 Stochastic system in Hilbert framework

Some useful lemmas

In this section we have compiled some basic but essential facts that will be used in the proof for introductory results to state main Theorem 5.9. Of particular importance is the *Martingale Repre-*

sentation Theorem 2.34 which allows us to find a solution for the backward stochastic equation.

Lemma 2.32 *The integral* (2.8) *can be approximated as follows: for* $n \in \mathbb{N}$ *consider the Wiener process* $(W_t^n)_{t\in[0,T]}$ *in* (2.7)*, then*

$$\mathbb{E}\left(\sup_{t\in[0,T]}\left\|\int_0^t \Phi_s dW_s - \int_0^t \Phi_s dW_s^n\right\|^2\right) \to 0 \quad as \quad N\to\infty,$$

for any $(\Phi_s)_{s\in[0,T]} \in \mathscr{N}_W([0,T]; L_2(V_0,K))$.

Lemma 2.33 *Let* $n\in\mathbb{N}$ *and* $(\Phi_s)_{s\in[0,T]} \in \mathscr{N}_W(0,T;L_2(V_0,H))$*, then the following holds,*

$$\int_0^t \Phi(s)dW_s^n = \sum_{j=1}^n \int_0^t \Phi(s)(Q^{1/2}f_j)d\beta_s^j.$$

Where W^n *is given by* (2.7)*.*

Proof: First, note that we have n integrals of H-valued processes with respect to real valued standard Brownian Motions (the associated covariance operator in this case is just 1). In our case, the space that gives sense to these integrals is $\mathscr{N}_W(0,T;L_2(\mathbb{R},H))$. It is straightforward that $L_2(\mathbb{R},H) = H$. We proceed by proving the property for elementary processes and conclude by taking the proper limit. For that purpose let $N\in\mathbb{N}$, $\{t_i\}_{i=0}^N$ be a partition of $[0,T]$ with $t_0=0$ and $t_N=T$, $\{\Phi_i\}_{i=1}^N \subset L(V,H)$ and an elementary process Φ defined as

$$\Phi(s) = \sum_{i=1}^N \Phi_i \mathbb{1}_{[t_{i-1},t_i)}(s).$$

Then by using the linearity of the operators Φ_i, definition (2.7) and $Q^{1/2}f_k = \lambda^{1/2}f_k$,

$$\begin{aligned}\int_0^t \Phi_s dW_s^n &= \sum_{i=1}^N \Phi_i(W^n_{t_{i+1}\wedge t} - W^n_{t_i\wedge t}) = \sum_{i=1}^N \Phi_i\left(\sum_{k=1}^n \sqrt{\lambda_k} f_k \beta^k_{t_{i+1}\wedge t} - \sum_{k=1}^n \sqrt{\lambda_k} f_k \beta^k_{t_i\wedge t}\right)\\
&= \sum_{i=1}^N \Phi_i\left(\sum_{k=1}^n (Q^{1/2}f_k)(\beta^k_{t_i\wedge t} - \beta^k_{t_{i-1}\wedge t})\right) = \sum_{k=1}^n\sum_{i=1}^N \Phi_i(Q^{1/2}f_k)(\beta^k_{t_i\wedge t} - \beta^k_{t_{i-1}\wedge t})\\
&= \sum_{k=1}^n \int_0^t \Phi_s(Q^{1/2}f_k)d\beta_s^k.\end{aligned}$$

It is easy to see that for every $j\in\mathbb{N}$, $(\Phi_s(Q^{1/2}f_j))_{s\in[0,T]}$ is an elementary process in $\mathscr{N}_W(0,T;L_2(\mathbb{R},H))$; therefore, the property is satisfied for those processes. Now, given a sequence of elementary processes such that $\Phi^k \to \Phi$ in $\mathscr{N}_W(0,T;L_2(V_0,H))$, we also have that for every $j\in\mathbb{N}$ $\Phi^k(Q^{1/2}f_j) \to \Phi(Q^{1/2}f_j)$ in $\mathscr{N}_W(0,T;L_2(\mathbb{R},H))$. For any $k\in\mathbb{N}$ it holds that,

$$\int_0^\cdot \Phi_s^k dW_s^N = \sum_{j=1}^n \int_0^\cdot \Phi_s^k(Q^{1/2}f_j)d\beta_s^j.$$

The property follows by taking limit in $\mathscr{M}_T^2(H)$ as $k\to\infty$ in both sides. □

Theorem 2.34 (*Martingale Representation Theorem*) *Let W be a Hilbert space and $r, s \in [0,T]$ with $r < s$. Then, for every $X \in L^2(\Omega, \mathcal{F}_s, \mathbb{P}; W)$ there exists $(Z_t)_{t\in[r,s]} \in \mathcal{N}_W([r,s]; L_2^0(V,W))$ such that*

$$X = \mathbb{E}(X|\mathcal{F}_t) + \int_t^s Z_u dW_u,\ t \in [r,s].$$

PROOF. See for instance [FT02, Proposition 4.1].

□

The forward process

Now we recall the mathematical structure associated to the forward process (X_t) in (1.11), where A, B and F were specified in Assumptions 2.12. For further details, the reader can consult [DPZ92].

Definition 2.35 (Strong and mild solutions)

1. *A predictable H-valued stochastic process $(X_t)_{t\in[0,T]}$ is said to be a* **strong solution** *of (1.11) if for all $t \in [0,T]$ $X_t \in \mathcal{D}(A)$ $\mathbb{P}$-a.e.,*

$$\int_0^T \|AX_s\|_H \, ds < \infty, \quad \mathbb{P}\text{-a.e.}$$

and equation (1.11) is satisfied for all $t \in [0,T]$.

2. *A predictable H-valued stochastic process $(X_t)_{t\in[0,T]}$ is said to be a* **mild solution** *of (1.11) if*

$$\mathbb{P}\left(\int_0^T \|X_s\|_H^2 \, ds < \infty\right) = 1,$$

and for all $t \in [0,T]$ we have the weak formulation of (1.11):

$$X_t = S(t)x + \int_0^t S(t-s)F(s,X_s)ds + \int_0^t S(t-s)B(s,X_s)dW_s, \quad \mathbb{P}\text{-a.e.} \tag{2.12}$$

The following result gives existence of mild solutions in a very general setting.

Theorem 2.36 *There exist a unique mild solution $(X_t)_{t\in[0,T]}$ to (1.11), unique among the stochastic processes satisfying,*

$$\mathbb{P}\left(\int_0^T \|X_s\|_H^2 \, ds < \infty\right) = 1.$$

Moreover, X possesses a continuous modification and for any $p \geq 2$ there exists a constant $C = C(p,T) > 0$ such that,

$$\sup_{s\in[0,T]} \mathbb{E}\,\|X_s\|_H^p \leq C(1 + \|x\|_H^p).$$

PROOF. See [DPZ92, Theorem 7.2]. □

Now we provide a proof of existence of strong solutions to (1.11), which follows closely [AGM+16, Theorem 2].

Proposition 2.37 *Assuming Assumptions 2.12 there exists a strong solution* $(X_t)_{t\in[0,T]}$ *to the equation* (1.11) *and* $C = C(T)$ *such that*

$$\sup_{s\in[0,T]} \mathbb{E}\,\|X_s\|_H^2 \leq C \quad \textit{and} \quad \mathbb{P}\left(\int_0^T \|X_s\|_H^2\, ds < \infty\right) = 1. \tag{2.13}$$

PROOF. By applying Theorem 2.36 we have a mild solution already satisfying (2.13) and then, due to Assumptions 2.12, from (2.12) we get that for all $t \in [0,T]$, $X_t \in \mathcal{D}(A)$ $\mathbb{P}$-a.e. and

$$\int_0^t AX_s = \int_0^t AS(s)x ds + \underbrace{\int_0^t \int_0^s AS(s-r)F(r,X_r)drds}_{\mathbf{I}} + \underbrace{\int_0^t \int_0^s AS(s-r)B(r,X_r)dW_r ds}_{\mathbf{II}}.$$

Basically, the idea here is to use Fubini theorem and its stochastic version (see [DPZ92, Section 4.5]) together with the fact that $S(t)y - y = \int_0^t AS(s)ds$ for $y \in \mathcal{D}(A)$. The bounds that F and B satisfy in Assumptions 2.12 imply that,

$$\begin{aligned}\int_0^T \int_0^s \|AS(s-r)F(r,X_r)\|_H\, drds &\leq \int_0^T \int_0^s g_1(s-r)drds + \int_0^T \int_0^s g_1(s-r)\,\|X_r\|_H\, drds \\ &\leq \|g_1\|_{L^1([0,T])}\left(T + \int_0^T \|X_r\|_H\, dr\right) < \infty \quad \mathbb{P}\text{-a.e..}\end{aligned}$$

And,

$$\begin{aligned}\int_0^T \mathbb{E}\int_0^s \|AS(s-r)B(r,X_r)\|_{L_2(V_0,H)}^2\, drds &\leq \int_0^T \int_0^s g_2(s-r)drds + \int_0^T \mathbb{E}\int_0^s g_2(s-r)\,\|X_r\|_H^2\, drds \\ &\leq \|g_1\|_{L^1([0,T])}\left(1 + T\mathbb{E}\left[\sup_{r\in[0,T]}\|X_r\|_H^2\right]\right) < \infty.\end{aligned}$$

Then, by Fubini Theorem,

$$\begin{aligned}\mathbf{I} &= \int_0^t S(t-r)F(r,X_r)dr - \int_0^t F(r,X_r)dr \quad \text{and,} \\ \mathbf{II} &= \int_0^t S(t-r)B(r,X_r)dW_r - \int_0^t B(r,X_r)dW_r.\end{aligned}$$

Therefore,

$$\begin{aligned}\int_0^t AX_s ds = S(t)x - x &+ \int_0^t S(t-r)F(r,X_r)dr - \int_0^t F(r,X_r)dr \\ &+ \int_0^t S(t-r)B(r,X_r)dW_r - \int_0^t B(r,X_r)dW_r.\end{aligned}$$

Hence,

$$X_t = x + \int_0^t AX_s ds + \int_0^t F(r,X_r)dr + \int_0^t B(r,X_r)dW_r \quad \mathbb{P}\text{-a.e.,}$$

and the proof is complete. □

The backward process

Now we provide existence results for the backward process (1.12), following ideas in [FT02, Lemma 4.2].

Lemma 2.38 *Let $\eta \in L^2(\Omega, \mathcal{F}_T, \mathbb{P})$ and $f \in \mathcal{N}_W(0,T;\mathbb{R})$. Then there exist a unique pair $(Y,Z) \in \mathcal{S}_T^2(\mathbb{R}) \times \mathcal{N}_W(0,T;L_2^0(V,\mathbb{R}))$ such that,*

$$Y_t = \eta + \int_t^T f_s ds - \int_t^T \langle Z_s, \cdot \rangle_0 dW_s. \tag{2.14}$$

Furthermore, the following bounds are satisfied,

$$\mathbb{E}\left(\int_0^T e^{2\beta s} \|Z_s\|_0^2 \, ds\right) \wedge \mathbb{E}\left(\sup_{s\in[0,T]} e^{2\beta s}|Y_s|^2\right) \leq \frac{4}{\beta}\mathbb{E}\int_0^T e^{2\beta s}|f_s|^2 ds + 8e^{2\beta T}\mathbb{E}|\eta|^2. \tag{2.15}$$

Where $\wedge$ indicates the maximum between both quantities.

Proof. For uniqueness to the first part of [FT02, Lemma 4.2]. First, we prove existence, define $\xi = \eta + \int_0^T f_s ds \in L^2(\Omega, \mathcal{F}_T, \mathbb{P})$. Then, by Theorem (2.34), there exists $Z \in \mathcal{N}_W(0,T;L_2^0(V,\mathbb{R}))$ such that

$$\xi = \mathbb{E}(\xi|\mathcal{F}_t) + \int_t^T \langle Z_s, \cdot \rangle_0 dW_s, \tag{2.16}$$

where we applied Remark 2.8 to notice that $L_2(V_0,\mathbb{R}) = V_0$. Define now $Y_t = \mathbb{E}(\xi|\mathcal{F}_t) - \int_0^t f_s ds$, follows that

$$Y_t = \eta + \int_t^T f_s ds - \int_t^T \langle Z_s, \cdot \rangle_0 dW_s. \tag{2.17}$$

To conclude that $(Y_t)_{t\in[0,T]} \in \mathcal{S}_T^2(\mathbb{R})$ we just note that by (2.16), (2.17) and the definition of ξ, one has for every $t \in [0,T]$

$$\mathbb{E}|Y_t|^2 \leq 3\left(\mathbb{E}|\eta|^2 + T\mathbb{E}\int_0^T |f_s|^2 ds + \mathbb{E}\int_0^T \|Z_s\|_0^2 \, ds\right) \leq 27\left(\mathbb{E}|\eta|^2 + \mathbb{E}\int_0^T f_s^2 ds\right) < \infty.$$

In order to prove estimate (2.15), we bound both quantities at left side by the right side. Sssume the existence and uniqueness of a solution (Y,Z) and note that for almost all $s \in [0,T]$, $\mathbb{E}|f_s|^2 < \infty$, thus by Theorem 2.34 there exists $(K(u,s))_{u\in[0,s]} \in \mathcal{N}_W(0,s;L_2^0(V,\mathbb{R}))$ such that,

$$f_s = \mathbb{E}(f_s|\mathcal{F}_t) + \int_t^s K(u,s) dW_u, \ t \in [0,s]. \tag{2.18}$$

We extend K to $[0,T]\times[0,T]$ in the following way,

$$\begin{aligned} K : [0,T]\times[0,T]\times\Omega &\longrightarrow L_2^0(V,\mathbb{R}) \\ (u,s,\omega) &\longmapsto K(u,s)(\omega)\mathbb{1}_{[0,s]}(u) = \begin{cases} K(u,s)(\omega), & u \leq s \\ 0, & \sim . \end{cases} \end{aligned}$$

$(\mathcal{P}_T \times \mathcal{B}([0,T]))$-measurability of K is discussed in [FT02], but it is no difficult to convince oneself of this. In the same way there exists $(L_t)_{t\in[0,T]} \in \mathscr{N}_W(0,T;L_2^0(V,\mathbb{R}))$ such that,

$$\eta = \mathbb{E}(\eta|\mathcal{F}_t) + \int_t^T L_s dW_s, \quad t \in [0,T]. \tag{2.19}$$

By taking $\mathbb{E}(\cdot|\mathcal{F}_t)$ in (2.14) then using conditional Fubini's theorem, and replacing (2.18) and (2.19) we have that for all $t \in [0,T]$,

$$Y_t = \eta - \int_t^T f_s ds - \int_t^T L_s dW_s + \int_t^T \int_t^T K(u,s)\mathbb{1}_{[t,s]}(u) dW_u ds.$$

Due to $\int_t^T \mathbb{E}\int_t^T \|K(u,s)\|_0^2 \mathbb{1}_{[t,s]}(u) du ds < \infty$ (it can be bounded by a factor of $\|f\|_{\mathscr{N}}$), we may apply stochastic Fubini theorem (see [DPZ92, Section 4.5]) getting,

$$Y_t = \eta - \int_t^T f_s ds - \int_t^T \left(L_u - \int_u^T K(u,s) ds\right) dW_s.$$

Then by uniqueness,

$$Z_u = L_u - \int_u^T K(u,s) ds, \quad \forall u \in [0,T],$$

which allows us to compute,

$$\mathbb{E}\int_0^T e^{2\beta u} \|Z_u\|_0^2 du = \underbrace{2\mathbb{E}\int_0^T e^{2\beta u}\|L_u\|_0^2 du}_{\mathbf{I}} + \underbrace{2\mathbb{E}\int_0^T e^{2\beta u}\left\|\int_u^T K(u,s)ds\right\|_0^2 du}_{\mathbf{II}}.$$

By standard procedures and using (2.19) we get $\mathbf{I} \le 8e^{2\beta T}\mathbb{E}|\eta|^2$. To work with $\mathbf{II}$ we first note that for any $u \in [0,T]$,

$$\left\|\int_u^T K(u,s)ds\right\|_0^2 \le \int_u^T e^{-2\beta s} ds \int_u^T e^{2\beta s}\|K(u,s)\|_0^2 ds \le \frac{e^{-2\beta u}}{2\beta}\int_u^T e^{2\beta s}\|K(u,s)\|_0^2 ds,$$

where we applied Bochner's estimate ($\left\|\int f\right\| \le \int \|f\|$) and Hölder's inequality. Then, by replacing the last relation in $\mathbf{II}$ and using Fubini theorem,

$$\begin{aligned}\mathbf{II} &\le \frac{1}{\beta}\mathbb{E}\int_0^T\int_u^T e^{2\beta s}\|K(u,s)\|_0^2 ds du = \frac{1}{\beta}\mathbb{E}\int_0^T\int_0^T e^{2\beta s}\|K(u,s)\|_0^2 \mathbb{1}_{[u,T]}(s) ds du \\ &= \frac{1}{\beta}\int_0^T e^{2\beta s}\mathbb{E}\left(\int_0^s \|K(u,s)\|_0^2 du\right) ds \le \frac{4}{\beta}\int_0^T e^{2\beta s}\mathbb{E}|f_s|^2 ds\end{aligned}$$

Now for the second bound we first note that by taking $\mathbb{E}(\cdot|\mathcal{F}_t)$ we have,

$$Y_t = \mathbb{E}(\eta|\mathcal{F}_t) - \mathbb{E}\left(\int_t^T f_s ds \middle| \mathcal{F}_t\right),$$

and then,

$$\mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}|Y_t|^2 \leq \underbrace{2\mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}|\mathbb{E}(\eta|\mathcal{F}_t)|^2}_{\mathbf{A}} + \underbrace{2\mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}\left|\mathbb{E}\left(\int_t^T f_s ds\middle|\mathcal{F}_t\right)\right|^2}_{\mathbf{B}}.$$

Using Doob's inequality we get $\mathbf{A} \leq 8e^{2\beta T}\mathbb{E}|\eta|^2$. For the second term,

$$\begin{aligned}
\mathbf{B} &\leq 2\mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}\left|\mathbb{E}\left(\sqrt{\int_t^T e^{-2\beta s}ds}\sqrt{\int_t^T e^{2\beta s}|f_s|^2 ds}\middle|\mathcal{F}_t\right)\right|^2 \\
&\leq \frac{1}{\beta}\mathbb{E}\sup_{t\in[0,T]}\left|\mathbb{E}\left(\sqrt{\int_0^T e^{2\beta s}|f_s|^2 ds}\middle|\mathcal{F}_t\right)\right|^2 \\
&\leq \frac{4}{\beta}\mathbb{E}\int_0^T e^{2\beta s}|f_s|^2 ds.
\end{aligned}$$

Where we used Doob's inequality on the last inequality. By putting all together we conclude the proof. □

Existence for the Forward-Backward system

The existence and uniqueness of a solution (Y, Z) to the backward equation (1.12) is well-known, here we follow the proof given in [FT02]. The argument, as we are working in a non-linear framework, relies on an application of Banach's fixed point theorem. The problem is that with the parameters as they are, the fixed-point functional does not necessarily contract. A solution to this issue is possible by giving equivalent norms to $\mathscr{N}_W(0,T;L_2^0(V;\mathbb{R}))$ and $\mathscr{S}_T^2(\mathbb{R})$ parameterized by a positive real number β. Let $\beta > 0$, consider

$$\|Y\|^2_{\mathscr{S}^2_{T,\beta}} = \mathbb{E}\left(\sup_{s\in[0,T]} e^{2\beta s}|Y|^2\right) \quad \text{and} \quad \|Z\|^2_{\mathscr{N}_{W,\beta}} = \mathbb{E}\left(\int_0^T e^{2\beta s}\|Z\|_0^2 ds\right).$$

With a bit of work we can see that $\|\cdot\|_{\mathscr{S}^2_{T,\beta}}$ and $\|\cdot\|_{\mathscr{N}_{W,\beta}}$ are equivalent to $\|\cdot\|_{\mathscr{S}^2_T}$ and $\|\cdot\|_{\mathscr{N}_W}$, respectively.

Proposition 2.39 *Given a H-valued stochastic process $(X_t)_{t\in[0,T]}$ such that*

$$\mathbb{E}\left(\int_0^T \psi(s, X_s, 0, 0)^2 ds\right) < \infty, \tag{2.20}$$

there exist a unique solution $(Y, Z) \in \mathscr{S}_T^2(\mathbb{R}) \times \mathscr{N}_W(0,T;L_2^0(V,\mathbb{R}))$ to equation (1.12) and there exists $C = C(K,T) > 0$ such that,

$$\|Y\|^2_{\mathcal{S}^2_T} + \|Z\|^2_{\mathcal{N}_W} \leq C\left(\mathbb{E}\phi(X_T)^2 + \mathbb{E}\int_0^T \psi(s, X_s, 0, 0)^2 ds\right). \tag{2.21}$$

PROOF. Again, we follow the proof given in [FT02, Proposition 4.3]. The following result is proven as the majority of existence of solutions to non-linear equations results, this is, by considering an adequate operator from a Banach space to itself and applying Banach's fixed point Theorem. For $\beta > 0$ consider $\mathscr{K}_\beta = \mathscr{S}_T^2(\mathbb{R}) \times \mathscr{N}_W(0,T; L_2^0(V,\mathbb{R}))$ which is a Banach space endowed with,

$$\begin{aligned}\|(Y,Z)\|_{\mathscr{K}_\beta}^2 &= \|Y\|_{\mathscr{S}_{T,\beta}^2}^2 + \|Z\|_{\mathscr{N}_{W,\beta}}^2 \\ &= \mathbb{E}\sup_{s\in[0,T]} e^{2\beta s}|Y_s|^2 + \mathbb{E}\int_0^T e^{2\beta s}\|Z_s\|_0^2\, ds.\end{aligned}$$

Let $\Psi\colon \mathscr{K}_\beta \to \mathscr{K}_\beta$ be defined as $\Psi(U,V) = (Y,Z)$ where (Y,Z) is such that,

$$Y_t + \int_t^T \langle Z_s, \cdot\rangle_0 dW_s = \phi(X_T) + \int_t^T \psi(s, X_s, U_s, V_s) ds.$$

Given $(U,V) \in \mathscr{K}_\beta$, $\Psi(U,V)$ is well-defined by Lemma 2.38 taking $(f_s)_{s\in[0,T]} = (\psi(s,X_s,U_s,V_s))_{s\in[0,T]}$ which is an element of $\mathscr{N}_W(0,T;\mathbb{R})$ due to the Lipschitz condition imposed on ψ and (2.20), the existence is proven if we show that Ψ is a contraction. Let $(U,V),(\bar U,\bar V),(Y,Z),(\bar Y,\bar Z) \in \mathscr{K}_\beta$ be such that $\Psi(U,V) = (Y,Z)$ and $\Psi(\bar U,\bar V) = (\bar Y,\bar Z)$, follows that for all $t\in[0,T]$,

$$Y_t - \bar Y_t + \int_t^T \langle Z_t - \bar Z_t, \cdot\rangle_0 dW_S = \int_t^T \left(\psi(s,X_s,U_s,V_s) - \psi(s,X_s,\bar U_s,\bar V_s)\right) ds.$$

This means that $(Y-\bar Y, Z-\bar Z)$ satisfies Lemma 2.38 with $\eta = 0$ and $f_s = \psi(s,X_s,U_s,V_s) - \psi(s,X_s,\bar U_s,\bar V_s)$. Thus

$$\begin{aligned}\left\|\Psi(U,V) - \Psi(\bar U,\bar V)\right\|_{\mathscr{K}_\beta}^2 &\le \frac{8K}{\beta}\mathbb{E}\int_0^T e^{2\beta s}\left(|U_s-\bar U_s|^2 + \left\|V_s-\bar V_s\right\|_0^2\right) ds \\ &\le \frac{8K}{\beta}\mathbb{E}\left(T\sup_{s\in[0,T]} e^{2\beta s}|U_s-\bar U_s|^2 + \int_0^T e^{2\beta s}\left\|V_s-\bar V_s\right\|_0^2 ds\right) \\ &\le \frac{8K(T+1)}{\beta}\left\|(U,V)-(\bar U,\bar V)\right\|_{\mathscr{K}_\beta}^2.\end{aligned}$$

By taking $\beta = 17K(T+1)$ we show that Ψ is a contraction, and therefore, the existence is proven. Uniqueness follows easily by standard arguments. Consider now the solution (Y,Z) by estimates (2.15),

$$\|Y\|_{\mathcal{S}_{T,\beta}^2}^2 + \|Z\|_{\mathcal{N}_{W,\beta}}^2 \le 16e^{2\beta T}\mathbb{E}\phi(X_T)^2 + \frac{8}{\beta}\underbrace{\mathbb{E}\int_0^T \psi(s,X_s,Y_s,Z_s)^2 ds}_{\mathbf{I}}.$$

Now, by the Lipschitz condition,

$$\begin{aligned}\mathbf{I} &\le 2K\mathbb{E}\int_0^T e^{2\beta s}\left(|Y_s|^2 + \|Z_s\|_0^2\right) + 2e^{2\beta T}\mathbb{E}\int_0^T \psi(s,X_s,0,0)^2 ds \\ &\le 2K(T+1)\left(\|Y\|_{\mathcal{S}_{T,\beta}^2}^2 + \|Z\|_{\mathcal{N}_{W,\beta}}^2\right) + 2e^{2\beta T}\mathbb{E}\int_0^T \psi(s,X_s,0,0)^2 ds.\end{aligned}$$

Hence,

$$\|Y\|^2_{\mathcal{S}^2_{T,\beta}} + \|Z\|^2_{\mathcal{N}_{W,\beta}} \leq 16e^{2\beta T}\mathbb{E}\phi(X_T)^2 + \frac{16}{\beta}e^{2\beta T}\mathbb{E}\int_0^T \psi(s, X_s, 0, 0)^2 ds + \frac{16K(T+1)}{\beta}\left(\|Y\|^2_{\mathcal{S}^2_{T,\beta}} + \|Z\|^2_{\mathcal{N}_{W,\beta}}\right).$$

Chosen β ensure that $16K(T+1)/\beta < 1$ and therefore,

$$\begin{aligned}\|Y\|^2_{\mathcal{S}^2_T} + \|Z\|^2_{\mathcal{N}_W} &\leq \|Y\|^2_{\mathcal{S}^2_{T,\beta}} + \|Z\|^2_{\mathcal{N}_{W,\beta}} \\ &\leq \left[1 - \frac{16K(T+1)}{\beta}\right]^{-1}\left(16e^{2\beta T}\mathbb{E}\phi(X_T)^2 + \frac{16}{\beta}e^{2\beta T}\mathbb{E}\int_0^T \psi(s, X_s, 0, 0)^2 ds\right).\end{aligned}$$

Hence, estimate (2.21) follows. The method that we have used remains valid if we intend to prove the existence of solutions $(Y, Z) \in \mathscr{S}^2_T(K) \times \mathscr{N}_W(0, T; L^0_2(V, K))$ and ψ, ϕ also taking values in the Hilbert space K. □

Previous proposition lets us state, given our assumptions (2.12), that from now on we can refer to a solution (X, Y, Z) of the system (1.11)-(1.12) with $(Y, Z) \in \mathscr{S}^2(\mathbb{R}) \times \mathscr{N}_W(0, T; L^0_2(V, \mathbb{R}))$ and X a strong solution of the forward equation (1.11) given by Proposition 2.37.

Chapter 3

Universal Approximation Theorems and Deep-H-Onets

In this chapter, our main objective is to obtain precise bounds on the terms ε_i^v, ε_i^z and ε_i^γ that appears on both main theorems. These bounds will be given in terms of (in)finite dimensional neural networks. Our main result for this section, Theorem 3.19, will provide the required control.

3.1 Finite Dimensional Neural Networks

The mathematical framework presented here is inspired by [HJKN20], we provide a slightly simpler development that adapts to our motivations. Finite dimensional Neural Networks are building blocks to their infinite dimensional version, we refer to this generalization as Infinite Dimensional NN (NN$^\infty$ for short). NNs are also used as an intermediate step in the proof of the Universal Approximation theorem for NN$^\infty$.

To fix ideas, in this section we focus on a setting where the input and output variables belong to multidimensional real spaces $\mathbb{R}^d$ and $\mathbb{R}^m$ respectively with $d, m \in \mathbb{N}$. The following definition is general and introduce the notion of finite dimensional Neural Network with an arbitrary activation function.

Definition 3.1 *Consider $L+1 \in \mathbb{N}$ as the number of layers within the network with $l_i \in \mathbb{N}$ neurons each for $i \in \{0, ..., L\}$ where $l_0 = d$ and $l_L = m$, weight matrices $\left\{W_i \in \mathbb{R}^{l_i \times l_{i-1}}\right\}_{i=1}^L$, bias vectors $\left\{b_i \in \mathbb{R}^{l_i}\right\}_{i=1}^L$, and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Let $\theta = (W_i, b_i)_{i=1}^L$, which can be seen as an element of $\mathbb{R}^\kappa$ with $\kappa = \sum_{i=1}^L (l_i l_{i-1} + l_i)$, and a function $\sigma : \mathbb{R} \to \mathbb{R}$. We define the neural network $f^{\theta,\sigma} : \mathbb{R}^{l_0} \to \mathbb{R}^{l_L}$ as the following composition,*

$$f^{\theta,\sigma}(x) = \left(A_L \circ \sigma \circ A_{L-1} \circ \cdots \circ A_2 \circ \sigma \circ A_1\right)(x),$$

where $A_i : \mathbb{R}^{l_{i-1}} \to \mathbb{R}^{l_i}$ is an affine linear function such that $A_i(x) = W_i x + b_i$ for $i \in \{1, ..., L\}$ and σ is applied component-wise. One says that the function $f^{\theta,\sigma}$ is the realization of the parameter θ as a NN. Numbers $(l_i)_{i \in \{0,...,L\}}$ represents the amount of units on each layer, note that the first

layer has $l_0 = d$ units and the last one has $l_L = m$ as they stand for the input and output variables respectively, the remaining $L-1$ layers are also known as hidden layers.

In previous definition, nothing prevent us from taking $L = 1$ and getting 0 hidden layers, observe that such neural network is just an affine linear function. We introduce some necessary conditions concerning activation functions. We follow the definitions given in [CC95].

Definition 3.2 *A function $\sigma : \mathbb{R} \to \mathbb{R}$ is called TW (Tauber-Wiener) if the set*

$$\left\langle \left\{ \sum_{i=1}^{N} c_i \sigma(\lambda_i x + \theta_i) \middle| \lambda_i, \theta_i, c_i \in \mathbb{R}\ i \in \{1, ..., N\} \right\} \right\rangle$$

is dense in $C([a, b])$ for $a, b \in \mathbb{R}$ and $a < b$.

From the definition it is not obvious how to determine if a function is TW, Chen and Chen [CC95, Theorem 1] provide us with a result that makes it easier to know.

Theorem 3.3 *Suppose that σ is a continuous function and that $\sigma \in S'(\mathbb{R})$, the set of tempered distribution. Then, σ is TW if and only if σ is not a polynomial.*

In this paper we work with an activation function known as ReLu denoted by $\sigma_{\text{ReLu}} : \mathbb{R} \to \mathbb{R}$ and is such that $\sigma_{\text{ReLu}}(x) = \max(x, 0)$ for all $x \in \mathbb{R}$. We can see that this function satisfies hypothesis of Theorem 3.3. In the following we make a formal definition of neural network and the set of parameters that defines them.

Definition 3.4 *The set of parameters of Neural Networks associated to $l_0 = d, l_L = m \in \mathbb{N}$ and a function $\sigma : \mathbb{R} \to \mathbb{R}$ is defined by,*

$$\mathcal{N}_{\sigma,L,d,m} = \bigcup_{\kappa \in \mathbb{N}} \mathcal{N}_{\sigma,L,d,m,\kappa}$$

where,

$$\mathcal{N}_{\sigma,L,d,m,\kappa} = \Big\{ \theta \in \mathbb{R}^{\kappa} \Big| \theta = \{W_i, b_i\}_{i=1}^{L},\ l_0 = d,\ l_L = m,\ W_i \in \mathbb{R}^{l_i \times l_{i-1}},\ b_i \in \mathbb{R}^{l_i},\ l_i \in \mathbb{N},$$
$$i \in \{1, ..., L\},\ \kappa = \sum_{i=1}^{L} (l_i l_{i-1} + l_i) \Big\}.$$

Naturally,

$$\mathcal{N}_{\sigma,d,m,\kappa} = \bigcup_{L \in \mathbb{N}} \mathcal{N}_{\sigma,L,d,m,\kappa} \quad \text{and} \quad \mathcal{N}_{\sigma,d,m} = \bigcup_{L \in \mathbb{N}} \bigcup_{\kappa \in \mathbb{N}} \mathcal{N}_{\sigma,L,d,m,\kappa}$$

Note that a parameter is eliminated when the union is taken over that parameter. For a set of parameters $\mathcal{N} \in \{\mathcal{N}_{\sigma,d,m}, \mathcal{N}_{\sigma,L,d,m}, \mathcal{N}_{\sigma,d,m,\kappa}\}$, the set of Neural Networks is then defined by,

$$\mathcal{R}(\mathcal{N}) = \left\{ f^{\theta,\sigma} \middle| \theta \in \mathcal{N} \right\}.$$

Here $f^{\theta,\sigma} : \mathbb{R}^d \to \mathbb{R}^m$.

Now, for completeness,we present two basic but important results. The first shows that the composition of two NNs produce another NN bellowing to certain space $\mathcal{N}_{\sigma,L,d,m}$ and he second proves that NNs have a growth that is controlled by its parameters and the activation function . We write $f^\theta = f^{\theta,\sigma}$ when no confusion arise.

Lemma 3.5 *Let $f^\gamma \in \mathcal{R}(\mathcal{N}_{\sigma,M,m,n})$ and $f^\theta \in \mathcal{R}(\mathcal{N}_{\sigma,L,d,m})$, then $f^\gamma \circ f^\theta \in \mathcal{R}(\mathcal{N}_{\sigma,L+M,d,n})$.*

PROOF. Let,

$$f^\gamma = B_M \circ \sigma \cdots \sigma \circ B_1$$
$$f^\theta = A_L \circ \sigma \cdots \sigma \circ A_1.$$

Then,

$$f^\gamma \circ f^\theta = B_M \circ \sigma \cdots \sigma \circ B_1 \circ A_L \circ \sigma \cdots \sigma \circ A_1.$$

Therefore the composition produce an additive property on the number of layers and $f^\gamma \circ f^\theta \in \mathcal{R}(\mathcal{N}_{\sigma,L+M,d,n})$. □

Previous lemma hints that the composition of NNs translate as a concatenation operation for its parameters, we introduce this notion in Definition 3.6:

Definition 3.6 *For σ, d, m we define the concatenation of parameters $\circ\colon \mathcal{N}_{\sigma,M,m,n} \times \mathcal{N}_{\sigma,L,d,m} \to \mathcal{N}_{\sigma,L+M,d,n}$ as,*

$$\{V_i, c_i\}_{i=1}^M \circ \{W_i, b_i\}_{i=1}^L = \{W_1, b_1, ..., W_L, b_L, V_1, c_1, \ldots, V_M, c_M\}. \tag{3.1}$$

Then we have that for $\theta \in \mathcal{N}_{\sigma,L,d,m}$ and $\gamma \in \mathcal{N}_{\sigma,M,m,n}$ $f^\theta \circ f^\gamma = f^{\theta\circ\gamma}$.

Remark 3.7 *Note that the order of composition at the left side of equation (3.1) differs from that of the right side. This is because the composition of functions is written in the opposite direction to the flow in a neural network (left to right).*

Lemma 3.8 *Assume that $|\sigma(a)| \leq |a|$ for any $a \in \mathbb{R}$. Let $\theta \in \mathcal{N}_{\sigma,L,d,m}$ with $L \geq 1$. Then there exist positive constants c_1, c_2, depending on θ, such that,*

$$\left\|f^{\theta,\sigma}(x)\right\|^2 \leq c_1 \|x\|^2 + c_2, \ \forall x \in \mathbb{R}^d.$$

PROOF. Let $A \in \mathbb{R}^{m\times n}$, here we denote $\|A\|^2 = \sum_{i=1,j=1}^{m,n} A_{i,j}^2$, the Frobenius matrix norm. We proceed by induction; the base case is $L = 1$, the NN takes the form of an affine linear function from $\mathbb{R}^d$ to $\mathbb{R}^m$ which satisfy the property. For the inductive step, let $\theta \in \mathcal{N}_{\sigma,L+1,d,m}$ and assume that that property holds for L, then

$$\begin{aligned}
\left\|f^\theta(x)\right\|^p &= \|(A_{L+1} \circ \sigma \circ \cdots \sigma \circ A_1)(x)\|^p \\
&= \left\|W_{L+1}\left(\left[\sigma \circ A_L \circ \cdots \circ A_1\right](x)\right) + b_{L+1}\right\|^p \\
&\leq 2^{p-1} d^{p/2} \|W_{L+1}\|^p \|(A_L \circ \sigma \circ \cdots \circ \sigma \circ A_1)(x)\|^p + 2^{p-1} \|b_{L+1}\|^p \\
&\leq 2^{p-1} d^{p/2} \|W_{L+1}\|^p \left(c_1^L \|x\|^p + c_2^L\right) + 2^{p-1} \|b_{L+1}\|^p .
\end{aligned}$$

Note that we applied the property for $\hat{\theta} = (W_i, b_i)_{i=1}^{L}$ with constants c_1^L and c_2^L. Now, defining

$$c_1^{L+1} = c_1^L 2^{p-1} d^{p/2} \|W_{L+1}\|^p \quad \text{and} \quad c_2^{L+1} = 2^{p-1} \left(d^{d/2} \|W_{L+1}\|^p c_2^L + \|b_{L+1}\|^p \right),$$

we finish the induction proof. □

If the activation function σ is continuous, the elements in $\mathcal{R}(\mathcal{N}_{\sigma,d,m})$ are continuous functions bellowing to $C(\mathbb{R}^d; \mathbb{R}^m)$. This is because they are composition of continuous mappings itself. Definition 3.4 is general, the first approximation theorem presented here is written a subset $\mathcal{H}$ of $\mathcal{N}_{\sigma,2,d,1}$ defined by

$$\mathcal{H} = \mathcal{N}_{\sigma,2,d,1} \cap \left\{ \theta \in \mathcal{N}_{\sigma,2,d,1} \mid \theta = \{W_1, b_1, W_2\} \in \mathbb{R}^{nd+n+n},\ b_2 = 0,\ n \in \mathbb{N} \right\}. \tag{3.2}$$

Note that in this definition the free parameter κ from definition 3.4 depends on the size $n \in \mathbb{N}$ of the first (and only) hidden layer in the following way, $\kappa = \sum_{i=1}^{2}(l_i l_{i-1} + l_i) = nd + n + n + 1$. It is straightforward that a function $f^{\theta,\sigma} \in \mathcal{R}(\mathcal{H})$, set of real-valued mappings, takes the following form

$$f^{\theta,\sigma}(x) = W_2 \cdot \sigma(W_1 x + b_1) = \sum_{i=1}^{n} W_{2,i} \sigma \left(\sum_{j=1}^{d} W_{2,i,j} x_j + b_{1,i} \right),$$

for $\theta = \{W_1, b_1, W_2\} \in \mathbb{R}^{nd+n+n}$, $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$. One of the first rigorous proof about the capabilities of Neural Networks is due to the work of Hornik in [Hor91] where the following result is given.

Theorem 3.9 ([Hor91], Theorem 1) *If $\sigma \colon \mathbb{R} \to \mathbb{R}$ is bounded and non-constant, then $\mathcal{R}(\mathcal{H})$ is dense in $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu; \mathbb{R})$ for every finite measure μ in $\mathbb{R}^d$.*

Let $m \in \mathbb{N}$. For a measure μ on $\mathbb{R}^d$, consider the space $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu; \mathbb{R}^m)$ of square integrable vector valued functions endowed with the norm

$$\|h\|^2_{L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu; \mathbb{R}^m)} = \int_{\mathbb{R}^d} \sum_{i=1}^{m} |h_i(x)|^2 \mu(dx)$$

for $h = (h_1, ..., h_m)$ and h_i a scalar function for $i \in \{1, ..., m\}$. We also need to approximate the derivative ∇u of the solution u to PIDE (1.3), the following proposition proves density of NNs in the space of square integrable vector valued functions taking advantage of Theorem 3.9.

Lemma 3.10 *Let $m \in \mathbb{N}$ with $m \geq 1$. If the activation function σ is bounded and non-constant, then $\mathcal{N}_{\sigma,2,d,m}$ is dense in $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu; \mathbb{R}^m)$ for every finite measure μ on $\mathbb{R}^d$.*

PROOF. Given $\varepsilon > 0$ and a function $h = (h_1, ..., h_m) \in L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu; \mathbb{R}^m)$ we need to find $f^{\theta,\sigma} = (f_1, ..., f_m) \in \mathcal{N}_{\sigma,2,d,m}$ such that

$$\int_{\mathbb{R}^d} |h(x) - f^{\theta,\sigma}(x)|^2 \mu(dx) < \varepsilon.$$

First, observe that $\mathcal{H} \subset \mathcal{N}_{\sigma,2,d,1}$ which implies, by using Theorem 3.9, that $\mathcal{N}_{\sigma,2,d,1}$ is also dense in $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu; \mathbb{R})$ and therefore for every $i \in \{1, ..., m\}$ we can find $f_i^{\theta^i,\sigma}(\cdot)$ with $\theta^i = (W_1^i, b_1^i, W_2^i, b_2^i)$ and $\kappa^i = n^i d + n^i + n^i + 1$, depending on ε, such that

$$\int_{\mathbb{R}^d} |h_i(x) - f_i^{\theta^i,\sigma}(x)|^2 \mu(dx) < \frac{\varepsilon}{m}.$$

Consider $f \in \mathcal{N}_{\sigma,2,d,m}$ defined by $\widehat{\theta} = \left(\widehat{W}_1, \widehat{b}_1, \widehat{W}_2, \widehat{b}_2\right)$ with

$$\widehat{W}_1 = \begin{pmatrix} W_1^1 \\ \vdots \\ W_1^m \end{pmatrix} \in \mathbb{R}^{\left(\sum_{i=1}^m n^i\right) \times d}, \quad \widehat{b}_1 = \begin{pmatrix} b_1^1 \\ \vdots \\ b_1^m \end{pmatrix} \in \mathbb{R}^{\sum_{i=1}^m n^i}$$

$$\widehat{W}_2 = \begin{pmatrix} W_2^{1,T} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & W_2^{m,T} \end{pmatrix} \in \mathbb{R}^{m \times \sum_{i=1}^m n^i}, \quad \widehat{b}_2 = \begin{pmatrix} b_2^1 \\ \vdots \\ b_2^m \end{pmatrix} \in \mathbb{R}^m,$$

and which satisfies that for $x \in \mathbb{R}^d$

$$f^{\widehat{\theta},\sigma}(x) = \widehat{W}_2 \phi(\widehat{W}_1 x + \widehat{b}_1) + \widehat{b}_2 = \begin{pmatrix} W_2^{1,T} \sigma(W_1^1 x + b_1^1) + b_2^1 \\ \vdots \\ W_2^{m,T} \sigma(W_1^m x + b_1^m) + b_2^m \end{pmatrix} = \begin{pmatrix} f_1^{\theta^1,\sigma}(x) \\ \vdots \\ f_m^{\theta^m,\sigma}(x) \end{pmatrix}.$$

Therefore,

$$\int_{\mathbb{R}^d} |h(x) - f^{\widehat{\theta},\sigma}(x)|^2 \mu(dx) = \int_{\mathbb{R}^d} \sum_{i=1}^m |h_i(x) - f_i^{\theta^i,\sigma}(x)| \mu(dx) < \varepsilon.$$

This ends the proof. □

Lemma 3.10 allows us to state that if we take some function $h : \mathbb{R}^d \to \mathbb{R}^m$ in $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu; \mathbb{R}^m)$, then the quantity

$$\inf_{\theta \in \mathbb{R}^\kappa} \int_{\mathbb{R}^d} |f^{\theta,\phi}(x) - h(x)|^2 \mu(dx) \tag{3.3}$$

can be made arbitrarily small by possible making κ growing sufficiently large, whenever μ is a finite measure on $\mathbb{R}^d$ and the activation function that defines the NN is bounded and non-constant.

The second universal approximation theorem that we present here, is due to Chen and Chen in [CC95, Theorem 3]. The main difference with Theorem 3.9 is the type of distance to measure de approximation.

Theorem 3.11 *Let K be a compact set in $\mathbb{R}^d$, U a compact set in $C(K)$ and $\sigma \colon \mathbb{R} \to \mathbb{R}$ a TW activation function. Then, for all $\varepsilon > 0$ there exists a parameter θ depending on $g \in U$ as $\theta(g) = \{W_1, b_1, W_2(g)\} \in \mathcal{H}$ such that*

$$\sup_{x \in K, g \in U} |g(x) - f^{\theta(g)}(x)| < \varepsilon.$$

In particular, the latter theorem states that $\mathcal{R}(\mathcal{H})$ is dense in $C(K)$ endowed with the uniform topology in the sense that for every ε there exist a NN with a sufficiently large hidden layer that meets the said accuracy in uniform distance. The following lemma extends Theorem 3.11 proving the density of $\mathcal{R}(\mathcal{N}_{\sigma,2,d,m})$ in $C(K,\mathbb{R}^m)$ for a compact $K\subset\mathbb{R}^d$ and $m\geq 1$.

Lemma 3.12 *Let $m\in\mathbb{N}$ with $m\geq 1$ and K a compact set in $\mathbb{R}^d$. If the activation function σ is TW, then $\mathcal{R}(\mathcal{N}_{\sigma,2,d,m})$ is dense in $C(K,\mathbb{R}^m)$.*

PROOF. Given $\varepsilon>0$ and a function $h=(h_1,...,h_m)\in C(K;\mathbb{R}^m)$ we need to find $f^{\theta,\sigma}=(f_1,...,f_m)\in\mathcal{R}(\mathcal{N}_{\sigma,2,d,m})$ such that

$$\sup_{x\in K}\left\|h(x)-f^{\theta,\sigma}(x)\right\|<\varepsilon.$$

First, observe that $\mathcal{R}(\mathcal{H})\subset\mathcal{R}(\mathcal{N}_{\sigma,2,d,1})$ which implies, by using Theorem 3.11, that $\mathcal{R}(\mathcal{N}_{\sigma,2,d,1})$ is also dense in $C(K)$ and therefore for every $i\in\{1,...,m\}$ we can find $f^{\theta^i,\sigma}$ with $\theta^i=\{W_1^i,b_1^i,W_2^i,b_2^i\}$ and $\kappa^i=n^id+n^i+n^i+1$, depending on ε, such that

$$\sup_{x\in K}|h_i(x)-f^{\theta^i,\sigma}(x)|<\frac{\varepsilon}{\sqrt{m}}.$$

Consider $\widehat{\theta}\in\mathcal{N}_{\sigma,2,d,m}$ with $\widehat{\theta}=\left\{\widehat{W}_1,\widehat{b}_1,\widehat{W}_2,\widehat{b}_2\right\}$ defined by

$$\widehat{W}_1=\begin{pmatrix}W_1^1\\ \vdots\\ W_1^m\end{pmatrix}\in\mathbb{R}^{\left(\sum_{i=1}^m n^i\right)\times d},\quad \widehat{b}_1=\begin{pmatrix}b_1^1\\ \vdots\\ b_1^m\end{pmatrix}\in\mathbb{R}^{\sum_{i=1}^m n^i}$$

$$\widehat{W}_2=\begin{pmatrix}W_2^{1,T} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & W_2^{m,T}\end{pmatrix}\in\mathbb{R}^{m\times\sum_{i=1}^m n^i},\quad \widehat{b}_2=\begin{pmatrix}b_2^1\\ \vdots\\ b_2^m\end{pmatrix}\in\mathbb{R}^m,$$

and which satisfies that for $x\in\mathbb{R}^d$

$$f^{\widehat{\theta},\sigma}(x)=\widehat{W}_2\sigma(\widehat{W}_1x+\widehat{b}_1)+\widehat{b}_2=\begin{pmatrix}W_2^{1,T}\sigma(W_1^1x+b_1^1)+b_2^1\\ \vdots\\ W_2^{m,T}\sigma(W_1^mx+b_1^m)+b_2^m\end{pmatrix}=\begin{pmatrix}f^{\theta^1,\sigma}(x)\\ \vdots\\ f^{\theta^m,\sigma}(x)\end{pmatrix}.$$

Therefore,

$$\sup_{x\in K}\left\|h(x)-f^{\widehat{\theta},\sigma}(x)\right\|=\sup_{x\in K}\left(\sum_{i=1}^m|h_i(x)-f^{\theta_i,\sigma}(x)|^2\right)^{1/2}<\varepsilon.$$

This ends the proof. □

The following lemma will be useful in the section devoted to NN$^\infty$, it is presented in [LMK21, Lemma C.1] as the Clipping lemma. Here we follow their proof as we need the explicit form of the NN given in the lemma.

Lemma 3.13 *Let $\varepsilon > 0$, $d \in \mathbb{N}$ and fix $0 < R_1 < R_2$. There exist a ReLu NN parameter $\theta \in \mathcal{N}_{\sigma_{ReLu},5,d,d}$, depending on ε and R_1, such that*

$$\begin{cases} \left\| f^\theta(x) - x \right\| < \varepsilon, & \|x\| \leq R_1, \\ \left\| f^\theta(x) \right\| < R_2, & \forall x \in \mathbb{R}^d. \end{cases}$$

Remark 3.14 *The previous lemma is used in the proof of more general universal approximation theorems (See the following section), therefore it force us to stick to ReLu NNs from now on.*

PROOF. For any $a \in \mathbb{R}$, $\vec{a}$ represents the vector $(a, \dots, a) \in \mathbb{R}^d$ and as we are only working with ReLu activation function, we drop the σ_{ReLu} from the NNs notation. Without loss of generality we may assume $\varepsilon < R_2 - R_1$. Consider $\gamma \colon \mathbb{R}^d \to [-R_1, R_1]^d$ defined for $x \in \mathbb{R}^d$ as $\gamma(x) = \min(\max(x, -R_1), R_1)$, which depends on R_1 and can be represented exactly by a ReLu NN in $\mathcal{N}_{\sigma_{\text{ReLu}},3,d,d}$ as,

$$\gamma(x) = -\max\left(-\max\left(x + \vec{R_1}, 0\right) + 2\vec{R_1}, 0\right) + \vec{R_1}.$$

Taking $\theta_\gamma = \left\{ I_d, \vec{R_1}, -I_d, 2\vec{R_1}, -I, \vec{R_1} \right\}$ follows that $\gamma = f^{\theta_\gamma}$. Note that for any $x \in [-R_1, R_1]^d$, $f^{\theta_\gamma}(x) = x$. The next step is to define a continuous function $\phi \colon \mathbb{R}^d \to \mathbb{R}^d$ by,

$$\phi(x) = \begin{cases} x, & \|x\| \leq R_1 \\ R_1 \frac{x}{\|x\|}, & \|x\| > R_1. \end{cases}$$

We have that $\phi \in C([-R_1, R_1]^d)$, then, by Theorem 3.12, there exists $f^{\theta_\varepsilon} \in \mathcal{N}_{\sigma_{\text{ReLu}},2,d,d}$ such that,

$$\sup_{x \in [-R_1, R_1]^d} \left\| \phi(x) - f^{\theta_\varepsilon}(x) \right\| < \varepsilon.$$

Define now $\theta = \theta_\varepsilon \circ \theta_{R_1}$, which is well defined and belong to $\mathcal{N}_{\sigma_{\text{ReLu}},5,d,d}$ by Lemma 3.5 and Definition 3.6. Then, for any $\|x\| \leq R_1$.

$$\left\| f^\theta(x) - x \right\| = \left\| f^{\theta_\varepsilon}(f^{\theta_\gamma}(x)) - \phi(x) \right\| = \left\| f^{\theta_\varepsilon}(x) - \phi(x) \right\| < \varepsilon,$$

and,

$$\sup_{x \in \mathbb{R}^d} \left\| f^\theta(x) \right\| \leq \sup_{x \in [-R_1, R_1]^d} \left\| f^{\theta_\varepsilon}(x) - \phi(x) \right\| + R_1 < R_2.$$

This finishes the proof. □

3.2 Infinite Dimensional Neural Networks: Hilbert-valued Deep-Onets

In this section we work with a particular type of NN^∞ called DeepOnets. Based on the definitions given in [LMK21], we provide a proper and rigorous treatment of this object and prove important results that allows them to be used on our PDE and stochastic setting.

Through this entire section $(H, \langle\cdot,\cdot\rangle_H, \|\cdot\|_H)$ and $(W, \langle\cdot,\cdot\rangle_W, \|\cdot\|_W)$ will denote separable Hilbert space with orthonormal basis $(e_i)_{i\in\mathbb{N}}$ and $(g_i)_{i\in\mathbb{N}}$ respectively, H is equipped with a Borel probability measure μ. In the following we are devoted to study the approximation of functionals of the form $F\colon H \to W$ by functions parameterized by finite dimensional parameters. The main idea to define such functions is to take a sufficiently large $d \in \mathbb{N}$ such that the approximations $\sum_{i=1}^{d}\langle x, e_i\rangle_H e_i$ are good enough to approximate $x \in H$ and encode x as the vector $(\langle x, e_1\rangle_H, ..., \langle x, e_n\rangle_H) \in \mathbb{R}^d$, then use a finite dimensional neural network to go from $\mathbb{R}^d$ to $\mathbb{R}^m$ for some $m \in \mathbb{N}$. At last, we take the resulting vector to W by considering its m components as coefficients for $\{g_1, ..., g_m\}$. The structure of Hilbert spaces allow us to take advantage of results such as Lemma 3.15, which we present below with a proof due to Aris Daniilidis. Note that it is valid for every Hilbert space.

Lemma 3.15 (Daniilidis) *Let K be a compact set on H. For every $k \in \mathbb{N}$ consider the operator $P_k : H \to H$ defined as $P_k(x) = \sum_{i=1}^{k}\langle x, e_i\rangle_H e_i$ for $x \in H$. Then, for every $\varepsilon > 0$ there exists $k \in \mathbb{N}$ such that for all $x \in K$,*

$$\|P_k x - x\|_H \leq \varepsilon.$$

PROOF. First, lets establish that for all $k \in \mathbb{N}$, $P_k \in L(H)$ and $\|P_k\|_H \leq 1$. P_k is clearly linear, to prove the bound let x be any non-zero vector in H,

$$\|P_k x\|_H^2 = \left\|\sum_{i=1}^{k}\langle x, e_i\rangle_H e_i\right\|_H^2 = \sum_{i=1}^{k}|\langle x, e_i\rangle_H|^2 \leq \sum_{i=1}^{\infty}|\langle x, e_i\rangle_H|^2 = \|x\|_H^2 .$$

This means that $\|P_k\|_{L(H)} \leq 1$.

We argue by contradiction. Suppose that there exists $\varepsilon > 0$ such that for all $n \in \mathbb{N}$ we can find $x_n \in K$ verifying $\|P_n(x_n) - x_n\|_H \geq \varepsilon$. Due to the compactness of K, there is a subsequence that converges to some $x \in H$, we denote this subsequence as x_n as well. Then,

$$\begin{aligned}\|P_n(x_n) - x_n\|_H &\leq \|P_n(x_n) - P_n(x)\|_H + \|P_n(x) - x\|_H + \|x - x_n\|_H \\ &\leq 2\|x_n - x\|_H + \|P_n(x) - x\|_H .\end{aligned}$$

The first term can be made as small as we want due to the convergence of x_n to x and the second because we have that $P_n(x) \to x$ in H as $n \to \infty$. Then, for some large n we can break the bound and thus, the contradiction. □

From now on we fix $\sigma = \sigma_{\text{ReLu}}$.

Definition 3.16 *Recall Definition 3.4. Given $L, d, m \in \mathbb{N}$ consider the functions*

$$\begin{aligned}\mathcal{E}_{H,d} : H &\longrightarrow \mathbb{R}^d \\ x &\longmapsto \Big(\langle x, e_i\rangle_H\Big)_{i=1}^{d},\end{aligned} \qquad \begin{aligned}\widehat{\mathcal{E}}_{W,m} : \mathbb{R}^m &\longrightarrow W \\ a &\longmapsto \sum_{i=1}^{m} a_i g_i.\end{aligned}$$

Let $\theta \in \mathcal{N}_{\sigma,L,d,m}$, for (H, d, θ, m, W) we define the DeepOnet $F^{H,d,\theta,m,W} : H \to W$ by

$$F^{H,d,\theta,m,W} = \widehat{\mathcal{E}}_{W,m} \circ f^{\theta} \circ \mathcal{E}_{H,d}. \tag{3.4}$$

Unless is extremely necessary, we omit H, W and just use $F^{d,\theta,m}$. Also, define the following sets of DeepOnets parameters,

$$\mathcal{N}_{\sigma}^{H\to W} = \bigcup_{d,m\in\mathbb{N}} \{d\} \times \mathcal{N}_{\sigma,d,m} \times \{m\},$$
$$\mathcal{N}_{\sigma,L}^{H\to W} = \bigcup_{d,m\in\mathbb{N}} \{d\} \times \mathcal{N}_{\sigma,L,d,m} \times \{m\}.$$

With $L \in \mathbb{N}$, observe that $\mathcal{N}_{\sigma,L}^{H\to W} \subset \mathcal{N}_{\sigma}^{H\to W}$ (the less parameters specified, the bigger the set). Let $\mathcal{N} = \mathcal{N}_{\sigma}^{H\to W}$ or $\mathcal{N} = \mathcal{N}_{\sigma,L}^{H\to W}$, it is straightforward to define,

$$\mathcal{R}(\mathcal{N}) = \left\{ F^{H,d,\theta,m,W} \;\middle|\; (d,\theta,m) \in \mathcal{N} \right\}.$$

Note that d is not readable as an input dimension, here it becomes a parameter of the DeepOnet and represents how many elements of the base $(e_i)_{i\in\mathbb{N}}$ we are using to project with in order to get the finite dimensional representation $(\langle x, e_i\rangle_H)_{i=1}^d$ for $x \in H$. Last action is carried out by mapping $\mathcal{E}_{H,d}$. The same goes for m but in the opposite direction and in this case, it is done by $\widehat{\mathcal{E}}_{W,m}$, which allows us to take a collection of real numbers to a Hilbert space. Observe that functions in $\mathcal{R}(\mathcal{N})$ are continuous because they are composition of continuous functions itself.

Remark 3.17 *We remark the following,*

- *We have that $\mathcal{E}_{\mathbb{R}^d,d} = I_d$ and $\widehat{\mathcal{E}}_{\mathbb{R}^d,d} = I_d$. Note that with this consideration we recover the finite dimensional theory by taking $H = \mathbb{R}^d$ and $W = \mathbb{R}^m$.*
- *We could just denote $F^{d,\theta,m}$ as F^{θ} because the information about the input and output dimension of the NN is codified in the parameter θ, but we decide to specify d, m for a better understanding. Also the order of the parameters makes clearer in which order the composition are taken.*
- *Note that the number of parameters to define a DeepOnet is the same as of NNs only adding d, m.*
- *If H is a functional space such as $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), dx)$, DeepOnets also admits a "neural network representation" where the first layer is in some sense dense as has an infinite number of units which are all captured by $\langle\cdot,\cdot\rangle$ to be transferred to the next finite layer.*

Proposition 3.18 (See e.g. Theorem 4 in [CC95]) *Let $m \in \mathbb{N}$, $K \subset H$ be a compact set and $f : K \to \mathbb{R}^m$ be a continuous function. Then, for any $\varepsilon > 0$ there exists $(d,\theta,m) \in \mathcal{N}_{\sigma,2}^{H\to\mathbb{R}^m}$ such that,*

$$\sup_{x\in K} \left\| F^{d,\theta,m}(x) - f(x) \right\| \leq \varepsilon.$$

In other words, $\left\{ F|_K : F \in \mathcal{R}(\mathcal{N}_{\sigma,2}^{\infty}) \right\}$ is dense in $C(K)$ endowed with the uniform norm.

PROOF. Consider the operators P_k from Lemma 3.15. The said Lemma tells us that for $\delta_k \searrow 0$ we can find a set of natural numbers $(n_k := n(\delta_k))_{k\in\mathbb{N}}$ such that,

$$\forall k \in \mathbb{N},\ \forall u \in K, \|P_{n_k}(u) - u\|_H < \delta_k.$$

Given the continuity of P_k, $P_k(K)$ is also a compact set in H for all $k \in \mathbb{N}$. Now we prove that the set

$$A := \left(\bigcup_{k=1}^{\infty} P_{n_k}(K)\right) \cup V,$$

is also compact in H. Indeed, let $(x_i)_{i\in\mathbb{N}}$ be a sequence in A. If there exists a subsequence such that it remains in K, there is nothing to prove because K is compact. The other case is that we can extract an infinite subsequence that lies in the infinite union. This means that there exists $(k_i)_{i\in\mathbb{N}} \subset \mathbb{N}$ and $(u_i)_{i\in\mathbb{N}} \subset K$ such that,

$$x_i = \sum_{j=1}^{n_{k_i}} \langle u_i, e_j\rangle e_j.$$

Due to compactness of K, up to a subsequence that we also denote $(u_i)_{i\in\mathbb{N}}$ as well, $(u_i)_{i\in\mathbb{N}}$ converges to some $u \in K$. We have two options, the first is that the sequence $(k_i)_{i\in\mathbb{N}}$ does not grow to infinite when $i \nearrow \infty$ and thus, up to a subsequence on i, we can find ι such that $\forall i \geq \iota$, $k_i = k_\iota$ which implies that, for $i \geq \iota$,

$$x_i = \sum_{j=1}^{n_{k_\iota}} \langle u_i, e_j\rangle_H e_j \longrightarrow \sum_{j=1}^{n_{k_\iota}} \langle u, e_j\rangle_H e_j \in P_{n_{k_\iota}} \subset A.$$

The second option is that up to a subsequence, $k_i \nearrow \infty$ as $i \nearrow \infty$, note that

$$x_i = \sum_{j=1}^{n_{k_i}} \langle u_i, e_j\rangle_H e_j = P_{n_{k_i}}(u_i),$$

and then,

$$\|x_i - u\|_H \leq \left\|P_{n_{k_i}}(u_i) - u_i\right\|_H + \|u_i - u\|_H \leq \delta_{k_i} + \|u_i - u\|_H,$$

where, taking $i \to \infty$ we prove that, up to a subsequence, $x_i \longrightarrow u \in K \subset A$. Thus, A is compact in H.

The next step is to use the well-known Tietze-Urysohn theorem [Mun00, Chapter 4, Theorem 35.1] which gives us a continuous extension $f_{\text{ex}} : A \to \mathbb{R}^m$ with $f_{\text{ex}}(x) = f(x)$ for $x \in K$. The compactness of A implies that f_{ex} is uniformly continuous, then, for $\varepsilon > 0$ we can find $\delta > 0$ depending only on ε such that $\|x - y\|_H < \delta$ implies $\|f_{\text{ex}}(x) - f_{\text{ex}}(y)\| < \varepsilon$. Lets fix $k \in \mathbb{N}$ such that $\delta_k < \delta$, let $F\colon K \to \mathbb{R}^m$ be a function to be specified later and x any element of K, then

$$\|f(x) - F(x)\| \leq \|f_{\text{ex}}(x) - f_{\text{ex}}(P_{n_k}(x))\| + \|f_{\text{ex}}(P_{n_k}(x)) - F(x)\| < \frac{\varepsilon}{2} + \|f_{\text{ex}}(P_{n_k}(x)) - F(x)\|.$$

By the continuity of $\mathcal{E}_{H,n_k}$ follows that $\mathcal{E}_{H,n_k}(K)$ is a compact set in $\mathbb{R}^{n_k}$. Consider the function $\bar{f}$ defined by

$$\begin{aligned}\bar{f}\colon \mathcal{E}_{H,n_k}(K) &\longrightarrow \mathbb{R}^m \\ y \quad &\longmapsto \bar{f}(y) = f_{\text{ex}}\left(\sum_{j=1}^{n_k} y_j e_j\right).\end{aligned}$$

Note that the extension is essential because $\mathcal{E}_{H,n_k}$ could not be a subset of K, where f is defined. By the universal approximation Theorem 3.12 there exists $\theta \in \mathcal{N}_{\sigma,2,n_k,m}$ such that

$$\begin{aligned}\sup_{y\in\mathcal{E}_{H,n_k}(K)} \left\|\bar{f}(y) - f^\theta(y)\right\| &= \sup_{y\in\mathcal{E}_{H,n_k}(K)} \left\| f_{\text{ex}}\left(\sum_{i=1}^{n_k} y_i e_i\right) - f^\theta(y)\right\| \\ &= \sup_{x\in K} \left\| f_{\text{ex}}\left(\sum_{i=1}^{n_k} \langle x, e_i\rangle_H e_i\right) - f^\theta\left(\left(\langle x, e_i\rangle_H\right)_{i=1}^{n_k}\right)\right\| \\ &= \sup_{x\in K} \left\| f_{\text{ex}}\left(P_{n_k}(x)\right) - \left(\widehat{\mathcal{E}}_{\mathbb{R}^m,m} \circ f^\theta \circ \mathcal{E}_{H,n_k}\right)(x)\right\| < \frac{\varepsilon}{2}.\end{aligned}$$

Recall the first point in Remark 3.17. It suffices to take $(n_k, \theta, m) \in \mathcal{N}_{\sigma,2}^{W\to\mathbb{R}^m}$ which concludes the proof. □

The main result of this section, concerning the approximation of a square integrable functional, is presented below and is closely related to the approximation of PDEs by DL techniques. We divide the proof in steps for a clear reading and follow the lines of [LMK21, Theorem 3.1].

Theorem 3.19 *Let $(W, \langle\cdot,\cdot\rangle_W, \|\cdot\|_W)$ be a separable Hilbert space with orthonormal basis $(g_i)_{i\in\mathbb{N}}$. Let $G\colon H \to W$ be a $L^2(H,\mu;W)$ mapping. Then, for any $\varepsilon > 0$ there exist a DO $F^{d,\theta,m} : H \to W$ such that,*

$$\int_H \left\|G(x) - F^{d,\theta,m}(x)\right\|_W^2 \mu(dx) \le \varepsilon.$$

PROOF. **Step 1.** Let $\varepsilon > 0$ and define $\delta = \sqrt{\varepsilon/8}$. First we prove that without loss of generality we can assume that G is bounded. Consider $M > 0$ and

$$G_M(x) := \begin{cases} G(x), & \|G(x)\|_W \le M \\ M\dfrac{G(x)}{\|G(x)\|_W}, & \sim \end{cases}$$

Then, for any function $F\colon H \to W$ we get,

$$\|G - F\|_{L^2(H,\mu;W)} \le \|G - G_M\|_{L^2(H,\mu;W)} + \|G_M - F\|_{L^2(H,\mu;W)}.$$

We have that $\|G_M - G\|_W^2 \to 0$ and $\|G_M - G\|_W^2 \le 4\|G\|_W^2$ μ-a.e., so applying dominate convergence theorem we take M such that,

$$\|G - F\|_{L^2(H,\mu;W)} \le \delta + \|G_M - F\|_{L^2(H,\mu;W)}.$$

Then, assuming $\|G\|_W \leq M$ on H, we prove that $\|G - F\|_{L^2(H,\mu;W)} < \delta$ for certain DeepOnet F.

Step 2. By Lusin's ([Bog07]) theorem, there exists a compact set $K = K(\delta, M) \subset H$ such that $G|_K$ is continuous and $\mu(H \setminus K) < \frac{\delta^2}{M^2}$. Now, consider the compact set $K' = G(K) \subset W$. In virtue of Lemma 3.15, there exist $\kappa = \kappa(K') \in \mathbb{N}$ such that,

$$\sup_{w \in K'} \|w - P_\kappa(w)\|_W \leq \delta.$$

Let $\widetilde{G} = P_\kappa \circ G \colon K \to W$. Note that,

$$\sup_{x \in K} \left\| G(x) - \widetilde{G}(x) \right\|_W = \sup_{w \in K'} \|w - P_\kappa(w)\|_W \leq \delta.$$

Step 3. Applying Proposition 3.18 for the continuous function $\mathcal{E}_{W,\kappa} \circ \widetilde{G} \colon K \to \mathbb{R}^\kappa$, we can take $(d, \theta_1, \kappa) \in \mathcal{N}_{\sigma,2}^{H \to \mathbb{R}^\kappa}$ such that,

$$\sup_{x \in K} \left\| F^{H,d,\theta_1,\kappa,\mathbb{R}^\kappa}(x) - (\mathcal{E}_{W,\kappa} \circ \widetilde{G})(x) \right\| < \delta.$$

Take any $x \in K$ and the DO generated by $(H, d, \theta_1, \kappa, W)$,

$$\begin{aligned}
\left\| F^{(H,d,\theta_1,\kappa,W)}(x) - \widetilde{G}(x) \right\|_W &= \left\| (\widehat{\mathcal{E}}_{W,\kappa} \circ f^\theta \circ \mathcal{E}_{H,d})(x) - \widetilde{G}(x) \right\|_W \\
&= \left\| \sum_{i=1}^{\kappa} (f^\theta \circ \mathcal{E}_{H,d})(x)_i g_i - \sum_{i=1}^{\kappa} \langle G(x), g_i \rangle_W g_i \right\|_W \\
&= \left\| (f^\theta \circ \mathcal{E}_{H,d})(x) - (\langle G(x), g_i \rangle_W)_{i=1}^{\kappa} \right\|_{\mathbb{R}^\kappa} \\
&= \left\| F^{H,d,\theta_1,\kappa,\mathbb{R}^\kappa}(x) - (\mathcal{E}_{H,\kappa} \circ \widetilde{G})(x) \right\|_{\mathbb{R}^\kappa} < \delta. \qquad (3.5)
\end{aligned}$$

Then, by using previous estimate, Lemma 3.15 and that G is bounded, one has the following bound

$$\left\| F^{H,d,\theta_1,\kappa,W}(x) \right\|_W \leq \left\| F^{H,d,\theta_1,\kappa,W}(x) - \widetilde{G}(x) \right\|_W + \left\| \widetilde{G}(x) - G(x) \right\|_W + \|G(x)\|_W < 2\delta + M.$$

Step 4. Applying the clipping Lemma 3.13 with δ, κ, $R_1 = M + 2\delta$ and $R_2 = 2M$, note that we can assume δ small enough such that $R_1 < R_2$, we can take $\theta_2 \in \mathcal{N}_{\sigma,5,\kappa,\kappa}$ such that,

$$\begin{cases} \left\| f^{\theta_2}(x) - x \right\| < \delta, & \|x\| < M + 2\delta \\ \left\| f^{\theta_2}(x) \right\| \leq 2M, & \forall x \in \mathbb{R}^\kappa. \end{cases} \qquad (3.6)$$

Recall that the norm used in previous equation is the usual norm in $\mathbb{R}^\kappa$ and that during this entire section, $\sigma = \sigma_{\text{ReLu}}$. Consider the following composition and its equivalences,

$$\widehat{\mathcal{E}}_{W,\kappa} \circ f^{\theta_2} \circ \widehat{\mathcal{E}}_{\mathbb{R}^\kappa} \circ f^{\theta_1} \circ \mathcal{E}_{H,d} = \widehat{\mathcal{E}}_{W,\kappa} \circ f^{\theta_2 \circ \theta_1} \circ \mathcal{E}_{H,d} = F^{H,d,\theta_1 \circ \theta_2,\kappa,W}.$$

Where we made use of Definition 3.6. Such DO satisfies the following,

$$\begin{aligned}
\left\| F^{H,d,\theta_2 \circ \theta_1,\kappa,W}(x) - \widetilde{G}(x) \right\|_W &\leq \left\| F^{H,d,\theta_2 \circ \theta_1,\kappa,W}(x) - F^{H,d,\theta_1,\kappa,W}(x) \right\|_W + \left\| F^{H,d,\theta_1,\kappa,W}(x) - \widetilde{G}(x) \right\|_W \\
&\leq \left\| \sum_{i=1}^{\kappa} f_i^{\theta_2} \left(f^{\theta_1} \left(\mathcal{E}_{H,d}(x) \right) \right) g_i - \sum_{i=1}^{\kappa} \left(f^{\theta_1} \circ \mathcal{E}_{H,d} \right)_i (x) g_i \right\|_W + \delta \\
&\leq \left\| f^{\theta_2} \left(f^{\theta_1} \left(\mathcal{E}_{H,d}(x) \right) \right) - f^{\theta_1} \left(\mathcal{E}_{H,d}(x) \right) \right\|_{\mathbb{R}^\kappa} + \delta < 2\delta,
\end{aligned}$$

where we used estimates (3.5) and (3.6).

Step 5. Now we use all previous bounds, let $F = F^{H,d,\theta_2\circ\theta_1,\kappa,W}$ with $(d, \theta_2 \circ \theta_1, \kappa) \in \{d\} \times \mathcal{N}_{\sigma,7,d,\kappa} \times \{\kappa\}$, then

$$\begin{aligned}\int_H \|G(x) - F(x)\|_W^2 \,\mu(dx) &= \int_{H\setminus K} \|G(x) - F(x)\|_W^2 \,\mu(dx) + \int_K \|G(x) - F(x)\|_W^2 \,\mu(dx)\\ &\leq 2\int_{H\setminus K} \|G(x)\|_W^2 \,\mu(dx) + 2\int_{H\setminus K} \|F(x)\|_W^2 \,\mu(dx)\\ &\quad + 2\int_K \left\|G(x) - \widetilde{G}(x)\right\|_W^2 \mu(dx) + 2\int_K \left\|\widetilde{G}(x) - F(x)\right\|_W^2 \mu(dx)\\ &\leq \mu\left(H \setminus K\right)\left(2M^2 + 2M^2\right) + 2\delta^2 + 2\delta^2 \leq 8\delta^2 = \varepsilon,\end{aligned}$$

which is the desired conclusion. □

Note that the theorem above only contribute with the existence of a parameter (d, θ, m) such that the generated DO is a good approximation, in order to overcome the said *curse of dimensionality* we may have to provide proper bounds on the size of (d, θ, m). Following lemma provides us with a useful bound for DeepOnets.

Remark 3.20 *Recall the notation from Step* 5 *from the proof above. Given the parameters* $(d, \theta_2 \circ \theta_1, \kappa) \in \{d\} \times \mathcal{N}_{\sigma,7,d,\kappa} \times \{\kappa\}$*, we have that* $\theta_2 \circ \theta_1 \in \mathbb{R}^\eta$ *for some* $\eta \in \mathbb{N}$*; therefore*

$$\inf_{(p,\theta,q)\in\mathbb{N}\times\mathbb{R}^\eta\times\mathbb{N}} \int_H \left\|G(x) - F^{p,\theta,q}(x)\right\|_W^2 \mu(dx) \leq \int_H \left\|G(x) - F^{d,\theta_2\circ\theta_1,\kappa}(x)\right\|_W^2 \mu(dx) \leq \varepsilon. \quad (3.7)$$

This observation allows us to state that for any $\varepsilon > 0$ *we can find a sufficiently large* $\eta \in \mathbb{N}$ *such that the left side of* (3.7) *is bounded by* ε*.*

Lemma 3.21 *Let* $p \geq 2$ *and* $(d, \theta, m) \in \mathcal{N}_{\sigma,2}^{H\to W}$*, then there exists* $c_1, c_2 > 0$ *such that* $|F^{\theta,d}(x)|^p \leq c_1 \|x\|_H^p + c_2$ *for every* $x \in H$*.*

PROOF. Let $x \in H$, then by using Lemma 3.8 there exists $a_1, a_2 > 0$ such that,

Defining $c_1 = 2^{\frac{p-2}{2}} a_1^{p/2}$ and $c_2 = 2^{\frac{p-2}{2}} a_2^{p/2}$ concludes the proof. □

Chapter 4

Non-local Kolmogorov equation on finite dimensions

In this chapter we introduce the scheme for the Kolomogorov equation introduced in Section 1.1.2. In Section 4.1 we derive an iterative scheme based on NNs, this scheme is intended to approximate strong solution of (1.3). In Section 4.2 we define auxiliary processes that are necessary for starting the proof of the main result and prove properties of these. Finally, Section 4.3 is devoted to the consistency proof of the presented scheme.

4.1 Numerical scheme

By applying Itô's lemma (see [Del13, Thm 2.3.4]) to the solution X_t in (1.6) and a $\mathcal{C}^{1,2}([0,T]\times\mathbb{R}^d)$ solution u of PIDE (1.3) as Y_t in (1.7), we obtain the compact stochastic formulation of (1.3):

$$\begin{aligned} u(t,X_t) = u(0,X_0) &- \int_0^t f(s,X_{s^-},u(s,X_{s^-}),\sigma(X_{s^-})\nabla u(s,X_{s^-}),\mathcal{I}[u](s,X_{s^-}))ds, \\ &+ \int_0^t [\sigma(X_{s^-})\nabla u(s,X_{s^-})]\cdot dW_s + \int_0^t\int_{\mathbb{R}^d}[u(s,X_{s^-}+\beta(X_{s^-},y))-u(s,X_{s^-})]\overline{\mu}(ds,dy), \end{aligned} \tag{4.1}$$

valid for $t\in[0,T]$. This tells us that whatever we use as approximations of

$$u(t,X_t), \qquad \sigma(X_t)\nabla u(t,X_t) \qquad \text{and} \qquad u(t,X_t+\beta(X_t,\cdot))-u(t,X_t),$$

must satisfy (4.1) in some proper metric. An important statement here is that the conditions (2.10) ensure the existence of a *viscosity solution* $u\in\mathcal{C}([0,T]\times\mathbb{R}^d)$ with at most polynomial growth such that $u(t,X_t)=Y_t$ (see [BBP97, Thm 3.4]), and this is the reason why our scheme seek to approximate the solution to the FBSDEJ (1.6)-(1.7). Recall Chapter 3 where we introduce Neural Networks.

From now on, fix a constant step partition of the interval $[0,T]$, defined as $\pi=\left\{\frac{iT}{N}\right\}_{i\in\{0,\dots,N\}}$, $t_i=\frac{iT}{N}$, and set $\Delta W_i = W_{t_{i+1}}-W_{t_i}$. Also, define $h:=\frac{T}{N}$ and (with a slight abuse of notation),

$\Delta t_i = (t_i, t_{i+1}]$. Recall the compensated measure $\overline{\mu}$ from Definition 2.25. Let

$$M_t := \overline{\mu}((0,t], \mathbb{R}^d) \quad \text{and} \quad \Delta M_i =: \overline{\mu}((t_i, t_{i+1}], \mathbb{R}^d) = \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \overline{\mu}(ds, dy). \tag{4.2}$$

It is well-known that an Euler scheme for the first equation in (1.6) obeys the form

$$X_0^\pi = x, \tag{4.3}$$

$$X_{t_{i+1}}^\pi = X_{t_i}^\pi + b(X_{t_i}^\pi)h + \sigma(X_{t_i}^\pi)\Delta W_i + \int_{\mathbb{R}^d} \beta(X_{t_i}, y)\overline{\mu}((t_i, t_{i+1}], dy). \tag{4.4}$$

Note that this scheme neglects the left limits that appears on the original equation, although, it satisfies the next error bound (see [Del13, Thm. 5.1.1], [BE08] or [GS21a]),

$$\max_{i=1,\dots,N} \mathbb{E}\left(\sup_{t\in[t_i,t_{i+1}]} \|X_t - X_{t_i}^\pi\|^2 \right) = O(h). \tag{4.5}$$

Under suitable conditions, mostly Lipschitz and linear growth assumptions, it can be proved that the constant behind $O(h)$ in (4.5) does not depend exponentially on d, see Lemma 4.3 in [GS21a]. Adapting the argument of [HPW19] to the non-local case, and in view of (4.1), we propose the following modified Euler scheme: for $i = 0, 1, \dots, N$,

$$u(t_{i+1}, X_{t_{i+1}}^\pi) \approx F_i\Big(t_i, X_{t_i}^\pi, u(t_i, X_{t_i}^\pi), \sigma(X_{t_i}^\pi)\nabla u(t_i, X_{t_i}^\pi), u(t_i, X_{t_i}^\pi + \beta(X_{t_i}^\pi, \cdot)) - u(t_i, X_{t_i}^\pi), h, \Delta W_i\Big),$$

where $F_i : \Omega \times [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda) \times \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$F_i(\omega, t, x, y, z, \psi, h, w) := y - hf\left(t, x, y, z, \int_{\mathbb{R}^d} \psi(y)\lambda(dy)\right) + w \cdot z + \int_{\mathbb{R}^d} \psi(y)\bar{\mu}\left((t_i, t_{i+1}], dy\right).$$

Note that ω is passed to F_i through its dependence on the compensated measure $\bar{\mu}$. The function F_i is, indeed, a random variable.

Remark 4.1 *Note that the nonlocal term in (1.3) forces us to define F_i in such a way that its fifth argument must be a function ψ in $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)$. In view of the integrals involved in F_i, it appears that we are again facing the same high dimensional problem; however this problem may be instead treated with Monte Carlo approximations, see below.*

Remark 4.2 *In the nonlocal setting, the function F_i also depends on the time interval $(t_i, t_{i+1}]$ in terms of the integrated measure $\bar{\mu}\left((t_i, t_{i+1}], dy\right)$. This is an important change in the Euler scheme, since we do not approximate the nonlocal term at time t_i in this case, but instead take into account how the measure $\bar{\mu}$ behaves on the time interval $(t_i, t_{i+1}]$.*

Recall Theorem 3.19 and let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a ReLu activation function. From now on we will be using NNs with a single hidden layer parameterized by $\theta \in \Xi$, where $\Xi = \mathbb{R}^\kappa$ for some free parameter $\kappa \in \mathbb{N}$ depending on the size of the hidden layer. For every time t_i on the grid consider,

$$u_i^\theta : \mathbb{R}^d \to \mathbb{R} \tag{4.6}$$

$$z_i^\theta : \mathbb{R}^d \to \mathbb{R}^d \tag{4.7}$$

$$w_i^\theta : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \tag{4.8}$$

with $u_i^\theta \in \mathcal{N}_{\phi,2,d,1,\kappa}$, $z_i^\theta \in \mathcal{N}_{\phi,2,d,d,\kappa}$ and $w_i^\theta \in \mathcal{N}_{\phi,2,d+d,1,\kappa}$ approximating

$$(u(t_i,\cdot),\ \sigma(\cdot)\nabla u(t_i,\cdot),\ u(t_i,\cdot+\beta(\cdot,\circ))-u(t_i,\cdot)),$$

respectively, in some sense to be specified. Let also

$$\langle w_i^\theta\rangle(x) = \int_{\mathbb{R}^d} w_i^\theta(x,y)\lambda(dy). \tag{4.9}$$

We propose an extension of the DBDP1 algorithm presented on [HPW19]. The main idea of the algorithm is that the NNs, evaluated on $X_{t_i}^\pi$, are good approximations of the processes solving the FBSDEJ. Let $\widehat{u}_{i+1}$ be the optimal approximation for step $i+1$ and let L_i be a cost function defined for $\theta \in \Xi$ as

$$L_i(\theta) = \mathbb{E}\left|\widehat{u}_{i+1}(X_{t_{i+1}}^\pi) - F_i(t_i, X_{t_i}^\pi, u_i^\theta(X_{t_i}^\pi), z_i^\theta(X_{t_i}^\pi), w_i^\theta(X_{t_i}^\pi,\cdot), h, \Delta W_i)\right|^2. \tag{4.10}$$

Algorithm 1: DBDP1 PIDE extension

Start with $\widehat{u}_N(\cdot) = g(\cdot)$;
for $i \in \{N-1, ..., 1\}$ **do**
  Given $\widehat{u}_{i+1}$;
  Compute $\theta^* = \underset{\theta}{\operatorname{argmin}}\ L_i(\theta)$;
  Update $(\widehat{u}_i, \widehat{z}_i, \widehat{w}_i) = (u_i^{\theta^*}, z_i^{\theta^*}, w_i^{\theta^*})$;
end

For the minimization step we need to calculate an expected value, but this is a complicated task due to the non linearity and the fact that the distribution of the random variables involved are not always known. To overcome this situation, as well as in [HPW19], one has to use a Monte Carlo approximation together with Stochastic Gradient Descent (SGD). See also Remark 4.1.

4.2 Previous Definitions and Results

First, introduce the conditional expectations of the averaged processes

$$\overline{Z}_{t_i} = \frac{1}{h}\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} Z_t dt\right), \quad \overline{\Gamma}_{t_i} = \frac{1}{h}\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \Gamma_t dt\right). \tag{4.11}$$

these quantities allows us to define the L^2-regularity of the solutions (Z,Γ) (see [BE08] and [HPW19]) as follows

$$\begin{aligned} e(Z,(\overline{Z}_t)_{t\in\pi}) &:= \mathbb{E}\left(\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} \|Z_t - \overline{Z}_{t_i}\|^2 dt\right), \\ e(\Gamma,(\overline{\Gamma}_t)_{t\in\pi}) &:= \mathbb{E}\left(\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} |\Gamma_t - \overline{\Gamma}_{t_i}|^2 dt\right). \end{aligned} \tag{4.12}$$

Both quantities can be made arbitrarily small as it is shown on [BE08] and presented in the following theorem.

Theorem 4.3 *Under Assumptions 2.10, there exists a constant $C > 0$ such that*

$$e(\Gamma, (\overline{\Gamma}_t)_{t\in\pi}) \leq Ch \qquad \textit{and} \qquad e(Z, (\overline{Z}_t)_{t\in\pi}) \leq Ch.$$

PROOF. See [BE08, Theorem 2.1 (i)] for the bound on $e(\Gamma, (\overline{\Gamma}_t)_{t\in\pi})$ and [BE08, Theorem 2.1 (ii)] for the bound on $e(Z, (\overline{Z}_t)_{t\in\pi})$. Note that in the cited reference this result is presented, using our notation, as follows,

$$\left\|\Gamma - \overline{\Gamma}\right\|^2_{\mathcal{N}^2_W(0,T;\mathbb{R})} \leq CN^{-1} \qquad \textit{and} \qquad \left\|Z - \overline{Z}\right\|^2_{\mathcal{N}^2_W(0,T;\mathbb{R}^d)} \leq CN^{-1}.$$

Where $\overline{\Gamma}_t = \overline{\Gamma}_{t_i}$ for $t \in [t_i, t_{i+1})$ and $\overline{Z}_t = \overline{Z}_{t_i}$ for $t \in [t_i, t_{i+1})$. □

We introduce a somehow auxiliary scheme that at the same time depends on the main one. Let $i \in \{0, \dots, N-1\}$. We follow the procedure taken in [HPW19], with key modifications. Let us use the ideas of [BE08] to define $\mathcal{F}$-adapted discrete processes

$$\widehat{\mathcal{V}}_{t_i} = \mathbb{E}_i\left(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right) + f\left(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}, \overline{\widehat{\Gamma}}_{t_i}\right)h, \tag{4.13}$$

$$\overline{\widehat{Z}}_{t_i} = \frac{1}{h}\mathbb{E}_i\left(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\Delta W_i\right), \tag{4.14}$$

$$\overline{\widehat{\Gamma}}_{t_i} = \frac{1}{h}\mathbb{E}_i\left(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\Delta M_i\right), \tag{4.15}$$

where $\widehat{\mathcal{V}}_{t_i}$ is well-defined for sufficiently small h by Lemma 4.4 and the variables $\overline{\widehat{Z}}_{t_i}$, $\overline{\widehat{\Gamma}}_{t_i}$ are defined below.

Lemma 4.4 *The process $\widehat{\mathcal{V}}_{t_i}$ is well-defined.*

PROOF. Let $i \in \{0, ..., N-1\}$ and $\psi : L^2(\Omega, \mathcal{F}, \mathbb{P}) \to L^2(\Omega, \mathcal{F}, \mathbb{P})$ be defined as

$$\psi(\xi)(\omega) = \mathbb{E}_i\left(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right)(\omega) + f\left(t_i, X^\pi_{t_i}(\omega), \xi(\omega), \overline{\widehat{Z}}_{t_i}(\omega), \overline{\widehat{\Gamma}}_{t_i}(\omega)\right)h.$$

For all $\xi \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and $\omega \in \Omega$. This function is well-defined by the properties of f and Lemma 3.8. Let $\xi, \overline{\xi} \in L^2$, then $\mathbb{P}$ a.s $|\psi(\xi) - \psi(\overline{\xi})| \leq h|\xi - \overline{\xi}|$, therefore

$$\left\|\psi(\xi) - \psi(\overline{\xi})\right\|_{L^2(\Omega,\mathcal{F},\mathbb{P})} \leq h\left\|\xi - \overline{\xi}\right\|_{L^2(\Omega,\mathcal{F},\mathbb{P})}$$

Taking sufficiently small h we can see that this function is a contraction on $L^2(\Omega, \mathcal{F}, \mathbb{P})$, and therefore, by applying Banach's fixed point theorem, we conclude the proof. □

For fixed $i \in \{0, ..., N\}$, let N_t be a process defined as $N_t := \mathbb{E}\left(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\middle|\mathcal{F}_t\right)$ for $t \in [t_i, t_{i+1}]$. Using Lemma 3.8, it is not difficult to see that N_t is a square integrable martingale

and therefore, by Martingale Representation Theorem (see Lemma 2.29), there exist $(\widehat{Z}, \widehat{U}) \in \mathscr{N}_W^2(0,T;\mathbb{R}^d) \times \mathscr{N}_\mu^2(0,T;\mathbb{R})$ such that

$$N_t = N_{t_i} + \int_{t_i}^t \widehat{Z}_s \cdot dW_s + \int_{t_i}^t \int_{\mathbb{R}^d} \widehat{U}_s(y)\overline{\mu}(ds,dy).$$

By taking $t = t_{i+1}$ and recalling that $\mathbb{E}_i(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_{t_i})$,

$$\widehat{u}_{i+1}(X^\pi_{t_{i+1}}) = \mathbb{E}_i\left(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right) + \int_{t_i}^{t_{i+1}} \widehat{Z}_s \cdot dW_s + \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \widehat{U}_s(y)\overline{\mu}(ds,dy).$$

By multiplying by ΔW_i and ΔM_i, then taking $\mathbb{E}_i$ and using Itô isometry,

$$\begin{aligned}
\overline{\widehat{Z}}_{t_i} &= \frac{1}{h}\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \widehat{Z}_s ds\right), \\
\overline{\widehat{\Gamma}}_{t_i} &= \frac{1}{h}\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \widehat{U}_s(y)\lambda(dy)ds\right).
\end{aligned}$$

Let

$$\overline{\widehat{U}}_{t_i}(y) := \frac{1}{h}\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \widehat{U}_s(y) ds\right). \tag{4.16}$$

By Lemma 2.31 one can see that

$$\overline{\widehat{\Gamma}}_{t_i} = \frac{1}{h}\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \widehat{U}_s(y)\lambda(dy)ds\right) = \int_{\mathbb{R}^d} \overline{\widehat{U}}_{t_i}(y)\lambda(dy). \tag{4.17}$$

The last equality can be seen as an analogous to (1.8) and makes sense with the notation $\overline{\widehat{\Gamma}}_{t_i} = \langle \overline{\widehat{U}}_{t_i} \rangle$. Also, we can establish the following useful bound:

$$\mathbb{E}\left|\overline{\widehat{\Gamma}}_{t_i} - \langle w_i^\theta \rangle(X^\pi_{t_i})\right|^2 \lesssim \mathbb{E}\left(\left\|\overline{\widehat{U}}_{t_i}(\cdot) - w_i^\theta(X^\pi_{t_i},\cdot)\right\|^2_{L^2(\mathbb{R}^d,\lambda)}\right).$$

Indeed, from (4.17) and (4.9), Hölder inequality and the fact that λ is a finite measure

$$\begin{aligned}
\mathbb{E}\left|\overline{\widehat{\Gamma}}_{t_i} - \langle w_i^\theta \rangle(X^\pi_{t_i})\right|^2 &= \mathbb{E}\left|\int_{\mathbb{R}^d} \overline{\widehat{U}}_{t_i}(y)\lambda(dy) - \int_{\mathbb{R}^d} w_i^\theta(X^\pi_{t_i},y)\lambda(dy)\right|^2 \\
&\lesssim \mathbb{E}\left(\left\|\overline{\widehat{U}}_{t_i}(\cdot) - w_i^\theta(X^\pi_{t_i},\cdot)\right\|^2_{L^2(\mathbb{R}^d,\lambda)}\right).
\end{aligned}$$

Following [HPW19], we can find deterministic functions v_i, ζ_i, γ_i such that $v_i(X^\pi_{t_i}) = \widehat{\mathcal{V}}_{t_i}, \zeta_i(X^\pi_{t_i}) = \overline{\widehat{Z}}_{t_i}$ and $\gamma_i(y, X^\pi_{t_i}) = \overline{\widehat{U}}_{t_i}(y)$ for $y \in \mathbb{R}^d$. The correspondent L^2-integrability of these functions is ensured by the properties of $\widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}$ and $\overline{\widehat{U}}_{t_i}$. With the previous setup, the natural extension of the terms to estimate the error of the scheme shown on [HPW19] must be

$$\begin{aligned}
\varepsilon_i^{v,\kappa} &= \inf_{\theta\in\mathbb{R}^\kappa} \mathbb{E}\left|v_i(X^\pi_{t_i}) - u_i^\theta(X^\pi_{t_i})\right|^2, \quad \varepsilon_i^{\zeta,\kappa} = \inf_{\theta\in\mathbb{R}^\kappa} \mathbb{E}\left|\zeta_i(X^\pi_{t_i}) - z_i^\theta(X^\pi_{t_i})\right|^2 \\
\varepsilon_i^{\gamma,\kappa} &= \inf_{\theta\in\mathbb{R}^\kappa} \mathbb{E}\left(\int_{\mathbb{R}^d} \left|\gamma_i(y, X^\pi_{t_i}) - w_i^\theta(X^\pi_{t_i},y)\right|^2 \lambda(dy)\right).
\end{aligned} \tag{4.18}$$

The expected values can be written as a integral with respect a probability measure in $\mathbb{R}^d$ and therefore, applying the Theorem 3.9, these quantities can be made arbitrarily small as κ increases.

The following results will be useful in the proof of the main result. In Section 2.5 of [BE08], it is explained that the results presented there still hold for a time-dependent non-linearity.

Proposition 4.5 *There exists a constant $C > 0$ independent of the step h such that*

$$\sum_{i=0}^{N-1} \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds\right) \leq Ch.$$

PROOF. See [BE08, Proposition 2.1]. □

We will also need the following result.

Lemma 4.6 *Consider $(X, Y, Z, U) \in \mathscr{S}_T^2(\mathbb{R}^d) \times \mathscr{B}^2$ the solution to (1.6) - (1.7), Γ defined as in (1.8) and $\Theta_s = (s, X_s, Y_s, Z_s, \Gamma_s)$. Then,*

$$\mathbb{E}\left(\int_0^T |f(\Theta_s)|^2 ds\right) < \infty.$$

PROOF. First, note that by using useful bound (4.19) we have that for every $s \in [0, T]$

$$|f(\Theta_s)|^2 \leq 2(|f(\Theta_s) - f(s, 0, 0, 0, 0)|^2 + |f(s, 0, 0, 0, 0)|^2).$$

Applying again (4.19) and the Lipschitz bound on f,

$$|f(\Theta_s)|^2 \leq 2\Big[K^2 5(\|X_s\|^2 + |Y_s|^2 + \|Z_s\|^2 + |\Gamma_s|^2) + |f(s, 0, 0, 0, 0)|^2\Big].$$

Then, integrating on $\Omega \times [0, T]$ with respect to $d\mathbb{P} \times ds$, using Hölder inequality and bound (2.10),

$$\begin{aligned}
\mathbb{E}\left(\int_0^T |f(\Theta_s)|^2 ds\right) &\leq 10K^2 T \mathbb{E}\left(\sup_{s\in[0,T]} \|X_s\|^2 + \sup_{s\in[0,T]} |Y_s|^2\right) \\
&\quad + 10K^2 \left(\mathbb{E}\int_0^T \|Z_s\|^2 ds + \lambda(\mathbb{R}^d)\mathbb{E}\int_0^T \int_{\mathbb{R}^d} |U_s(y)|^2 \lambda(dy) ds\right) \\
&\quad + 2T \sup_{s\in[0,T]} |f(s, 0, 0, 0, 0)|^2 \\
&\quad < \infty
\end{aligned}$$

this finishes the proof. □

4.3 Main Result

As stated previously, the proof of our main result, Theorem 4.7, is deeply inspired in the case without jumps considered in [HPW19]. We follow the lines of that proof with some important

differences because of the nonlocal character of our problem. Also, along the proof we use several times that for $x_1, ..., x_k \in \mathbb{R}$, the following holds

$$(x_1 + \cdots + x_k)^2 \leq k(x_1^2 + \cdots + x_k^2). \tag{4.19}$$

Theorem 4.7 *Under Assumptions 2.10, there exists a constant $C > 0$ independent of the partition such that for sufficiently small h,*

$$\begin{aligned}\max_{i=0,\dots,N-1} \mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 + \sum_{i=0}^{N-1} \mathbb{E}\left(\int_{t_i}^{t_{i+1}} \left[\|Z_t - \widehat{z}_i(X^\pi_{t_i})\|^2 + |\Gamma_t - \langle \widehat{w}_i\rangle(X^\pi_{t_i})|^2\right] dt\right) \\ \leq C\left[h + \sum_{i=0}^{N-1}(N\varepsilon_i^v + \varepsilon_i^\zeta + \varepsilon_i^\gamma) + e(Z, (\overline{Z}_t)_{t\in\pi}) + e(\Gamma, (\overline{\Gamma}_t)_{t\in\pi}) + \mathbb{E}\left|g(X_T) - g(X^\pi_T)\right|^2\right],\end{aligned}$$

with ε_i^v, ε_i^ζ and ε_i^γ given in (4.18), and $e(Z, (\overline{Z}_t)_{t\in\pi})$ and $e(\Gamma, (\overline{\Gamma}_t)_{t\in\pi})$ defined in (4.12).

PROOF. **Step 1**

Recall $\widehat{\mathcal{V}}_{t_i}$ introduced in (4.13). The purpose of this part is to obtain a suitable bound of the term $\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2$ in terms of more tractable terms. We have

Lemma 4.8 *There exists $C > 0$ fixed such that for any $0 < h < 1$ sufficiently small, one has*

$$\begin{aligned}\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 \leq Ch^2 + C\mathbb{E}\left(\int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds\right) + C\mathbb{E}\left(\int_{t_i}^{t_{i+1}} \|Z_s - \overline{Z}_{t_i}\|^2 ds\right) \\ + C\mathbb{E}\left(\int_{t_i}^{t_{i+1}} |\Gamma_s - \overline{\Gamma}_{t_i}|^2 ds\right) + Ch\mathbb{E}\left(\int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr\right) \\ + C(1 + Ch)\mathbb{E}\left|Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right|^2,\end{aligned} \tag{4.20}$$

with $\Theta_r = (r, X_r, Y_r, Z_r, \Gamma_r)$.

The rest of this subsection is devoted to the proof of this result.

PROOF. Subtracting the equation (1.7) between t_i and t_{i+1}, we obtain

$$\Delta Y_i = Y_{t_{i+1}} - Y_{t_i} = -\int_{t_i}^{t_{i+1}} f(\Theta_s) ds + \int_{t_i}^{t_{i+1}} Z_s \cdot dW_s + \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} U_s(y)\overline{\mu}(ds, dy). \tag{4.21}$$

Using the definition of $\widehat{\mathcal{V}}_{t_i}$ in (4.13),

$$\begin{aligned}Y_{t_i} - \widehat{\mathcal{V}}_{t_i} &= Y_{t_{i+1}} - \Delta Y_i - \widehat{\mathcal{V}}_{t_i} \\ &= Y_{t_{i+1}} + \int_{t_i}^{t_{i+1}} [f(\Theta_s) - f(\widehat{\Theta}_{t_i})] ds - \int_{t_i}^{t_{i+1}} Z_s \cdot dW_s - \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} U_s(y)\overline{\mu}(ds, dy) \\ &\quad - \mathbb{E}_i(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})).\end{aligned}$$

Here $\widehat{\Theta}_{t_i} = (t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}, \overline{\widehat{\Gamma}}_{t_i})$. Then, by applying the conditional expectation for time t_i given by $\mathbb{E}_i$ and using that, in this case, the stochastic integrals are martingales

$$Y_{t_i} - \widehat{\mathcal{V}}_{t_i} = \mathbb{E}_i(Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})) + \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} [f(\Theta_s) - f(\widehat{\Theta}_{t_i})]ds\right) = a + b.$$

Using the classical inequality $(a+b)^2 \leq (1+\gamma h)a^2 + (1+\frac{1}{\gamma h})b^2$ for $\gamma > 0$ to be chosen, we get

$$\begin{aligned}\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 \leq (1+\gamma h)\mathbb{E}\left[\mathbb{E}_i\left(Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right)\right]^2 \\ + \left(1 + \frac{1}{\gamma h}\right)\mathbb{E}\left[\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} [f(\Theta_s) - f(\widehat{\Theta}_{t_i})]ds\right)\right]^2.\end{aligned} \tag{4.22}$$

With no lose of generality, because we are looking for bounds, we can replace $[f(\Theta_s) - f(\widehat{\Theta}_{t_i})]$ by $|f(\Theta_s) - f(\widehat{\Theta}_{t_i})|$. Also, we can drop the $\mathbb{E}_i$ due to the law of total expectation. The Lipschitz condition on f in (2.10) allows us to give a bound in terms of the difference between Θ_s and $\widehat{\Theta}_{t_i}$. Indeed, for a fixed constant $K > 0$,

$$|f(\Theta_s) - f(\widehat{\Theta}_{t_i})| \leq K\left(|s - t_i|^{1/2} + \|X_s - X^\pi_{t_i}\| + |Y_s - \widehat{\mathcal{V}}_{t_i}| + \|Z_s - \overline{\widehat{Z}}_{t_i}\| + |\Gamma_s - \overline{\widehat{\Gamma}}_{t_i}|\right).$$

Therefore, we have the bound

$$\begin{aligned}\mathbb{E}\left(\int_{t_i}^{t_{i+1}} |f(\Theta_s) - f(\widehat{\Theta}_{t_i})|ds\right)^2 \leq Ch\left[h^2 + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} \|X_s - X^\pi_{t_i}\|^2 ds\right) + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |Y_s - \widehat{\mathcal{V}}_{t_i}|^2 ds\right)\right. \\ \left. + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} \|Z_s - \overline{\widehat{Z}}_{t_i}\|^2 ds\right) + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |\Gamma_s - \overline{\widehat{\Gamma}}_{t_i}|^2 ds\right)\right],\end{aligned}$$

where the Lipschitz constant K was absorbed by C. Using now the triangle inequality $|Y_s - \widehat{\mathcal{V}}_{t_i}|^2 \leq 2|Y_s - Y_{t_i}|^2 + 2|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}|^2$, and the approximation error of the X scheme (4.5), we find

$$\mathbb{E}\left(\int_{t_i}^{t_{i+1}} |f(\Theta_s) - f(\widehat{\Theta}_{t_i})|ds\right)^2 \tag{4.23}$$

$$\begin{aligned}\leq Ch\left[h^2 + 2\mathbb{E}\left(\int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds\right) + 2h\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2\right. \\ \left. + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} \|Z_s - \overline{\widehat{Z}}_{t_i}\|^2 ds\right) + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |\Gamma_s - \overline{\widehat{\Gamma}}_{t_i}|^2 ds\right)\right],\end{aligned} \tag{4.24}$$

and therefore, replacing in (4.22),

$$\begin{aligned}&\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 \\ &\leq (1+\gamma h)\,\mathbb{E}\left|\mathbb{E}_i\left[Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right]\right|^2 \\ &\quad + (1+\gamma h)\,\frac{C}{\gamma}\left[h^2 + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds\right) + h\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2\right. \\ &\qquad \left. + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} \|Z_s - \overline{\widehat{Z}}_{t_i}\|^2 ds\right) + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |\Gamma_s - \overline{\widehat{\Gamma}}_{t_i}|^2 ds\right)\right].\end{aligned} \tag{4.25}$$

Recall $\overline{Z}_{t_i}$ and $\overline{\Gamma}_{t_i}$ introduced in (4.11). Now, we are going to prove the following

$$\mathbb{E}\left(\int_{t_i}^{t_{i+1}} \|Z_s - \overline{\widehat{Z}}_{t_i}\|^2 ds\right) = \mathbb{E}\left(\int_{t_i}^{t_{i+1}} \|Z_s - \overline{Z}_{t_i}\|^2 ds\right) + h\mathbb{E}\left|\overline{Z}_{t_i} - \overline{\widehat{Z}}_{t_i}\right|^2. \tag{4.26}$$

$$\mathbb{E}\left(\int_{t_i}^{t_{i+1}} |\Gamma_s - \overline{\widehat{\Gamma}}_{t_i}|^2 ds\right) = \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |\Gamma_s - \overline{\Gamma}_{t_i}|^2 ds\right) + h\mathbb{E}\left|\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i}\right|^2. \tag{4.27}$$

Let us prove the latter, the former is analogous. Recall that the Γ components represents the nonlocal part and therefore is one dimensional.

$$|\Gamma_t - \overline{\widehat{\Gamma}}_{t_i}|^2 = |(\Gamma_t - \overline{\Gamma}_{t_i}) + (\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i})|^2 = (\Gamma_t - \overline{\Gamma}_{t_i})^2 + (\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i})^2 + 2(\Gamma_t - \overline{\Gamma}_{t_i})(\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i}).$$

It is sufficient to establish that the double product is 0 when integrating and taking expectation. Recall that $\overline{\Gamma}_{t_i}$ from (4.11) is a $\mathcal{F}_{t_i}$ measurable random variable. Then,

$$\begin{aligned}\int_{t_i}^{t_{i+1}} \left(\Gamma_t - \overline{\Gamma}_{t_i}\right)(\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i})dt &= \left(\int_{t_i}^{t_{i+1}} (\Gamma_t - \overline{\Gamma}_{t_i})dt\right)(\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i}) \\ &= \left[\int_{t_i}^{t_{i+1}} \Gamma_t dt - \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \Gamma_t dt\right)\right](\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i}).\end{aligned}$$

Due to the $\mathcal{F}_{t_i}$-measurability of the right side of the last multiplication and the $L^2(\mathbb{P})$ orthogonality, taking expectation annihilates the last term. Therefore, equations (4.26) and (4.27) are proven. By multiplying (4.21) by ΔW_i and taking $\mathbb{E}_i$,

$$\begin{aligned}\mathbb{E}_i\left(\Delta W_i Y_{t_{i+1}}\right) + \mathbb{E}_i\left(\Delta W_i \int_{t_i}^{t_{i+1}} f(\Theta_r)dr\right) &= \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} dW_r \int_{t_i}^{t_{i+1}} Z_r \cdot dW_r\right) \\ &\quad + \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} U_r(y)\overline{\mu}(dy, dr) \int_{t_i}^{t_{i+1}} dW_r\right) \\ &= \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} Z_r dr\right) = h\overline{Z}_{t_i},\end{aligned}$$

where we have used Lemma 2.11. Then, subtracting $h\overline{\widehat{Z}}_{t_i} = \mathbb{E}_i(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\Delta W_i)$,

$$h(\overline{Z}_{t_i} - \overline{\widehat{Z}}_{t_i}) = \mathbb{E}_i\left[\Delta W_i(Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}}))\right] + \mathbb{E}_i\left(\Delta W_i \int_{t_i}^{t_{i+1}} h(\Theta_r)dr\right).$$

By multiplying (4.21) by ΔM_i and taking $\mathbb{E}_i$,

$$\begin{aligned}&\mathbb{E}_i\left(\Delta M_i Y_{t_{i+1}}\right) + \mathbb{E}_i\left(\Delta M_i \int_{t_i}^{t_{i+1}} f(\Theta_r)dr\right) \\ &= \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \overline{\mu}(ds, dy) \int_{t_i}^{t_{i+1}} Z_r \cdot dW_r\right) + \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \overline{\mu}(dr, dy) \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} U_r(y)\overline{\mu}(dr, dy)\right) \\ &= \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} U_r(y)\lambda(dy)ds\right) = h\overline{\Gamma}_{t_i}.\end{aligned}$$

Then, subtracting $h\overline{\widehat{\Gamma}}_{t_i} = \mathbb{E}_i\left(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\Delta M_i\right)$,

$$h(\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i}) = \mathbb{E}_i\left[\Delta M_i\left(Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right)\right] + \mathbb{E}_i\left(\Delta M_i \int_{t_i}^{t_{i+1}} f(\Theta_r)dr\right).$$

Summarizing, one has

$$h(\overline{Z}_{t_i} - \overline{\widehat{Z}}_{t_i}) = \mathbb{E}_i\left[\Delta W_i\left(Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}}) - \mathbb{E}_i\left[Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right]\right)\right] + \mathbb{E}_i\left[\Delta W_i \int_{t_i}^{t_{i+1}} f(\Theta_r)dr\right];$$

$$h(\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i}) = \mathbb{E}_i\left[\Delta M_i\left(Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}}) - \mathbb{E}_i\left[Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right]\right)\right] + \mathbb{E}_i\left[\Delta M_i \int_{t_i}^{t_{i+1}} f(\Theta_r)dr\right].$$

For the sake of brevity, define now

$$H_i := Y_{t_i} - \widehat{u}_i(X^\pi_{t_i}); \tag{4.28}$$

note that it depends on i. By the properties related with Itô isometry, from the previous identities we have

$$\mathbb{E}\left(h^2\|\overline{Z}_{t_i} - \overline{\widehat{Z}}_{t_i}\|^2\right) \leq 2dh\left(\mathbb{E}(H^2_{i+1}) - \mathbb{E}\,|\mathbb{E}_i(H_{i+1})|^2\right) + 2dh^2\mathbb{E}\left[\int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr\right]; \tag{4.29}$$

$$\mathbb{E}\left(h^2|\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i}|^2\right) \leq 2\lambda(\mathbb{R}^d)h\left(\mathbb{E}(H^2_{i+1}) - \mathbb{E}\,|\mathbb{E}_i(H_{i+1})|^2\right) + 2\lambda(\mathbb{R}^d)h^2\mathbb{E}\left[\int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr\right]. \tag{4.30}$$

Remark 4.9 *Note that in the previous bound is important the finiteness of the Levy measure λ. The case of more general integro-differential operators, such as the fractional Laplacian mentioned in the introduction, it is an interesting open problem.*

Let us work with equation (4.24). Using (4.26) and (4.27),

$$\begin{aligned}\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 \leq\ & (1+\gamma h)\,\mathbb{E}\,|\mathbb{E}_i(H_{i+1})|^2 \\ & + (1+\gamma h)\,\frac{C}{\gamma}\left[h^2 + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds\right) + h\mathbb{E}|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}|^2\right. \\ & \qquad + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} \|Z_s - \overline{Z}_{t_i}\|^2 ds\right) + h\mathbb{E}\|\overline{Z}_{t_i} - \overline{\widehat{Z}}_{t_i}\|^2 \\ & \qquad \left. + \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |\Gamma_s - \overline{\Gamma}_{t_i}|^2 ds\right) + h\mathbb{E}\left|\overline{\Gamma}_{t_i} - \overline{\widehat{\Gamma}}_{t_i}\right|^2\right].\end{aligned}$$

Now use (4.29) and (4.30) to find that

$$
\begin{aligned}
&\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2 \\
&\leq (1+\gamma h)\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \\
&\quad + (1+\gamma h)\frac{C}{\gamma}\left[h^2+\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|Y_s-Y_{t_i}|^2ds\right)+h\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2\right. \\
&\qquad\qquad + \mathbb{E}\left(\int_{t_i}^{t_{i+1}}\|Z_s-\overline{Z}_{t_i}\|^2ds\right)+\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|\Gamma_s-\overline{\Gamma}_{t_i}|^2ds\right) \\
&\qquad\qquad + 2d\left[\mathbb{E}\left(H_{i+1}^2\right)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right]+2dh\mathbb{E}\left(\int_{t_i}^{t_{i+1}}f(\Theta_r)^2dr\right) \\
&\qquad\qquad \left. + 2\lambda(\mathbb{R}^d)\left[\mathbb{E}\left(H_{i+1}^2\right)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right]+2\lambda(\mathbb{R}^d)h\mathbb{E}\left(\int_{t_i}^{t_{i+1}}f(\Theta_r)^2dr\right)\right].
\end{aligned}
$$

Let $\gamma = C(\lambda(\mathbb{R}^d)+d)$ and define $D := (1+\gamma h)\frac{C}{\gamma}$, then the above term is bounded by

$$
\begin{aligned}
&(1+\gamma h)\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2+Dh^2+D\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|Y_s-Y_{t_i}|^2\right)+Dh\mathbb{E}|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}|^2+D\mathbb{E}\left(\int_{t_i}^{t_{i+1}}\|Z-\overline{Z}_{t_i}\|^2ds\right) \\
&\quad + D\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|\Gamma_s-\overline{\Gamma}_{t_i}|^2ds\right)+(1+\gamma h)\frac{C}{\gamma}2d\mathbb{E}\left(H_{i+1}^2\right)+2dDh\mathbb{E}\left(\int_{t_i}^{t_{i+1}}f(\Theta_r)^2dr\right) \\
&\quad + (1+\gamma h)\frac{C}{\gamma}2\lambda(\mathbb{R}^d)\mathbb{E}\left(H_{i+1}^2\right)+2\lambda(\mathbb{R}^d)Dh\mathbb{E}\left(\int_{t_i}^{t_{i+1}}f(\Theta_r)^2dr\right)-2(1+\gamma h)\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2
\end{aligned}
$$

Note that the first and last term in the last expression are similar, therefore can be subtracted which yields a negative number that can be bounded from above by 0. Also, we have the similar terms on $\mathbb{E}\left(H_{i+1}^2\right)$ and the integral of f that we put together and bound respectively. Due to the definition of D, from now on the constant C has a linear dependence on the dimension d such that $D \leq C$. By replacing the last calculation and putting $\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2$ on the left side

$$
\begin{aligned}
&(1-Ch)\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2 \\
&\quad \leq Ch^2+C\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|Y_s-Y_{t_i}|^2ds\right)+C\mathbb{E}\left(\int_{t_i}^{t_{i+1}}\|Z_s-\overline{Z}_{t_i}\|^2ds\right) \\
&\quad + C\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|\Gamma_s-\overline{\Gamma}_{t_i}|^2ds\right)+C(1+Ch)\mathbb{E}\left(H_{i+1}^2\right)+Ch\mathbb{E}\left(\int_{t_i}^{t_{i+1}}f(\Theta_r)^2dr\right).
\end{aligned}
$$

Now we have to take h small such that, for example, $Ch \leq \frac{1}{2}$ and then

$$
\begin{aligned}
\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2 &\leq Ch^2+C\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|Y_s-Y_{t_i}|^2ds\right)+C\mathbb{E}\left(\int_{t_i}^{t_{i+1}}\|Z_s-\overline{Z}_{t_i}\|^2ds\right) \\
&\quad + C\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|\Gamma_s-\overline{\Gamma}_{t_i}|^2ds\right)+Ch\mathbb{E}\left(\int_{t_i}^{t_{i+1}}f(\Theta_r)^2dr\right)+C(1+Ch)\mathbb{E}\left(H_{i+1}^2\right).
\end{aligned}
$$

Finally, by recalling that $H_{i+1} = Y_{t_{i+1}} - \widehat{u}_{i+1}(X_{t_{i+1}}^{\pi})$, we have established (4.20). □

Step 2

The last term in (4.20),

$$C(1+Ch)\mathbb{E}\left|Y_{t_{i+1}}-\widehat{u}_{i+1}(X^{\pi}_{t_{i+1}})\right|^2,$$

was left without a control in previous step. Here in what follows we provide a control on this term. Recall the error terms $e(Z,(Z_t)_{t\in\pi})$ and $e(\Gamma,(\Gamma_t)_{t\in\pi})$ introduced in (4.12). The purpose of this section is to show the following estimate:

Lemma 4.10 *There exists a constant $C>0$ such that,*

$$\max_{i\in\{0,\dots,N-1\}}\mathbb{E}\left|Y_{t_i}-\widehat{u}_i(X^{\pi}_{t_i})\right|^2 \leq C\Bigg[N\sum_{i=0}^{N-1}\mathbb{E}\left|\widehat{u}_i(X^{\pi}_{t_i})-\widehat{\mathcal{V}}_{t_i}\right|^2+h+e(Z,(\overline{Z}_t)_{t\in\pi})+e(\Gamma,(\overline{\Gamma}_t)_{t\in\pi}) \\ +\mathbb{E}\left|g(X_T)-g(X^{\pi}_T)\right|^2\Bigg]. \tag{4.31}$$

The rest of this section is devoted to the proof of this result.

PROOF. (of Lemma 4.10) We have that $(a+b)^2\geq(1-h)a^2+(1-\frac{1}{h})b^2$ and

$$\begin{aligned}\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2&=\mathbb{E}\left|\left(Y_{t_i}-\widehat{u}_i(X^{\pi}_{t_i})\right)+\left(\widehat{u}_i(X^{\pi}_{t_i})-\widehat{\mathcal{V}}_{t_i}\right)\right|^2 \\ &\geq(1-h)\mathbb{E}\left|Y_{t_i}-\widehat{u}_i(X^{\pi}_{t_i})\right|^2+\left(1-\frac{1}{h}\right)\mathbb{E}\left|\widehat{u}_i(X^{\pi}_{t_i})-\widehat{\mathcal{V}}_{t_i}\right|^2.\end{aligned} \tag{4.32}$$

Therefore, we have an upper (4.20) and lower bound for $\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2$. By connecting these bounds,

$$\begin{aligned}&(1-h)\mathbb{E}\left|Y_{t_i}-\widehat{u}_i(X^{\pi}_{t_i})\right|^2+\left(1-\frac{1}{h}\right)\mathbb{E}\left|\widehat{u}_i(X^{\pi}_{t_i})-\widehat{\mathcal{V}}_{t_i}\right|^2 \\ &\quad\leq Ch^2+C\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|Y_s-Y_{t_i}|^2ds\right)+C\mathbb{E}\left(\int_{t_i}^{t_{i+1}}\|Z_s-\overline{Z}_{t_i}\|^2ds\right) \\ &\quad\quad+C\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|\Gamma_s-\overline{\Gamma}_{t_i}|^2ds\right)+Ch\mathbb{E}\left(\int_{t_i}^{t_{i+1}}f(\Theta_r)^2dr\right)+C(1+Ch)\mathbb{E}\left(H^2_{i+1}\right).\end{aligned}$$

Using that for sufficiently small h we have $(1-h)^{-1}\leq 2$, we get,

$$\begin{aligned}&\mathbb{E}\left|Y_{t_i}-\widehat{u}_i(X^{\pi}_{t_i})\right|^2 \\ &\leq CN\mathbb{E}\left|\widehat{u}_i(X^{\pi}_{t_i})-\widehat{\mathcal{V}}_{t_i}\right|^2+Ch^2 \\ &\quad+C\left[\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|Y_s-Y_{t_i}|^2ds\right)+\mathbb{E}\left(\int_{t_i}^{t_{i+1}}\|Z_s-\overline{Z}_{t_i}\|^2ds\right)+\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|\Gamma_s-\overline{\Gamma}_{t_i}|^2ds\right)\right] \\ &\quad+Ch\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|f(\Theta_s)|^2ds\right)+C\mathbb{E}\left|Y_{t_{i+1}}-\widehat{u}_{i+1}(X^{\pi}_{t_{i+1}})\right|^2.\end{aligned}$$

Notice that the expression on time t_i that we want to estimate, appears on the right side on time t_{i+1}, we can iterate the bound and get that $\forall\, i \in \{0, ..., N-1\}$

$$
\begin{aligned}
&\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 \\
&\leq NC \sum_{k=i}^{N-1} \mathbb{E}\left|\widehat{u}_k(X^\pi_{t_k}) - \widehat{\mathcal{V}}_{t_k}\right|^2 + C(N-i)h^2 \\
&\quad + C\sum_{k=i}^{N-1}\left[\mathbb{E}\left(\int_{t_k}^{t_{k+1}} |Y_s - Y_{t_k}|^2 ds\right) + \mathbb{E}\left(\int_{t_k}^{t_{k+1}} \|Z_s - \overline{Z}_{t_k}\|^2 ds\right) + \mathbb{E}\left(\int_{t_k}^{t_{k+1}} |\Gamma_s - \overline{\Gamma}_{t_k}|^2 ds\right)\right] \\
&\quad + Ch\sum_{k=i}^{N-1}\mathbb{E}\left(\int_{t_k}^{t_{k+1}} |f(\Theta_s)|^2 ds\right) + C\mathbb{E}\left|Y_{t_N} - g(X^\pi_{t_N})\right|^2 \\
&\leq NC \sum_{k=0}^{N-1} \mathbb{E}\left|\widehat{u}_k(X^\pi_{t_k}) - \widehat{\mathcal{V}}_{t_k}\right|^2 + CNh^2 \\
&\quad + C\sum_{k=0}^{N-1}\left[\mathbb{E}\left(\int_{t_k}^{t_{k+1}} |Y_s - Y_{t_k}|^2 ds\right) + \mathbb{E}\left(\int_{t_k}^{t_{k+1}} \|Z_s - \overline{Z}_{t_k}\|^2 ds\right) + \mathbb{E}\left(\int_{t_k}^{t_{k+1}} |\Gamma_s - \overline{\Gamma}_{t_k}|^2 ds\right)\right] \\
&\quad + Ch\sum_{k=0}^{N-1}\mathbb{E}\left(\int_{t_k}^{t_{k+1}} |f(\Theta_s)|^2 ds\right) + C\mathbb{E}\left|Y_{t_N} - g(X^\pi_{t_N})\right|^2.
\end{aligned}
$$

Applying maximum on $i \in \{0, ..., N-1\}$, recalling (4.12) and the bounds from Lemmas (4.6) and (4.5),

$$
\begin{aligned}
&\max_{i\in\{0,\dots,N-1\}} \mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 \\
&\leq C\left[N\sum_{i=0}^{N-1}\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2 + h + e(Z, (\overline{Z}_t)_{t\in\pi}) + e(\Gamma, (\overline{\Gamma}_t)_{t\in\pi}) + \mathbb{E}\left|g(X_T) - g(X^\pi_T)\right|^2\right].
\end{aligned}
$$

This is nothing that (4.31).

□

Remark 4.11 *The classic bound used at the beginning of step 2 could have been stated using a fixed parameter $\delta \in (0,1)$ in the form: $(a+b)^2 \geq (1-h^\delta)a^2 + (1-\frac{1}{h^\delta})b^2$. This change makes N become N^δ, which is better. However, at some point of the proof the value $\delta = 1$ is necessary.*

Step 3

Estimate (4.31) contains some uncontrolled terms on its RHS. Here the purpose is to bound the term

$$\sum_{i=0}^{N-1}\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2,$$

in terms of more tractable terms. In this step we will prove

Lemma 4.12 *There exists $C > 0$ such that,*

$$\max_{i\in\{0,\dots,N-1\}} \mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 \leq C\Bigg[h + \sum_{i=0}^{N-1}(N\varepsilon_i^{v,\kappa} + \varepsilon_i^{\zeta,\kappa} + \varepsilon_i^{\gamma,\kappa}) + e(Z,(\overline{Z}_t)_{t\in\pi}) + e(\Gamma,(\overline{\Gamma}_t)_{t\in\pi}) + \mathbb{E}\left|g(X_T) - g(X^\pi_T)\right|^2\Bigg], \tag{4.33}$$

with $\varepsilon_i^{v,\kappa}$, $\varepsilon_i^{\zeta,\kappa}$ and $\varepsilon_i^{\gamma,\kappa}$ defined in (4.18).

In what follows, we will prove 4.33.

PROOF. Fix $i \in \{0,...,N-1\}$. Recall the martingale $(N_t)_{t\in[t_i,t_{i+1}]}$ and take $t = t_{i+1}$,

$$\widehat{u}_{i+1}(X^\pi_{t_{i+1}}) = \mathbb{E}_i\left(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right) + \int_{t_i}^{t_{i+1}} \widehat{Z}_s \cdot dW_s + \int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d} \widehat{U}_s(y)\overline{\mu}(ds,dy).$$

Now we replace the definition of $\widehat{\mathcal{V}}_{t_i}$,

$$\widehat{u}_{i+1}(X^\pi_{t_{i+1}}) = \widehat{\mathcal{V}}_{t_i} - f(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}, \overline{\widehat{\Gamma}}_{t_i})h + \int_{t_i}^{t_{i+1}} \widehat{Z}_s \cdot dW_s + \int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d} \widehat{U}_s(y)\overline{\mu}(ds,dy). \tag{4.34}$$

In what follows recall the value of F in the loss function $L_i(\theta)$ (4.10) evaluated at the point

$$(t_i, X^\pi_{t_i}, u_i^\theta(X^\pi_{t_i}), z_i^\theta(X^\pi_{t_i}), w_i^\theta(X^\pi_{t_i};\cdot), h, \Delta W_i),$$

and that $\langle w_i^\theta\rangle(X_{t_i})$ is given in (4.9):

$$\begin{aligned}
&F\left(t_i, X^\pi_{t_i}, u_i^\theta(X^\pi_{t_i}), z_i^\theta(X^\pi_{t_i}), w_i^\theta(X^\pi_{t_i},\cdot), h, \Delta W_i\right)\\
&= u_i^\theta(X^\pi_{t_i}) - hf(t_i, X^\pi_{t_i}, u_i^\theta(X^\pi_{t_i}), z_i^\theta(X^\pi_{t_i}), \langle w_i^\theta\rangle_i(X_{t_i})) + z_i^\theta(X^\pi_{t_i})\cdot\Delta W_i + \int_{\mathbb{R}^d} w_i^\theta(X^\pi_{t_i},y)\overline{\mu}\big((t_i,t_{i+1}],dy\big).
\end{aligned}$$

Now fix a parameter θ and replace (4.34) on $L_i(\theta)$:

$$\begin{aligned}
&\mathbb{E}\Big|\widehat{u}_{i+1}(X^\pi_{t_{i+1}}) - F(t_i, X^\pi_{t_i}, u_i^\theta(X^\pi_{t_i}), z_i^\theta(X^\pi_{t_i}), w_i^\theta(X^\pi_{t_i},\cdot), \Delta t_i, \Delta W_i)\Big|^2\\
&= \mathbb{E}\Big|\widehat{\mathcal{V}}_{t_i} - f(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}, \overline{\widehat{\Gamma}}_{t_i})h + \int_{t_i}^{t_{i+1}} \widehat{Z}_s\cdot dW_s + \int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d}\widehat{U}_s(y)\overline{\mu}(ds,dy) - u_i^\theta(X^\pi_{t_i})\\
&\quad + hf(t_i, X^\pi_{t_i}, u_i^\theta(X_{t_i}), z_i^\theta(X_{t_i}), \langle w_i^\theta\rangle(X_{t_i})) - z_i^\theta(X^\pi_{t_i})\cdot\Delta W_i - \int_{\mathbb{R}^d} w_i^\theta(X^\pi_{t_i},y)\overline{\mu}(\Delta t_i,dy)\Big|^2\\
&= \mathbb{E}\Bigg|\left[\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X^\pi_{t_i}) + h\left(f(t_i, X^\pi_{t_i}, u_i^\theta(X^\pi_{t_i}), z_i^\theta(X^\pi_{t_i}), \langle w_i^\theta\rangle(X^\pi_{t_i})) - f(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}, \overline{\widehat{\Gamma}}_{t_i})\right)\right]\\
&\quad + \left[\int_{t_i}^{t_{i+1}} \widehat{Z}_s\cdot dW_s - \int_{t_i}^{t_{i+1}} z_i^\theta(X^\pi_{t_i})\cdot dW_s + \int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d}\widehat{U}(s,y)\overline{\mu}(ds,dy) - \int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d} w_i^\theta(X^\pi_{t_i},y)\overline{\mu}(ds,dy)\right]\Bigg|^2\\
&= \mathbb{E}\left|a+b\right|^2.
\end{aligned}$$

Note that b is a sum of martingale's differences and therefore $\mathbb{E}_i(b) = 0$. By independence of μ with W, we can deduce that

$$\mathbb{E}(b^2) = \mathbb{E}\left(\int_{t_i}^{t_{i+1}} [\widehat{Z}_s - z_i^\theta(X_{t_i}^\pi)]dW_s\right)^2 + \mathbb{E}\left(\int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d} [\widehat{U}(s,y) - w_i^\theta(X_{t_i}^\pi, y)]\overline{\mu}(ds,dy)\right)^2;$$

and, since the random variables that appears on a are $\mathcal{F}_{t_i}$-measurable, $\mathbb{E}(ab) = \mathbb{E}\left(\mathbb{E}_i(ab)\right) = \mathbb{E}\left(a\mathbb{E}_i(b)\right) = 0$, we have that

$$\begin{aligned} L_i(\theta) = \mathbb{E}\left(\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi) + h\left[f(t_i, X_{t_i}^\pi, u_i^\theta(X_{t_i}^\pi), z_i^\theta(X_{t_i}^\pi), \langle w_i^\theta\rangle(X_{t_i}^\pi))) - f(t_i, X_{t_i}^\pi, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}, \overline{\widehat{\Gamma}}_{t_i})\right]\right)^2 \\ \underbrace{+\,\mathbb{E}\left(\int_{t_i}^{t_{i+1}} [\widehat{Z}_s - z_i^\theta(X_{t_i}^\pi)]dW_s\right)^2 + \mathbb{E}\left(\int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d} [\widehat{U}(s,y) - w_i^\theta(X_{t_i}^\pi, y)]\overline{\mu}(ds,dy)\right)^2}_{c_0}. \end{aligned}$$

By the same arguments on equations (4.26) and (4.27),

$$\begin{aligned} c_0 = \mathbb{E}\left(\int_{t_i}^{t_{i+1}} |\widehat{Z}_s - \overline{\widehat{Z}}_{t_i}|^2 ds\right) + h\mathbb{E}\left|\overline{\widehat{Z}}_{t_i} - z_i^\theta(X_{t_i}^\pi)\right|^2 \\ + \mathbb{E}\left(\int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d} |\widehat{U}_s(y) - \overline{\widehat{U}}_{t_i}(y)|^2 \lambda(dy)ds\right) + h\mathbb{E}\left(\int_{\mathbb{R}^d} \left(\overline{\widehat{U}}_{t_i}(y) - w_i^\theta(X_{t_i}^\pi, y)\right)^2 \lambda(dy)\right). \end{aligned}$$

With this decomposition of $L_i(\theta)$, for optimization reasons, we can ignore the part that does not depend on the optimization parameter θ. Let

$$\begin{aligned} &\hat{L}_i(\theta) \\ &= \mathbb{E}\left(\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi) + h\left[f(t_i, X_{t_i}^\pi, u_i^\theta(X_{t_i}^\pi), z_i^\theta(X_{t_i}^\pi), \langle w_i^\theta\rangle(X_{t_i}^\pi))) - f(t_i, X_{t_i}^\pi, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}, \overline{\widehat{\Gamma}}_{t_i})\right]\right)^2 \\ &\quad + h\,\mathbb{E}\left|\overline{\widehat{Z}}_{t_i} - z_i^\theta(X_{t_i}^\pi)\right|^2 + h\,\mathbb{E}\left(\int_{\mathbb{R}^d} \left(\overline{\widehat{U}}_{t_i}(y) - w_i^\theta(X_{t_i}^\pi, y)\right)^2 \lambda(dy)\right). \end{aligned}$$

Let $\gamma > 0$ and use Young inequality and the Lipschitz condition on f to find that

$$\begin{aligned} &\mathbb{E}\left(\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi) + h\left[f(t_i, X_{t_i}^\pi, u_i^\theta(X_{t_i}^\pi), z_i^\theta(X_{t_i}^\pi), \langle w_i^\theta\rangle(X_{t_i}^\pi))) - f(t_i, X_{t_i}^\pi, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}, \overline{\widehat{\Gamma}}_{t_i})\right]\right)^2 \\ &\leq (1+\gamma h)\,\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi)\right|^2 \\ &\quad + \left(1 + \frac{1}{\gamma h}\right) h^2 K^2 \mathbb{E}\left(|\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi)|^2 + |z_i^\theta(X_{t_i}^\pi) - \overline{\widehat{Z}}_{t_i}|^2 + |\langle w_i^\theta\rangle(X_{t_i}^\pi) - \overline{\widehat{\Gamma}}_{t_i}|^2\right) \\ &\leq (1+Ch)\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi)\right|^2 + Ch\left[\mathbb{E}|z_i^\theta(X_{t_i}^\pi) - \overline{\widehat{Z}}_{t_i}|^2 + \mathbb{E}\left(\left\|\overline{\widehat{U}}_{t_i}(\cdot) - w_i^\theta(X_{t_i}^\pi, \cdot)\right\|^2_{L^2(\mathbb{R}^d,\lambda)}\right)\right]. \end{aligned}$$

Therefore, we have an upper bound on $L(\theta)$ for all θ

$$\hat{L}(\theta) \leq C\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi)\right|^2 + h\left(\mathbb{E}|z_i^\theta(X_{t_i}^\pi) - \overline{\widehat{Z}}_{t_i}|^2\right) + h\mathbb{E}\left(\left\|\overline{\widehat{U}}_{t_i}(\cdot) - w_i^\theta(X_{t_i}^\pi, \cdot)\right\|^2_{L^2(\lambda)}\right).$$

To find a lower bound, we use $(a+b)^2 \geq (1-\gamma h)a^2 + \left(1-\frac{1}{\gamma h}\right)b^2$ with $\gamma > 0$

$$\mathbb{E}\left(\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi) + h\left[f(t_i, X_{t_i}^\pi, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}, \overline{\widehat{\Gamma}}_{t_i}) - f(t_i, X_{t_i}^\pi, u_i^\theta(X_{t_i}^\pi), z_i^\theta(X_{t_i}^\pi), \langle w_i^\theta\rangle(X_{t_i}^\pi))\right]\right)^2$$
$$\geq (1-Ch)\,\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi)\right|^2 - \frac{h}{2}\left(\mathbb{E}|z_i^\theta(X_{t_i}^\pi) - \overline{\widehat{Z}}_{t_i}|^2 + \mathbb{E}|\langle\mathcal{G}\rangle_i(X_{t_i}^\pi;\theta) - \overline{\widehat{\Gamma}}_{t_i}|^2\right);$$

where we used $\gamma = 6K^2$. Then,

$$\hat{L}(\theta) \geq (1-Ch)\,\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi)\right|^2 - \frac{h}{2}\left[\mathbb{E}|z_i^\theta(X_{t_i}^\pi) - \overline{\widehat{Z}}_{t_i}|^2 + \mathbb{E}\left(\int_{\mathbb{R}^d}\left(\overline{\widehat{U}}_{t_i}(y) - w_i^\theta(X_{t_i}^\pi, y)\right)^2\lambda(dy)\right)\right].$$

Connecting this bounds using that $\hat{L}(\theta^*) \leq \hat{L}(\theta)$ yields that $\forall\theta$,

$$(1-Ch)\,\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u_i^{\theta^*}(X_{t_i}^\pi)\right|^2 + \frac{h}{2}\mathbb{E}|\overline{\widehat{Z}}_{t_i} - z_i^{\theta^*}(X_{t_i})|^2 + \frac{h}{2}\mathbb{E}\left(\int_{\mathbb{R}^d}\left(\overline{\widehat{U}}_{t_i}(y) - w_i^{\theta^*}(X_{t_i}^\pi, y)\right)^2\lambda(dy)\right)$$
$$\leq C\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u_i^\theta(X_{t_i}^\pi)\right|^2 + h\left(\mathbb{E}|\overline{\widehat{Z}}_{t_i} - z_i^\theta(X_{t_i}^\pi)|^2\right) + h\mathbb{E}\left(\left\|\overline{\widehat{U}}_{t_i}(\cdot) - w_i^\theta(X_{t_i}^\pi, \cdot)\right\|^2_{L^2(\mathbb{R}^d,\lambda)}\right).$$

By taking infimum on the right side and h small such that $(1-Ch) \geq \frac{1}{2}$

$$\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - \widehat{u}_i(X_{t_i}^\pi)\right|^2 + \frac{h}{2}\mathbb{E}|\overline{\widehat{Z}}_{t_i} - \widehat{z}_i(X_{t_i})|^2 + \frac{h}{2}\mathbb{E}\left(\int_{\mathbb{R}^d}\left(\overline{\widehat{U}}_{t_i}(y) - \widehat{w}_i(X_{t_i}^\pi, y)\right)^2\lambda(dy)\right)$$
$$\leq C\left(\varepsilon_i^{v,\kappa} + h\varepsilon_i^{\zeta,\kappa} + h\varepsilon_i^{\gamma,\kappa}\right). \tag{4.35}$$

Using this bound on what we found on steps 1 and 2, we find

$$\max_{i\in\{0,\dots,N-1\}}\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X_{t_i}^\pi)\right|^2 \leq C\left[h + \sum_{i=0}^{N-1}(N\varepsilon_i^{v,\kappa} + \varepsilon_i^{\zeta,\kappa} + \varepsilon_i^{\gamma,\kappa}) + \sum_{i=0}^{N-1}\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|Y_s - Y_{t_i}|^2 ds\right)\right.$$
$$\left. + e(Z, (\overline{Z}_t)_{t\in\pi}) + e(\Gamma, (\overline{\Gamma}_t)_{t\in\pi}) + \mathbb{E}\left|g(X_T) - g(X_T^\pi)\right|^2\right].$$

Finally, using Proposition 4.5, one ends the proof of (4.33). □

Step 4

We are going to show some bounds for the terms involving the Γ and U components, the same bounds holds for Z component and are shown in [HPW19]. By using (4.30) on (4.27),

$$\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|\Gamma_t - \overline{\widehat{\Gamma}}_{t_i}|^2 dt\right) \leq \mathbb{E}\left(\int_{t_i}^{t_{i+1}}|\Gamma_t - \overline{\Gamma}_{t_i}|^2 dt\right) + 2\lambda(\mathbb{R}^d)\left(\mathbb{E}\left(H_{i+1}^2\right) - \mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right)$$
$$+ 2h\lambda(\mathbb{R}^d)\mathbb{E}\left(\int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr\right),$$

this implies, after using (4.12) and (4.6),

$$\mathbb{E}\left(\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}|\Gamma_t-\overline{\widehat{\Gamma}}_{t_i}|^2dt\right)\leq\mathbb{E}\left(\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}|\Gamma_t-\overline{\Gamma}_{t_i}|^2dt\right)+C\sum_{i=0}^{N-1}\left(\mathbb{E}\left(H_{i+1}^2\right)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right)+Ch$$
$$=e(\Gamma,(\Gamma_t)_{t\in\pi})+Ch+C\sum_{i=0}^{N-1}\left(\mathbb{E}\left(H_{i+1}^2\right)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right).$$

From [HPW19] we get the analogous bound for the Z component, therefore, putting this two together yields

$$\mathbb{E}\left(\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\left(\|Z_t-\overline{\widehat{Z}}_{t_i}\|^2+|\Gamma_t-\overline{\widehat{\Gamma}}_{t_i}|^2\right)dt\right)\leq\quad e(Z,(\overline{Z}_t)_{t\in\pi})+e(\Gamma,(\overline{\Gamma}_t)_{t\in\pi})$$
$$+Ch+C\sum_{i=0}^{N-1}\left(\mathbb{E}\left(H_{i+1}^2\right)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right).\tag{4.36}$$

This tells us that the next mission in this proof is to give a suitable bound for $\mathbb{E}\left(H_{i+1}^2\right)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2$. Recall from (4.28) that $H_{i+1}=Y_{t_{i+1}}-\widehat{u}_{i+1}(X^\pi_{t_{i+1}})$, then

$$\begin{aligned}\sum_{i=0}^{N-1}\left(\mathbb{E}\left(H_{i+1}^2\right)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right)&=\sum_{i=0}^{N-1}\mathbb{E}(H_{i+1}^2)-\sum_{i=0}^{N-1}\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\\&=\mathbb{E}\left|Y_{t_N}-\widehat{u}_N(X^\pi_{t_N})\right|+\sum_{i=0}^{N-2}\mathbb{E}(H_{i+1}^2)-\sum_{i=0}^{N-1}\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\\&\leq\mathbb{E}\left|Y_{t_N}-\widehat{u}_N(X^\pi_{t_N})\right|+\mathbb{E}(H_0^2)+\sum_{i=1}^{N-1}\mathbb{E}(H_i^2)-\sum_{i=0}^{N-1}\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\\&=\mathbb{E}\left|g(X_T)-g(X_T^\pi)\right|^2+\sum_{i=0}^{N-1}\left(\mathbb{E}(H_i^2)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right).\end{aligned}\tag{4.37}$$

From (4.32) and (4.24) we have an upper and lower bound on $\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2$. Indeed, first one has

$$(1-h)\mathbb{E}\left|Y_{t_i}-\widehat{u}_i(X^\pi_{t_i})\right|^2\leq\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2+\left(\frac{1}{h}-1\right)\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i})-\widehat{\mathcal{V}}_{t_i}\right|^2.\tag{4.38}$$

Second, we have that for all $\gamma>0$

$$(1-h)\ \mathbb{E}\left|Y_{t_i}-\widehat{u}_i(X^\pi_{t_i})\right|^2\leq\left(\frac{1}{h}-1\right)\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i})-\widehat{\mathcal{V}}_{t_i}\right|^2+(1+\gamma h)\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2$$
$$+(1+\gamma h)\frac{C}{\gamma}\underbrace{\left[h^2+\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|Y_s-Y_{t_i}|^2ds\right)+h\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2+\mathbb{E}\left(\int_{t_i}^{t_{i+1}}\|Z_s-\overline{\widehat{Z}}_{t_i}\|^2ds\right)+\mathbb{E}\left(\int_{t_i}^{t_{i+1}}|\Gamma_s-\overline{\widehat{\Gamma}}_{t_i}|^2ds\right)\right]}_{B_i}.$$

Let us call the expression inside the squared brackets by B_i. Subtracting $(1-h)\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2$ and dividing by $(1-h)$,

$$\mathbb{E}(H_i^2)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\leq\frac{1}{h}\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i})-\widehat{\mathcal{V}}_{t_i}\right|^2+\left(\frac{h+\gamma h}{1-h}\right)\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2+\frac{C}{\gamma}\frac{(1+\gamma h)}{(1-h)}B_i.$$

For $\gamma = 3C$ and sufficiently small h, we can force,

$$\frac{C}{\gamma}\frac{(1+\gamma h)}{(1-h)} \leq \frac{1}{2} \qquad \text{and} \qquad \frac{1}{1-h} \leq \frac{1}{2}.$$

Hence,

$$\mathbb{E}(H_i^2) - \mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \leq \frac{1}{h}\mathbb{E}\left|\widehat{u}_i(X_{t_i}^\pi) - \widehat{\mathcal{V}}_{t_i}\right|^2 + Ch\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 + \frac{1}{2}B_i.$$

Finally, note that,

$$\sum_{i=0}^{N-1}\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \leq \mathbb{E}\left|g(X_T) - g(X_T^\pi)\right|^2 + N\max_{i=0,\dots,N-1}\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X_{t_i}^\pi)\right|^2. \tag{4.39}$$

Remark 4.13 *Note that in equation (4.39) appears N multiplying the last term. With the bounds that we have, is impossible to get rid of the N, and this is why the δ improvement mentioned on Remark 4.11 will not be of much help.*

Coming back to (4.37),

$$\sum_{i=0}^{N-1}\left(\mathbb{E}\left(H_{i+1}^2\right) - \mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right) \leq 2\mathbb{E}\left|g(X_T) - g(X_T^\pi)\right|^2 + N\sum_{i=0}^{N-1}\mathbb{E}\left|\widehat{u}_i(X_{t_i}^\pi) - \widehat{\mathcal{V}}_{t_i}\right|^2$$
$$+ ChN\max_{i=0,\dots,N-1}\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X_{t_i}^\pi)\right|^2 + \frac{1}{2}\sum_{i=0}^{N-1}B_i.$$

Therefore, by plugging this bound in (4.36), noting that $|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}|^2 \leq 2|Y_{t_i} - \widehat{u}_i(X_{t_i}^\pi)|^2 + 2|\widehat{u}_i(X_{t_i}^\pi) - \widehat{\mathcal{V}}_{t_i}|^2$, $hN = 1$, and using Lemma 4.5, we have for some $C > 0$,

$$\mathbb{E}\Big(\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}(\|Z_t - \overline{\widehat{Z}}_{t_i}\|^2 + |\Gamma_t - \overline{\widehat{\Gamma}}_{t_i}|^2)dt\Big) \leq C\Big[\mathbb{E}\left|g(X_T) - g(X_T^\pi)\right|^2 + e(Z,(\overline{Z}_t)_{t\in\pi}) + e(\Gamma,(\overline{\Gamma}_t)_{t\in\pi}) + h$$
$$+ N\sum_{i=0}^{N-1}\mathbb{E}\left|\widehat{u}_{t_i}(X_{t_i}^\pi) - \widehat{\mathcal{V}}_{t_i}\right|^2 + \max_{i=0,\dots,N-1}\mathbb{E}\left|Y_{t_i} - \widehat{u}_{t_i}(X_{t_i}^\pi)\right|^2\Big].$$

Now, use (4.35) together with Lemma 4.12 to get

$$\mathbb{E}\Big(\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\left(\|Z_t - \overline{\widehat{Z}}_{t_i}\|^2 + |\Gamma_t - \overline{\widehat{\Gamma}}_{t_i}|^2\right)dt\Big) \leq C\Big[\mathbb{E}\left|g(X_T) - g(X_T^\pi)\right|^2 + e(Z,(Z_t)_{t\in\pi}) + e(\Gamma,(\Gamma_t)_{t\in\pi})$$
$$+ h + \sum_{i=0}^{N-1}(N\varepsilon_i^{v,\kappa} + \varepsilon_i^{\zeta,\kappa} + \varepsilon_i^{\gamma,\kappa})\Big].$$

Again, recalling (4.35) using the previous bound and,

$$\sum_{i=0}^{N-1}\mathbb{E}\left(\int_{t_i}^{t_{i+1}}\left[\|Z_t - \widehat{z}_i(X_{t_i}^\pi)\|^2|\Gamma_t - \langle\widehat{w_i}\rangle(X_{t_i}^\pi)|^2\right]dt\right)$$
$$\leq \sum_{i=0}^{N-1}\mathbb{E}\left(\int_{t_i}^{t_{i+1}}\left[|Z_t - \overline{\widehat{Z}}_{t_i}|^2 + |\Gamma_t - \overline{\widehat{\Gamma}}_{t_i}|^2\right]dt\right) + \sum_{i=0}^{N-1}h\mathbb{E}\left(\left[\|\overline{\widehat{Z}}_{t_i} - \widehat{z}_i(X_{t_i}^\pi)\|^2 + \left\|\overline{\widehat{U}}_{t_i}(\cdot) - \widehat{w}_i(X_{t_i}^\pi,\cdot)\right\|^2_{L^2(\mathbb{R}^d,\lambda)}\right]dt\right),$$

we conclude that there exist $C > 0$, independent of the partition, such that for h sufficiently small,

$$\max_{i=0,\dots,N-1} \mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 + \sum_{i=0}^{N-1} \mathbb{E}\left(\int_{t_i}^{t_{i+1}} \left[\|Z_t - \widehat{z}_i(X^\pi_{t_i})\|^2 + |\Gamma_t - \langle \widehat{w}_i\rangle(X^\pi_{t_i})|^2\right] dt\right)$$
$$\leq C\left[h + \sum_{i=0}^{N-1}(N\varepsilon_i^{v,\kappa} + \varepsilon_i^{\zeta,\kappa} + \varepsilon_i^{\gamma,\kappa}) + e(Z, (\overline{Z}_t)_{t\in\pi}) + e(\Gamma, (\overline{\Gamma}_t)_{t\in\pi}) + \mathbb{E}\left|g(X_T) - g(X^\pi_T)\right|^2\right].$$

Thus it has been demonstrated. □

We state some remarks from the proof.

Remark 4.14 *Note that the terms $\mathcal{E}_i^v$, $\mathcal{E}_i^z$ and $\mathcal{E}_i^\gamma$, can be made arbitrarily small, in view of Lemma 3.10. The challenge here, and in almost every DL algorithm, is that we do not know how many units per layer, i.e., how large κ we need to take in order to achieve a fixed tolerance, we can only ensure the existence of a NN architecture satisfying the approximation property.*

Remark 4.15 *The main difficulty of the adaptation of the proof given in [HPW19], was to give a useful definition of the third NN with the mission of approximate the non local component. This was problematic because we have two options, the first is to define the NN to approximate the whole integral*

$$\int_{\mathbb{R}^d} [u(t_i, X^\pi_{t_i} + \beta(X^\pi_{t_i}, y)) - u(t_i, X^\pi_{t_i})]\lambda(dy),$$

this seems intuitive because this will lead our third NN to approximate the nonlocal part of the PIDE and, therefore, receive one parameter: $X^\pi_{t_i}$. But, we also need to approximate or been able to calculate the stochastic integral

$$\int_{\mathbb{R}^d} [u(t_i, X^\pi_{t_i} + \beta(X^\pi_{t_i}, y)) - u(t_i, X^\pi_{t_i})]\bar{\mu}\left((t_i, t_{i+1}], dy\right),$$

that cannot be done by only knowing the first integral. To overcome this issue, we proposed the idea to approximate what it is inside the integrals and solve the problem of actually integrate this function with another tools.

Remark 4.16 *The non local part of the PIDE (1.3) makes us add a Lévy process, which is a canonical tool when dealing with non local operators such as the one that appears on equation (1.3). This addition results in the natural definition of analogous objects from [HPW19] such as the $\Gamma, \bar{\Gamma}$ components for the nonlocal case.*

Remark 4.17 *The result of the theorem states that the better we can approximate v_i, z_i, γ_i by NN architectures, the better we can approximate $(Y_{t_i}, Z_{t_i}, \Gamma_{t_i})$ by $(\widehat{u}_i(X^\pi_{t_i}), \widehat{z}_i(X^\pi_{t_i}), \langle \widehat{w}_i\rangle(X^\pi_{t_i}))$.*

Remark 4.18 *Because of the finiteness of the measure λ, the case of the Fractional Laplacian mentioned in the introduction is not contained in Theorem 4.7. We hope to extend our results to this case in a forthcoming result.*

4.3.1 Optimization step of the algorithm

In this subsection we give a brief explanation on how to compute the loss function from Algorithm 1 in order to perform it. As usual, we extend the computation of the loss function shown on [HPW19] to our non local case for which we need to introduce the following definitions. For a cadlag process $(C_s)_{s\in[0,T]}$, $\Delta C_s := C_s - C_{s^-}$ stands for the jump of C at time $s \in [0,T]$ and for a process $U \in \mathscr{M}_\mu^2(0,T;\mathbb{R}^d)$ the definition of stochastic integral with respect to μ ([App09, Sections 2 and 4]) is as follows,

$$\int_s^t \int_{\mathbb{R}^d} U(r,y)\mu(ds,dy) := \sum_{r\in(s,t]} U(r,\Delta P_r)\mathbb{1}_{\mathbb{R}^d}(\Delta P_r),$$

where

$$\left(P_s = \int_{\mathbb{R}^d} x\mu(s,dx)\right)_s,$$

is a compound Poisson process (see [App09, Thm 2.3.10]). And therefore,

$$\int_s^t \int_{\mathbb{R}^d} U(r,y)\bar{\mu}(ds,dy) = \sum_{r\in(s,t]} U(r,\Delta P_r)\mathbb{1}_{\mathbb{R}^d}(\Delta P_r) - \int_s^t \int_{\mathbb{R}^d} U(r,y)\lambda(dy)dr.$$

For simplicity assume that λ is a probability measure absolutely continuous with respect to Lebesgue measure. As we will see, several simulation of Lévy process $(X_t)_{t\in[0,T]}$ are needed.

As shown on Algorithm 1, given $\widehat{u}_{i+1}$ for $i \in \{0,...,N-1\}$, we need to minimize $L_i(\cdot)$ and define the NNs for step i. Recall the definition of L_i in (4.10), the idea is to write the expected value from the loss function as an average of simulations. Let $M \in \mathbb{N}$ and $I = \{1,...,M\}$, generate simulations $\{x_k^i : k \in I\}$, $\{x_k^{i+1} : k \in I\}$, $\{\delta w_k : k \in I\}$ of $X_{t_i}^\pi$, $X_{t_{i+1}}^\pi$ and ΔW_i respectively. Then,

$$L_i(\theta) \approx \frac{1}{M}\sum_{k\in I}\left(\widehat{u}_{i+1}(x_k^{i+1}) - F(t_i, x_k^i, u_i^\theta(x_k^i), z_i^\theta(x_k^i), w_i^\theta(x_k^i,\cdot), h, \delta w_i)\right)^2.$$

Note that we are using an Euler scheme on the simulations of $(X_t)_{t\in[0,T]}$, nevertheless, there exists other methods depending on the structure of the diffusion, see [BR11] and [KHT10]. Recall that F needs two different integrals of $w_i^\theta(x_k^i,\cdot)$, to approximate these values let $L \in \mathbb{N}$ and $J = \{1,...,L\}$ and consider, for every $k \in I$, simulations $\{y_l^k : l \in J\}$ of a random variable $Y \sim \lambda$, here is important the finitness of the measure. Then, the quantities we need can be computed as follows,

$$\int_{\mathbb{R}^d} w_i^\theta(x_k^i,y)\lambda(dy) = \mathbb{E}(w_i^\theta(x_k^i,y)) \approx \frac{1}{L}\sum_{l\in J} w_i^\theta(x_k^i,y_l^k)$$

$$\begin{aligned}\int_{\mathbb{R}^d} w_i^\theta(x_k^i,y)\bar{\mu}\left((t_i,t_{i+1}],dy\right) &= \int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d} w_i^\theta(x_k^i,y)\mu\left(dt,dy\right) - \int_{t_i}^{t_{i+1}}\int_{\mathbb{R}^d} w_i^\theta(x_k^i,y)dt\lambda(dy)\\ &\approx \sum_{t_i\le s<t_{i+1}} w_i^\theta(x_k^i,\Delta P_s)\mathbb{1}_{\mathbb{R}^d}(\Delta P_s) - \frac{h}{L}\sum_{l\in J} w_i^\theta(x_k^i,y_l^k).\end{aligned}$$

Therefore, provided we can simulate: trajectories of $(X_t)_{t\in[0,T]}$ and $(W_t)_{t\in[0,T]}$, realizations of $Y \sim \lambda$ and the compound Poisson process $(P_t)_{t\in[0,T]}$, we can minimize L_i, find the optimal θ^* and define

$$(\widehat{u}_i, \widehat{z}_i, \widehat{w}_i) = (u_i^{\theta^*}, z_i^{\theta^*}, w_i^{\theta^*}).$$

Remark 4.19 *The nonlocal term in equation (1.3) adds complexity not only in the proof of the consistency of the algorithm but in the algorithm itself. As we saw, it is key that the measure λ is finite as well as the capability to simulate integrals with respect to Poisson random measures and trajectories of the Lévy process. The implementation of this method and an extension to PIDEs with more general integro-differential operators, such as fractional Laplacian, are left to future work.*

Chapter 5

Kolmogorov equation posed on a Hilbert space

In this chapter we introduce the scheme for the equation presented in Section 1.1.3. As well as in the previous chapter, in Section 5.1 we apply Itô lemma to a strong solution of (1.9) deriving an recursive algorithm. In Section 5.2 we define auxiliary processes that are necessary for starting the proof of the main result and prove properties of these. Finally, Section 5.3 is devoted to the consistency proof of the presented scheme.

5.1 Functional Numerical Scheme

Throughout this section we will work with functions that we call *approximators* and are parameterized by a finite dimensional parameter $\theta \in \Theta_\eta \subset \mathbb{R}^\eta$ for some $\eta \in \mathbb{N}$, also let $\Theta = \cup_{\eta\in\mathbb{N}}\Theta_\eta$. As the reader may anticipate, these functions will be the DeepOnets introduced in Section 3.2. We work in generality first, to then apply our results to this particular case.

The following is a key assumption for the validity of our main results.

Assumptions 5.1 *Assume we are given a function $u \in C^{1,2}([0,T] \times H)$ satisfying (1.9) and a strong solution $(X_t)_{t\in[0,T]}$ to (1.11).*

This assumption is natural in finite dimensions, but its validity in infinite dimensions is far from obvious. The scheme presented here is fully inspired by [HPW19] and relies on an application of Itô Lemma to $(u(t, X_t))_{t\in[0,T]}$ as follows (see [DPZ92, Theorem 4.32]),

$$
\begin{aligned}
&u(t, X_t)\\
&= u(0, X_0) + \int_0^t \langle \nabla u(s, X_s), B(s, X_s)(\cdot)\rangle_H dW_s - \int_0^t \psi\left(s, X_s, u(s, X_s), B^*(s, X_s)\nabla u(s, X_s)\right) ds\\
&= u(0, X_0) + \int_0^t \langle B^*(s, X_s)\nabla u(s, X_s), \cdot\rangle_0 dW_s - \int_0^t \psi\left(s, X_s, u(s, X_s), B^*(s, X_s)\nabla u(s, X_s)\right) ds.
\end{aligned}
$$

Consider now a uniform partition $\pi = \{t_0 = 0, ..., t_N = T\}$ with $t_i = \frac{iT}{N}$ such that $h = t_{i+1} - t_i > 0$ for all $i \in \{0, ..., N-1\}$, then

$$u(t_{i+1}, X_{t_{i+1}}) = u(t_i, X_{t_i}) + \int_{t_i}^{t_{i+1}} \langle B^*(s, X_s)\nabla u(s, X_s), \cdot\rangle_0 dW_s$$
$$- \int_{t_i}^{t_{i+1}} \psi\left(s, X_s, u(s, X_s), B^*(s, X_s)\nabla u(s, X_s)\right) ds.$$

Let $\eta \in \mathbb{N}$ be a fixed natural number and let $\Theta_\eta \subset \mathbb{R}^\eta$ be also a fixed set. Now, let us introduce some *approximators* as a collection of mappings $u_i^\theta \colon H \to \mathbb{R}$ for $i \in \{0, ..., N\}$ and $z_i^\theta \colon H \to V_0$ for $i \in \{0, ..., N-1\}$. Additionally, consider an scheme $X^\pi = (X_t^\pi)_{t\in\pi}$ for the equation (1.11) which we assume satisfies $\sigma(X_s^\pi \colon s \le t, s \in \pi) \subset \mathcal{F}_t$, $X_t^\pi \in L^4(\Omega, \mathcal{F}_t, \mathbb{P}; H)$ for $t \in \pi$. Here X^π is a Markov process. These approximators are assumed to be such that $\left\{u_i^\theta\right\}_{\theta\in\Theta}$ and $\left\{z_i^\theta\right\}_{\theta\in\Theta}$ are dense in $L^2(H, \mu_{X_{t_i}^\pi})$ and $L^2(H, \mu_{X_{t_i}^\pi}; V_0)$ respectively. Also assume that the approximators has polynomial growth at most.

Remark 5.2 *Hilbert valued DeepOnets are a set of approximators. This is obtained by defining*

$$\Theta_\eta = \bigcup_{d,m\in\mathbb{N}} \{d\} \times \mathcal{N}_{\sigma,7,d,m,\eta} \times \{m\}. \tag{5.1}$$

The size of the hidden layers of the NN (recall Definition 3.4) is the variable that may increase in order to have a better performance of the DO.

We propose a scheme in which we intend to find $\theta \in \Theta_\eta$ such that given $\hat{u}_{i+1}$, the following approximations hold as good as possible:

$$u_i^\theta(\cdot) \approx u(t_i, \cdot)$$
$$z_i^\theta(\cdot) \approx B^*(t_i, \cdot)\nabla u(t_i, \cdot)$$
$$\hat{u}_{i+1}(X_{t_{i+1}}^\pi) \approx u_i^\theta(X_{t_i}^\pi) + \int_{t_i}^{t_{i+1}} \langle z_i^\theta(X_{t_i}^\pi), \cdot\rangle_0 dW_s - \psi\left(t_i, X_{t_i}^\pi, u_i^\theta(X_{t_i}^\pi), z_i^\theta(X_{t_i}^\pi)\right) h,$$

each one in some proper measure for every $i \in \{1, ..., N-1\}$. The above approximations motivates the definition of a cost function, $L_i : \Theta_\eta \to [0, +\infty)$, associated to $\theta \in \Theta_\eta$:

$$L_i(\theta) = \mathbb{E}\left|\hat{u}_{i+1}(X_{t_{i+1}}^\pi) - u_i^\theta(X_{t_i}^\pi) - \int_{t_i}^{t_{i+1}} \langle z_i^\theta(X_{t_i}^\pi), \cdot\rangle_0 dW_s + \psi\left(t_i, X_{t_i}^\pi, u_i^\theta(X_{t_i}^\pi), z_i^\theta(X_{t_i}^\pi)\right) h\right|^2.$$

We present the following algorithm as an infinite-dimension extension of the one already presented in [HPW19] and [Cas21].

Algorithm 2: DBDP1 infinite-dimension extension

Start with $\hat{u}_N = \phi$;
for $i \in \{N-1, ..., 1\}$ **do**
  Given $\widehat{u}_{i+1}$;
  Compute $\theta^* = \underset{\theta\in\Theta_\eta}{\operatorname{argmin}}\, L_i(\theta)$;
  Update $(\widehat{u}_i, \hat{z}_i) = \left(u_i^{\theta^*}, z_i^{\theta^*}\right)$;
end

5.2 Previous Definitions and Results

Let us introduce the operator $\mathbb{E}_i = \mathbb{E}(\cdot|\mathcal{F}_{t_i})$ defined for every integrable real or vector valued random variable. For the consistency proof of the algorithm we need to introduce a somehow auxiliary scheme $(\widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i})_{i\in\{0,\dots,N-1\}}$ that is inspired by [BT04], used in [HPW19] and we generalize to the infinite-dimensional case as follows,

$$\widehat{\mathcal{V}}_{t_i} = \mathbb{E}_i(\hat{u}_{i+1}(X^\pi_{t_{i+1}})) + \psi(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i})h \tag{5.2}$$

$$\overline{\widehat{Z}}_{t_i} = \frac{1}{h}\mathbb{E}_i(\hat{u}_{i+1}(X^\pi_{t_{i+1}})\Delta W_i). \tag{5.3}$$

Observe that these processes are adapted to the discrete filtration $(\mathcal{F}_t)_{t\in\pi}$. The discrete process $\widehat{\mathcal{V}}_{t_i}$ for $i \in \{0, ..., N-1\}$ is well-defined for sufficiently small h as shown in Lemma 5.3 and by Markov property of X^π, there exists square integrable functions v_i, z_i for $i \in \{0, ..., N-1\}$ such that

$$\widehat{\mathcal{V}}_{t_i} = v_i(X^\pi_{t_i}) \quad \text{and} \quad \overline{\widehat{Z}}_{t_i} = z_i(X^\pi_{t_i}).$$

Lemma 5.3 *Assume that for sufficiently small h and every $i \in \{0, ..., N-1\}$, $\mathbb{E}|\hat{u}_{i+1}(X^\pi_{t_{i+1}})|^4 < +\infty$. Then there exists $\widehat{\mathcal{V}}_{t_i} \in L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P})$ such that (5.2) holds and $\overline{\widehat{Z}}_{t_i} \in L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P}; V)$.*

PROOF. Let $i \in \{0, ..., N-1\}$ and $f : L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P}) \to L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P})$ be defined as

$$f(\xi)(\omega) = \mathbb{E}_i\left(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right)(\omega) + \psi\left(t_i, X^\pi_{t_i}(\omega), \xi(\omega), \overline{\widehat{Z}}_{t_i}(\omega)\right)h.$$

For all $\xi \in L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P})$ and $\omega \in \Omega$. This function is well-defined by the properties of ψ and the approximators. Let $\xi, \overline{\xi} \in L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P})$, then $\mathbb{P}$ a.s $|\psi(\xi) - \psi(\overline{\xi})| \le h|\xi - \overline{\xi}|$, therefore

$$\left\|\psi(\xi) - \psi(\overline{\xi})\right\|_{L^2(\Omega,\mathcal{F}_{t_i},\mathbb{P})} \le h\left\|\xi - \overline{\xi}\right\|_{L^2(\Omega,\mathcal{F}_{t_i},\mathbb{P})}.$$

Taking $h < 1$, which is independent of i, we can see that this function is a contraction on $L^2(\Omega, \mathcal{F}, \mathbb{P})$, and therefore, by applying Banach's fixed point theorem, we conclude the first result of this lemma. By standard computations,

$$\begin{aligned}
\mathbb{E}\left\|\overline{\widehat{Z}}_{t_i}\right\|^2_V &= \mathbb{E}\left\|\frac{1}{h}\mathbb{E}_i\left(\hat{u}_{i+1}(X^\pi_{t_{i+1}})\Delta W_i\right)\right\|^2 \\
&\le \frac{1}{h^2}\mathbb{E}\left(\mathbb{E}_i\left\|\hat{u}_{i+1}(X^\pi_{t_{i+1}})\Delta W_i\right\|_V\right)^2 \le \frac{1}{h}\mathbb{E}\left(|\hat{u}_{i+1}(X^\pi_{t_{i+1}})|^2\left\|\Delta W_i\right\|^2_V\right) \\
&\le \frac{1}{h}\sqrt{\mathbb{E}|\hat{u}_{i+1}(X^\pi_{t_{i+1}})|^4}\sqrt{\mathbb{E}\left\|\Delta W_i\right\|^4_V} < \infty,
\end{aligned}$$

where we used the fact that $W_t \in L^4(\Omega, \mathcal{F}, \mathbb{P}; V)$. The proof is completed. □

We intent to write $\overline{\widehat{Z}}_{t_i}$ as the average of some other process on $[t_i, t_{i+1}]$, to be consistent with the overline notation this process has to be denoted as $\widehat{Z}_t$ for $t \in [t_i, t_{i+1}]$.

Lemma 5.4 *There exists a V_0-valued process $(\widehat{Z}_t)_{t\in[t_i,t_{i+1}]}$, which can be seen as an element of $\mathscr{N}_W(t_i,t_{i+1};L_2(V_0,\mathbb{R}))$, such that,*

$$\overline{\widehat{Z}}_{t_i}=\frac{1}{h}\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}}\widehat{Z}_s ds\right)\in L^2(\Omega,\mathcal{F}_{t_i},\mathbb{P};Q^{1/2}V).$$

PROOF. Consider $N_t=\mathbb{E}(\hat{u}_{i+1}(X^\pi_{t_{i+1}})|\mathcal{F}_t)$ for $t\in[t_i,t_{i+1}]$, this process is a square integrable martingale because $\hat{u}_{i+1}(X^\pi_{t_{i+1}})\in L^2(\Omega,\mathcal{F}_{t_{i+1}},\mathbb{P})$. By the martingale representation theorem 2.34 there exists $(\widehat{Z}_t)_{t\in[t_i,t_{i+1}]}\in\mathscr{N}_W(t_i,t_{i+1};L_2(V_0,\mathbb{R}))$, which ensures the a.e. Bochner integrability of $(\widehat{Z}_t)_{t\in[t_i,t_{i+1}]}$, such that,

$$N_t=N_{t_i}+\int_{t_i}^t\langle\widehat{Z}_s,\cdot\rangle_0 dW_s.$$

By taking $t=t_i$,

$$\widehat{u}_{i+1}(X^\pi_{t_{i+1}})=\mathbb{E}_i(\widehat{u}_{i+1}(X^\pi_{t_{i+1}}))+\int_{t_i}^{t_{i+1}}\langle\widehat{Z}_s,\cdot\rangle_0\, dW_s.$$

It follows that,

$$h\overline{\widehat{Z}}_{t_i}=\mathbb{E}_i(\mathbb{E}_i(\widehat{u}_{i+1}(X^\pi_{t_{i+1}}))\Delta W_i)+\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}}\langle\widehat{Z}_s,\cdot\rangle_0\, dW_s\left(W_{t_{i+1}}-W_{t_i}\right)\right).$$

Note that we took the equation from $\mathbb{R}$ to V. We can make the following elimination,

$$\mathbb{E}_i(\mathbb{E}_i(\widehat{u}_{i+1}(X^\pi_{t_{i+1}}))\Delta W_i)=\mathbb{E}_i(\widehat{u}_{i+1}(X^\pi_{t_{i+1}}))\mathbb{E}_i\Delta W_i=0,$$

which yields,

$$h\overline{\widehat{Z}}_{t_i}=\mathbb{E}_i\Big(\int_{t_i}^{t_{i+1}}\langle\widehat{Z}_s,\cdot\rangle_0\, dW_s\left(W_{t_{i+1}}-W_{t_i}\right)\Big).$$

Recall that the representation (2.6) allows us to write $W_{t_{i+1}}-W_{t_i}=\sum_{j=1}^\infty f_j\sqrt{\lambda_j}(\beta_j(t_{i+1})-\beta_j(t_i))$, where the series converges in $L^2(\Omega,\mathcal{F},\mathbb{P};V)$. Therefore, we can take the summation out of $\mathbb{E}_i$,

$$h\overline{\widehat{Z}}_{t_i}=\sum_{j=1}^\infty f_j\sqrt{\lambda_j}\mathbb{E}_i\Big(\int_{t_i}^{t_{i+1}}\langle\widehat{Z}_s,\cdot\rangle_0\, dW_s\int_{t_i}^{t_{i+1}}d\beta_j(s)\Big).$$

Using Lemma 2.33 and the same argument as before with the $L^2(\Omega,\mathcal{F},\mathbb{P})$ limit

$$\int_{t_i}^{t_{i+1}}\langle\widehat{Z}_s,\cdot\rangle_0 dW_s=\lim_{n\to\infty}\int_{t_i}^{t_{i+1}}\langle\widehat{Z}_s,\cdot\rangle_0 dW^n_s=\sum_{k=1}^\infty\int_{t_i}^{t_{i+1}}\langle\widehat{Z}_s,Q^{1/2}f_j\rangle_0 d\beta^k_s,$$

we get,

$$\begin{aligned}h\overline{\widehat{Z}}_{t_i}&=\sum_{j=1}^\infty\sum_{k=1}^\infty\lambda_j^{1/2}f_j\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}}\langle\widehat{Z}_s,Q^{1/2}f_k\rangle_0 d\beta^k_s\int_{t_i}^{t_{i+1}}d\beta^j_s\right)\\&=\sum_{j=1}^\infty\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}}\langle\widehat{Z}_s,Q^{1/2}f_j\rangle_0 Q^{1/2}f_j ds\right).\end{aligned}$$

Where we used conditional Ito isometry. Last step is proving the following limit in $L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P}; V)$,

$$\lim_{n\to\infty} \sum_{j=1}^{n} \int_{t_i}^{t_{i+1}} \langle \widehat{Z}_s, Q^{1/2} f_j \rangle_0 Q^{1/2} f_j ds = \int_{t_i}^{t_{i+1}} \widehat{Z}_s ds.$$

Indeed,

$$\begin{aligned}
&\mathbb{E} \left\| \int_{t_i}^{t_{i+1}} \widehat{Z}_s ds - \sum_{j=1}^{n} \int_{t_i}^{t_{i+1}} \langle \widehat{Z}_s, Q^{1/2} f_j \rangle_0 Q^{1/2} f_j ds \right\|_V^2 \\
&= \mathbb{E} \left\| \int_{t_i}^{t_{i+1}} \sum_{j=n+1}^{\infty} \langle \widehat{Z}_s, Q^{1/2} f_j \rangle_0 Q^{1/2} f_j ds \right\|_V^2 \leq h\mathbb{E} \int_{t_i}^{t_{i+1}} \left\| \sum_{j=n+1}^{\infty} \langle \widehat{Z}_s, Q^{1/2} f_j \rangle_0 Q^{1/2} f_j \right\|_V^2 ds \\
&= h\mathbb{E} \int_{t_i}^{t_{i+1}} \sum_{j=n+1}^{\infty} |\langle \widehat{Z}_s, Q^{1/2} f_j \rangle_0|^2 \langle Q^{1/2} \rangle ds \leq h\mathbb{E} \int_{t_i}^{t_{i+1}} \left\| \widehat{Z}_s \right\|_0^2 ds \left(\sum_{j=n+1}^{\infty} \lambda_j \right),
\end{aligned}$$

which approaches to 0 as $n \to \infty$ because of Q been trace class. □

Recall the uniform partition π with step h from Subsection 5.1 and that $\Delta W_i = W_{t_{i+1}} - W_{t_i}$.

Lemma 5.5 *The following holds:*

$$\mathbb{E}_i \left\| \Delta W_i \right\|_V^2 = tr(Q)h.$$

PROOF. Consider the identity mapping $I_V \colon V \to V$. By Ito isometry, one has that

$$\begin{aligned}
\mathbb{E} \left\| \Delta W_i \right\|_V^2 = \mathbb{E} \left\| \int_{t_i}^{t_{i+1}} I_V dW_s \right\|_V^2 &= \mathbb{E} \int_{t_i}^{t_{i+1}} \left\| I_V \right\|_{L_2(V_0,V)}^2 ds \\
&= h \left\| I_V \right\|_{L_2(V_0,V)}^2 = h \sum_{k=1}^{\infty} \lambda_k = \text{tr}(Q)h.
\end{aligned}$$

□

It is useful to state and prove our main result to consider the following definition:

Definition 5.6 *For $i \in \{0, ..., N-1\}$ let $(M_s)_{s\in[0,T]}$ be an integrable process and $(L_i)_{i\in\{0,...,N-1\}}$ be a set of random variables, all random objects taking values in some Hilbert K. We define,*

$$e_i(M, L_0) = \mathbb{E} \int_{t_i}^{t_{i+1}} \left\| M_s - L_0 \right\|_K^2 ds \qquad \textit{and} \qquad e(M, L) = \sum_{i=0}^{N-1} e_i(M, L_i). \tag{5.4}$$

Also,

$$\overline{Z}_{t_i} = \frac{1}{h} \mathbb{E}_i \int_{t_i}^{t_{i+1}} Z_s ds \ \in \ L^2(\Omega, \mathcal{F}_{t_i}, \mathbb{P}; Q^{1/2}V). \tag{5.5}$$

Let ε_i^v, ε_i^z given by

$$\varepsilon_i^{v,\eta} := \inf_{\theta\in\Theta_\eta} \mathbb{E}|v_i(X_{t_i}^\pi) - u_i^\theta(X_{t_i}^\pi)|^2, \qquad \varepsilon_i^{z,\eta} := \inf_{\theta\in\Theta_\eta} \mathbb{E}\left\|z_i(X_{t_i}^\pi) - z_i^\theta(X_{t_i}^\pi)\right\|_0^2. \tag{5.6}$$

Finally, consider

$$\varepsilon^{v,\eta} = \sum_{i=0}^{N-1} \varepsilon_i^v, \qquad \varepsilon^{z,\eta} = \sum_{i=0}^{N-1} \varepsilon_i^z. \tag{5.7}$$

Previous definitions are related to the error committed in our scheme. Given the previous notation, consider the following assumptions which depends on the behavior of solution (Y, Z) to stochastic equation (1.12) and how good the assumed scheme X^π is.

Assumptions 5.7 *Assume that the processes $(Y, Z) \in \mathscr{S}_T^2(\mathbb{R}) \times \mathscr{N}_W(0, T; L_2^0(V, \mathbb{R}))$ satisfy that there exist $C > 0$ and a function $\rho\colon (0,\infty) \to (0,\infty)$ such that,*

$$e(X, X^\pi) + e(Y, (Y_t)_{t\in\pi}) + e(Z, (\overline{Z}_t)_{t\in\pi}) \leq \rho(h), \tag{5.8}$$

where $\rho(h) \to 0$ as $h \to 0$.

This assumption holds in the finite dimensional case, where the control on regularity is precise and stipulated as a $\mathcal{O}(h)$. See e.g. [BE08, Theorem 2.1]. Note that in general the distance used to measure the component related to Y is always expressed in a L^∞-type of distance. Meanwhile, terms related to Z are measured in L^2-type of measure. The following is a typical but essential preliminary result:

Lemma 5.8 *Let $(X_t)_{t\in[0,T]}$ be such that $\sup_{s\in[0,T]} \mathbb{E}\|X_s\|_H^2 < \infty$ and $(Y, Z) \in \mathscr{S}_T^2(\mathbb{R}) \times \mathscr{N}_W(0, T; L_2^0(V))$. The following bound holds,*

$$\mathbb{E}\left(\int_0^T \psi(s, X_s, Y_s, Z_s)^2 ds\right) < \infty$$

Proof. First note that

$$\int_0^T \mathbb{E}\|X_s\|_H^2\, ds \leq T \sup_{s\in[0,T]} \mathbb{E}\|X_s\|_H^2 < \infty,$$

then, Fubini theorem can be applied together with the Lipschitz condition on ψ to get,

$$\begin{aligned}
\mathbb{E}\int_0^T \psi(s, X_s, Y_s, Z_s)^2 ds &\leq 2K\mathbb{E}\int_0^T (s + \|X_s\|_H^2 + |Y_s|^2 + \|Z_s\|_0^2) ds + 2T\psi(0,0,0,0)^2 \\
&\leq \frac{CT^2}{2} + CT \sup_{s\in[0,T]} \mathbb{E}\|X_s\|_H^2 + CT \sup_{s\in[0,T]} |Y_s|^2 + C\mathbb{E}\int_0^T \|Z_s\|_0^2\, ds + CT \\
&\leq C\left(1 + \sup_{s\in[0,T]} \mathbb{E}\|X_s\|_H^2 + \|Y\|_{\mathscr{S}_T^2(\mathbb{R})}^2 + \|Z\|_{\mathscr{N}_W(0,T;L_2^0(V,\mathbb{R}))}^2\right) < \infty.
\end{aligned}$$

Thus, the proof is completed. □

5.3 Main Result

Now we are ready to state and prove the main result of this paper. Recall the properties of approximators in Subsection 5.1.

Theorem 5.9 *Under Assumptions 2.12, 5.1 and 5.7, there exists a constant $C > 0$ independent of the partition such that for sufficiently small h,*

$$\max_{i=0,\dots,N-1} \mathbb{E}\left|Y_{t_i} - \hat{u}_i(X^\pi_{t_i})\right|^2 + \sum_{i=0}^{N-1} \mathbb{E}\left(\int_{t_i}^{t_{i+1}} \left\|Z_t - \hat{z}_i(X^\pi_{t_i})\right\|_0^2 dt\right)$$
$$\leq C\Big[h + \mathbb{E}\left|\phi(X_T) - \phi(X^\pi_T)\right|^2 + N\varepsilon^{v,\eta} + \varepsilon^{z,\eta} + \rho(h)\Big],$$

with $\varepsilon^{v,\eta}$, $\varepsilon^{z,\eta}$ given in (5.7).

PROOF. **Step 1**

Recall $\widehat{\mathcal{V}}_{t_i}$ introduced in (5.2). The purpose of this part is to obtain a suitable bound of the term $\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2$ in terms of more tractable terms. We have

Lemma 5.10 *There exists $C > 0$ fixed such that for any $0 < h < 1$ sufficiently small, one has*

$$\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 \leq Ch^2 + C\mathbb{E}\int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds + C\mathbb{E}\int_{t_i}^{t_{i+1}} \left\|Z_s - \overline{Z}_{t_i}\right\|_V^2 ds + Ch\mathbb{E}\int_{t_i}^{t_{i+1}} \psi(\Theta_r)^2 dr$$
$$+ C(1+Ch)\mathbb{E}\left|Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right|^2, \tag{5.9}$$

with $\Theta_r = (r, X_r, Y_r, Z_r)$.

The rest of this subsection is devoted to the proof of this result.

PROOF. Subtracting the equation (1.12) between t_i and t_{i+1}, we obtain

$$\Delta Y_i = Y_{t_{i+1}} - Y_{t_i} = -\int_{t_i}^{t_{i+1}} \psi(\Theta_s) ds + \int_{t_i}^{t_{i+1}} \langle Z_s, \cdot\rangle_0 dW_s. \tag{5.10}$$

Using the definition of $\widehat{\mathcal{V}}_{t_i}$ in 5.2,

$$\begin{aligned} Y_{t_i} - \widehat{\mathcal{V}}_{t_i} &= Y_{t_{i+1}} - \Delta Y_i - \widehat{\mathcal{V}}_{t_i} \\ &= Y_{t_{i+1}} + \int_{t_i}^{t_{i+1}} [\psi(\Theta_s) - \psi(\widehat{\Theta}_{t_i})] ds - \int_{t_i}^{t_{i+1}} \langle Z_s, \cdot\rangle_0 dW_s - \mathbb{E}_i \widehat{u}_{i+1}(X^\pi_{t_{i+1}}). \end{aligned}$$

Here $\widehat{\Theta}_{t_i} = (t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i})$. Then, by taking $\mathbb{E}_i$ and using that stochastic integration produces a martingale

$$Y_{t_i} - \widehat{\mathcal{V}}_{t_i} = \mathbb{E}_i(Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})) + \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}} [\psi(\Theta_s) - \psi(\widehat{\Theta}_{t_i})] ds\right) = a + b.$$

Using the classical inequality $(a+b)^2 \leq (1+\gamma h)a^2 + (1+\frac{1}{\gamma h})b^2$ for $\gamma > 0$ to be chosen, we get

$$\begin{aligned}\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 \leq (1+\gamma h)\mathbb{E}\left[\mathbb{E}_i\left(Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right)\right]^2 \\ + \left(1+\frac{1}{\gamma h}\right)\mathbb{E}\left[\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}}[\psi(\Theta_s) - \psi(\widehat{\Theta}_{t_i})]ds\right)\right]^2.\end{aligned} \tag{5.11}$$

With no lose of generality, as we are seeking for an upper bound, we can replace $[\psi(\Theta_s) - \psi(\widehat{\Theta}_{t_i})]$ by $|\psi(\Theta_s) - \psi(\widehat{\Theta}_{t_i})|$. Also, in the second term, we can drop the $\mathbb{E}_i$ due to the law of total expectation. The Lipschitz condition on ψ in Assumptions 2.12 allows us to give an upper bound in terms of the difference between Θ_s and $\widehat{\Theta}_{t_i}$. Indeed, we have that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}}[\psi(\Theta_s) - \psi(\widehat{\Theta}_{t_i})]ds\right)\right]^2 \leq Ch\left[h^2 + \mathbb{E}\int_{t_i}^{t_{i+1}}\left\|X_s - X^\pi_{t_i}\right\|_H^2 ds + \mathbb{E}\int_{t_i}^{t_{i+1}}|Y_s - \widehat{\mathcal{V}}_{t_i}|^2 ds\right. \\ \left. + \mathbb{E}\int_{t_i}^{t_{i+1}}\left\|Z_s - \overline{\widehat{Z}}_{t_i}\right\|_V^2 ds\right],\end{aligned}$$

where the Lipschitz constant of ψ was absorbed by C. Using now triangle inequality $|Y_s - \widehat{\mathcal{V}}_{t_i}| \leq |Y_s - Y_{t_i}| + |Y_{t_i} - \widehat{\mathcal{V}}_{t_i}|$ and the definition of e_i in (5.4), we find

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_i\left(\int_{t_i}^{t_{i+1}}[\psi(\Theta_s) - \psi(\widehat{\Theta}_{t_i})]ds\right)\right]^2 \leq Ch\left[h^2 + e_i(X, X^\pi_{t_i}) + e_i(Y, Y_{t_i}) + h\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2\right. \\ \left. + \mathbb{E}\int_{t_i}^{t_{i+1}}\left\|Z_s - \overline{\widehat{Z}}_{t_i}\right\|_V^2 ds\right].\end{aligned} \tag{5.12}$$

For the sake of brevity, define now

$$H_i := Y_{t_i} - \widehat{u}_i(X^\pi_{t_i}). \tag{5.13}$$

Therefore, replacing in (4.22),

$$\begin{aligned}\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 \leq (1+\gamma h)\,\mathbb{E}|\mathbb{E}_i H_{i+1}|^2 + (1+\gamma h)\,\frac{C}{\gamma}\left[h^2 + e_i(X, X^\pi_{t_i}) + e_i(Y, Y_{t_i}) + h\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2\right. \\ \left. + \mathbb{E}\int_{t_i}^{t_{i+1}}\left\|Z_s - \overline{\widehat{Z}}_{t_i}\right\|_V^2 ds\right].\end{aligned} \tag{5.14}$$

Recall $\overline{Z}_{t_i}$ introduced in equation (5.5). In order to work with last term in previous equation, we prove the following,

$$\mathbb{E}\int_{t_i}^{t_{i+1}}\left\|Z_s - \overline{\widehat{Z}}_{t_i}\right\|_V^2 ds = \mathbb{E}\int_{t_i}^{t_{i+1}}\left\|Z_s - \overline{Z}_{t_i}\right\|_V^2 ds + h\mathbb{E}\left\|\overline{Z}_{t_i} - \overline{\widehat{Z}}_{t_i}\right\|_V^2. \tag{5.15}$$

Indeed,

$$\begin{aligned}\left\|Z_t - \overline{\widehat{Z}}_{t_i}\right\|_V^2 &= \left\|(Z_t - \overline{Z}_{t_i}) + (\overline{Z}_{t_i} - \overline{\widehat{Z}}_{t_i})\right\|_V^2 \\ &= \left\|Z_t - \overline{Z}_{t_i}\right\|_V^2 + \left\|\overline{Z}_{t_i} - \overline{\widehat{Z}}_{t_i}\right\|_V^2 + 2\langle Z_t - \overline{Z}_{t_i}, \overline{Z}_{t_i} - \overline{\widehat{Z}}_{t_i}\rangle_V.\end{aligned}$$

It is sufficient to establish that the double product is null when we integrate and take expected valued. Recall that $\overline{Z}_{t_i}$ from (5.5) is a $\mathcal{F}_{t_i}$ measurable random variable. Then, by using elementary properties of Bochner integral,

$$\begin{aligned}\mathbb{E}\int_{t_i}^{t_{i+1}}\langle Z_t-\overline{Z}_{t_i},\overline{Z}_{t_i}-\overline{\widehat{Z}}_{t_i}\rangle_V dt &= \mathbb{E}\Big\langle\int_{t_i}^{t_{i+1}}(Z_s-\overline{Z}_{t_i})ds,\overline{Z}_{t_i}-\overline{\widehat{Z}}_{t_i}\Big\rangle_V\\ &= h\mathbb{E}\Big\langle\frac{1}{h}\int_{t_i}^{t_{i+1}}Z_s ds-\overline{Z}_{t_i},\overline{Z}_{t_i}-\overline{\widehat{Z}}_{t_i}\Big\rangle_V=0.\end{aligned}$$

The latter is due to the fact that $\overline{Z}_{t_i}-\overline{\widehat{Z}}_{t_i}\in L^2(\Omega,\mathcal{F}_{t_i},\mathbb{P};V)$ and $\frac{1}{h}\int_{t_i}^{t_{i+1}}Z_s ds-\overline{Z}_{t_i}$ is an orthogonal element to $L^2(\Omega,\mathcal{F}_{t_i},\mathbb{P};V)\subset L^2(\Omega,\mathcal{F},\mathbb{P};V)$. Therefore, equation (4.26) is established. By multiplying (4.21) by ΔW_i and taking $\mathbb{E}_i$,

$$\begin{aligned}\mathbb{E}_i\left(\Delta W_i Y_{t_{i+1}}\right)+\mathbb{E}_i\left(\Delta W_i\int_{t_i}^{t_{i+1}}\psi(\Theta_s)ds\right) &= \mathbb{E}_i\left(\int_{t_i}^{t_{i+1}}dW_s\int_{t_i}^{t_{i+1}}\langle Z_s,\cdot\rangle_0 dW_s\right)\\ &= \mathbb{E}_i\int_{t_i}^{t_{i+1}}Z_s ds=h\overline{Z}_{t_i},\end{aligned}$$

where we used the arguments from the proof of Lemma 5.4. Subtracting $h\overline{\widehat{Z}}_{t_i}=\mathbb{E}_i(\widehat{u}_{i+1}(X^\pi_{t_{i+1}})\Delta W_i)$ and then noting that $\mathbb{E}_i(\Delta W_i\mathbb{E}_i(H_{i+1}))=0$,

$$\begin{aligned}h(\overline{Z}_{t_i}-\overline{\widehat{Z}}_{t_i}) =&\mathbb{E}_i\left[\Delta W_i(Y_{t_{i+1}}-\widehat{u}_{i+1}(X^\pi_{t_{i+1}}))\right]+\mathbb{E}_i\left(\Delta W_i\int_{t_i}^{t_{i+1}}\psi(\Theta_s)ds\right)\\ =&\mathbb{E}_i\left[\Delta W_i(H_{i+1}-\mathbb{E}_i H_{i+1})\right]+\mathbb{E}_i\left(\Delta W_i\int_{t_i}^{t_{i+1}}\psi(\Theta_s)ds\right)\end{aligned}$$

By applying the conditional version of Holder inequality for the first term and its classical form to the second one, follows that

$$\begin{aligned}h^2\mathbb{E}\left\|\overline{Z}_{t_i}-\overline{\widehat{Z}}_{t_i}\right\|_V^2 &= \mathbb{E}\left\|\mathbb{E}_i\Big[\Delta W_i(H_{i+1}-\mathbb{E}_i H_{i+1})\Big]+\mathbb{E}_i\left(\Delta W_i\int_{t_i}^{t_{i+1}}\psi(\Theta_s)ds\right)\right\|_V^2\\ &\le 2\mathbb{E}\left(\mathbb{E}_i\left\|\Delta W_i\right\|_V^2\mathbb{E}_i[H_{i+1}-\mathbb{E}H_{i+1}]^2\right)+2\mathbb{E}\left(\mathbb{E}_i\left\|\Delta W_i\right\|_V^2\mathbb{E}_i\left[\int_{t_i}^{t_{i+1}}\psi(\Theta_s)ds\right]^2\right)\\ &\le C\mathrm{tr}(Q)\mathbb{E}\left(\mathbb{E}_i H_{i+1}^2-(\mathbb{E}_i H_{i+1})^2\right)+Ch\,\mathrm{tr}(Q)\mathbb{E}\int_{t_i}^{t_{i+1}}|\psi(\Theta_s)|^2 ds;\end{aligned}\tag{5.16}$$

Putting all together,

$$\begin{aligned}\mathbb{E}\left|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}\right|^2 \le&\,(1+\gamma h)\,\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\\ &+(1+\gamma h)\frac{C}{\gamma}\Big[h^2+e_i(X,X^\pi_{t_i})+e_i(Y,Y_{t_i})+e_i(Z,\overline{Z}_{t_i})+h\mathbb{E}|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}|^2\\ &\qquad+\mathrm{tr}(Q)\mathbb{E}H_{i+1}^2-\mathrm{tr}(Q)\mathbb{E}|\mathbb{E}_i H_{i+1}|^2\\ &\qquad+h\mathrm{tr}(Q)\mathbb{E}\int_{t_i}^{t_{i+1}}|\psi(\Theta_s)|^2 ds\Big].\end{aligned}$$

Where we also used that $Z_t, \overline{Z}_{t_i}$ are V_0-valued and implies $\left\|Z_t - \overline{Z}_{t_i}\right\|_V^2 \leq \left\|Q^{1/2}\right\|_{L(Q)}^2 \left\|Z_t - \overline{Z}_{t_i}\right\|_0^2$. Let $\gamma = C^2 \mathrm{tr}(Q)$ and note that $(1+\gamma h)\frac{C}{\gamma} \leq C$ and also $\gamma \leq C$, then the above term transform to

$$\begin{aligned} &Ch^2 + Ce_i(X, X_{t_i}^\pi) + Ce_i(Y, Y_{t_i}) + Ce_i(Z, \overline{Z}_{t_i}) \\ &+ Ch\mathbb{E}|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}|^2 + C(1+Ch)\mathbb{E}H_{i+1}^2 + Ch\mathbb{E}\int_{t_i}^{t_{i+1}} |\psi(\Theta_s)|^2 ds. \end{aligned}$$

Now we take h small such that $Ch < 1$ and then

$$\begin{aligned} \mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 \leq\ & Ch^2 + Ce_i(X, X_{t_i}^\pi) + Ce_i(Y, Y_{t_i}) + Ce_i(Z, \overline{Z}_{t_i}) \\ &+ C(1+Ch)\mathbb{E}H_{i+1}^2 + Ch\mathbb{E}\int_{t_i}^{t_{i+1}} |\psi(\Theta_s)|^2 ds. \end{aligned}$$

Finally, by recalling that $H_{i+1} = Y_{t_{i+1}} - \widehat{u}_{i+1}(X_{t_{i+1}}^\pi)$, we have established (4.20). □

Step 2

The term,

$$C(1+Ch)\mathbb{E}\left|Y_{t_{i+1}} - \widehat{u}_{i+1}(X_{t_{i+1}}^\pi)\right|^2,$$

in (4.20) was left without a control in previous step. Here in what follows we provide a control on this term. The purpose of this section is to show the following estimate:

Lemma 5.11 *There exists a constant $C > 0$ such that,*

$$\begin{aligned} \max_{i\in\{0,\dots,N-1\}} \mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X_{t_i}^\pi)\right|^2 \leq C\Bigg[& h + \mathbb{E}\left|\phi(X_T) - \phi(X_T^\pi)\right|^2 N + \sum_{i=0}^{N-1} \mathbb{E}\left|\widehat{u}_i(X_{t_i}^\pi) - \widehat{\mathcal{V}}_{t_i}\right|^2 \\ &+ e(X, X^\pi) + e(Y, (Y_t)_{t\in\pi}) + e(Z, (\overline{Z}_t)_{t\in\pi})\Bigg]. \end{aligned} \tag{5.17}$$

The rest of this section is devoted to the proof of this result.

PROOF OF LEMMA 5.11. Recall $H_{i+1} = Y_{t_{i+1}} - \widehat{u}_{i+1}(X_{t_{i+1}}^\pi)$. We have that $(a+b)^2 \geq (1-h)a^2 + (1-\frac{1}{h})b^2$ and

$$\begin{aligned} \mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 &= \mathbb{E}\left|\left(Y_{t_i} - \widehat{u}_i(X_{t_i}^\pi)\right) + \left(\widehat{u}_i(X_{t_i}^\pi) - \widehat{\mathcal{V}}_{t_i}\right)\right|^2 \\ &\geq (1-h)\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X_{t_i}^\pi)\right|^2 + \left(1 - \frac{1}{h}\right)\mathbb{E}\left|\widehat{u}_i(X_{t_i}^\pi) - \widehat{\mathcal{V}}_{t_i}\right|^2. \end{aligned} \tag{5.18}$$

Therefore, we have an upper (4.20) and lower (5.18) bound for $\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2$. By connecting these bounds,

$$\begin{aligned} (1-h)\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X_{t_i}^\pi)\right|^2 + \left(1 - \frac{1}{h}\right)\mathbb{E}\left|\widehat{u}_i(X_{t_i}^\pi) - \widehat{\mathcal{V}}_{t_i}\right|^2 \leq\ & Ch^2 + Ce_i(X, X_{t_i}^\pi) + Ce_i(Y, Y_{t_i}) + Ce_i(Z, \overline{Z}_{t_i}) \\ &+ Ch\mathbb{E}\int_{t_i}^{t_{i+1}} \psi(\Theta_s)^2 ds + C(1+Ch)\mathbb{E}\left(H_{i+1}^2\right). \end{aligned}$$

Using that for sufficiently small h we have $(1-h)^{-1} \leq 2 \leq C$, we get,

$$\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 \leq CN\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2 + Ch^2 + Ce_i(X, X^\pi_{t_i}) + Ce_i(Y, Y_{t_i}) + Ce_i(Z, \overline{Z}_{t_i}) + Ch\mathbb{E}\int_{t_i}^{t_{i+1}} |\psi(\Theta_s)|^2 ds + C\mathbb{E}\left|Y_{t_{i+1}} - \widehat{u}_{i+1}(X^\pi_{t_{i+1}})\right|^2.$$

Notice that the expression on time t_i that we want to estimate, appears on the right side on time t_{i+1}, we can iterate the bound and get that $\forall\, i \in \{0, ..., N-1\}$

$$\begin{aligned}
&\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 \\
&\leq CN\sum_{k=i}^{N-1}\mathbb{E}\left|\widehat{u}_k(X^\pi_{t_k}) - \widehat{\mathcal{V}}_{t_k}\right|^2 + C(N-i)h^2 + C\sum_{k=i}^{N-1}\left[e_i(X, X^\pi_{t_i}) + e_i(Y, Y_{t_i}) + e_i(Z, \overline{Z}_{t_i})\right] \\
&\quad + Ch\sum_{k=i}^{N-1}\mathbb{E}\int_{t_k}^{t_{k+1}} |\psi(\Theta_s)|^2 ds + C\mathbb{E}\left|Y_{t_N} - \phi(X^\pi_{t_N})\right|^2 \\
&\leq CN\sum_{k=0}^{N-1}\mathbb{E}\left|\widehat{u}_k(X^\pi_{t_k}) - \widehat{\mathcal{V}}_{t_k}\right|^2 + CNh^2 + C\left[e(X, X^\pi) + e(Y, (Y_t)_{t\in\pi}) + e(Z, (\overline{Z}_t)_{t\in\pi})\right] \\
&\quad + Ch\sum_{k=0}^{N-1}\mathbb{E}\int_{t_k}^{t_{k+1}} |\psi(\Theta_s)|^2 ds + C\mathbb{E}\left|Y_{t_N} - \phi(X^\pi_{t_N})\right|^2.
\end{aligned}$$

Applying maximum on $i \in \{0, ..., N-1\}$ and recalling bound from Lemma (5.8),

$$\max_{i\in\{0,\dots,N-1\}}\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 \leq C\Bigg[h + \mathbb{E}\left|\phi(X_T) - \phi(X^\pi_T)\right|^2 N + \sum_{i=0}^{N-1}\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2 + e(X, X^\pi) + e(Y, (Y_t)_{t\in\pi}) + e(Z, (\overline{Z}_t)_{t\in\pi})\Bigg].$$

This is nothing that (5.17). □

Step 3

Estimate (4.31) contains some uncontrolled terms on its RHS. Here the purpose is to bound the term

$$\sum_{i=0}^{N-1}\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2,$$

in terms of more tractable terms. In this step we will prove

Lemma 5.12 *It holds that,*

$$\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2 + h\mathbb{E}\left\|\widehat{\overline{Z}}_{t_i} - \hat{z}_i(X^\pi_{t_i})\right\|_0^2 \leq C\varepsilon_i^v + Ch\varepsilon_i^z, \tag{5.19}$$

with ε_i^v and ε_i^z defined in (5.6).

PROOF. Fix $i \in \{0, ..., N-1\}$. Recall the martingale $(N_t)_{t\in[t_i,t_{i+1}]}$ and take $t = t_{i+1}$,

$$\widehat{u}_{i+1}(X^\pi_{t_{i+1}}) = \mathbb{E}_i \widehat{u}_{i+1}(X^\pi_{t_{i+1}}) + \int_{t_i}^{t_{i+1}} \langle \widehat{Z}_s, \cdot\rangle_0 dW_s.$$

Now we replace the definition of $\widehat{\mathcal{V}}_{t_i}$ (5.2),

$$\widehat{u}_{i+1}(X^\pi_{t_{i+1}}) = \widehat{\mathcal{V}}_{t_i} - \psi(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i})h + \int_{t_i}^{t_{i+1}} \langle \widehat{Z}_s, \cdot\rangle_0 dW_s. \tag{5.20}$$

Now fix a parameter $\theta \in \Theta_\eta$ and replace (4.34) on $L_i(\theta)$:

$$L_i(\theta) = \mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i}) + \psi(t_i, X^\pi_{t_i}, u^\theta_i(X^\pi_{t_i}), z^\theta_i(X^\pi_{t_i}))h - \psi(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i})h + \int_{t_i}^{t_{i+1}} \langle \widehat{Z}_s - z^\theta_i(X^\pi_{t_i}), \cdot\rangle_0 dW_s\right|^2$$

Note that the four first terms are $\mathcal{F}_{t_i}$-measurable and the stochastic integral is a martingale difference, therefore

$$\begin{aligned} L_i(\theta) = \mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i}) + \psi(t_i, X^\pi_{t_i}, u^\theta_i(X^\pi_{t_i}), z^\theta_i(X^\pi_{t_i}))h - \psi(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i})h\right|^2 \\ + \mathbb{E}\int_{t_i}^{t_{i+1}} \left\|\widehat{Z}_s - \overline{\widehat{Z}}_{t_i}\right\|_0^2 ds + h\mathbb{E}\left\|\overline{\widehat{Z}}_{t_i} - z^\theta_i(X^\pi_{t_i})\right\|_0^2. \end{aligned}$$

Where we used Ito isometry and the same argument used on equation (4.26). With this decomposition of $L_i(\theta)$, we can easily see the part that depends on θ. Lets work with $\hat{L}_i$ defined as follows,

$$\hat{L}_i(\theta) = \mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i}) + \left(\psi(t_i, X^\pi_{t_i}, u^\theta_i(X^\pi_{t_i}), z^\theta_i(X^\pi_{t_i})) - \psi(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i})\right)h\right|^2 + h\mathbb{E}\left\|\overline{\widehat{Z}}_{t_i} - z^\theta_i(X^\pi_{t_i})\right\|_0^2.$$

Let $\gamma > 0$ and use Young inequality and the Lipschitz condition on ψ to find that

$$\begin{aligned} &\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i}) + \left(\psi(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}) - \psi(t_i, X^\pi_{t_i}, u^\theta_i(X^\pi_{t_i}), z^\theta_i(X^\pi_{t_i}))\right)\right|^2 \\ &\leq (1+\gamma h)\,\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i})\right|^2 + \left(1 + \frac{1}{\gamma h}\right) h^2 C\mathbb{E}\left(|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i})|^2 + \left\|z^\theta_i(X^\pi_{t_i}) - \overline{\widehat{Z}}_{t_i}\right\|_0^2\right) \\ &\leq C\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i})\right|^2 + Ch\mathbb{E}\left\|z^\theta_i(X^\pi_{t_i}) - \overline{\widehat{Z}}_{t_i}\right\|_0^2. \end{aligned}$$

Therefore, we have an upper bound on $L(\theta)$ for all $\theta \in \Theta_\eta$, to find a lower bound, we use $(a+b)^2 \geq (1-\gamma h)a^2 + \left(1 - \frac{1}{\gamma h}\right)b^2 \geq (1-\gamma h)a^2 - \frac{1}{\gamma h}b^2$ with $\gamma > 0$

$$\begin{aligned} \mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i}) + \left(\psi(t_i, X^\pi_{t_i}, \widehat{\mathcal{V}}_{t_i}, \overline{\widehat{Z}}_{t_i}) - \psi(t_i, X^\pi_{t_i}, u^\theta_i(X^\pi_{t_i}), z^\theta_i(X^\pi_{t_i}))\right)\right|^2 &\geq (1-Ch)\,\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i})\right|^2 \\ &\quad - \frac{h}{2}\mathbb{E}\left\|z^\theta_i(X^\pi_{t_i}) - \overline{\widehat{Z}}_{t_i}\right\|_0^2; \end{aligned}$$

where we used $\gamma = 2C$ in order to force the $\frac{1}{2}$ in the second term of the RHS. Then, connecting these bounds and using that $\forall\theta \in \Theta\ \hat{L}(\theta^*) \leq \hat{L}(\theta)$ yields,

$$(1-Ch)\,\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - \hat{u}_i(X^\pi_{t_i})\right|^2 + \frac{h}{2}\mathbb{E}\left\|\overline{\widehat{Z}}_{t_i} - \hat{z}_i(X^\pi_{t_i})\right\|_0^2 \leq C\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - u^\theta_i(X^\pi_{t_i})\right|^2 Ch\mathbb{E}\left\|\overline{\widehat{Z}}_{t_i} - z^\theta_i(X^\pi_{t_i})\right\|_0^2.$$

By taking h small such that $(1-Ch) \geq \frac{1}{2}$ and infimum on the right side with respect to $\theta \in \Theta_\eta$ we get (5.19),

$$\mathbb{E}\left|\widehat{\mathcal{V}}_{t_i} - \hat{u}_i(X^\pi_{t_i})\right|^2 + h\mathbb{E}\left\|\overline{\widehat{Z}}_{t_i} - \hat{z}_i(X^\pi_{t_i})\right\|_0^2 \leq C\varepsilon_i^{v,\eta} + Ch\varepsilon_i^{z,\eta} \tag{5.21}$$

Thus the proof is completed. □

Previous lemma and steps proves the following.

Lemma 5.13 *It holds that,*

$$\max_{i\in\{0,\dots,N-1\}} \mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 + \leq C\Bigg[h + \mathbb{E}\left|\phi(X_T) - \phi(X^\pi_T)\right|^2 + N\varepsilon^{v,\eta} + \varepsilon^{z,\eta} + e(X, X^\pi) + e(Y, (Y_t)_{t\in\pi}) + e(Z, (\overline{Z}_t)_{t\in\pi})\Bigg]. \tag{5.22}$$

Step 4

In this step we show the desire bound for the remaining component.

Lemma 5.14 *It holds that,*

$$\sum_{i=0}^{N-1}\mathbb{E}\int_{t_i}^{t_{i+1}}\left\|Z_s - \hat{z}_i(X^\pi_{t_i})\right\|_0^2 ds \leq C\Bigg[h + \mathbb{E}\left|\phi(X_T) - \phi(X^\pi_T)\right|^2 + N\varepsilon^{v,\eta} + \varepsilon^{z,\eta} + e(X, X^\pi) + e(Y, (Y_t)_{t\in\pi}) + e(Z, (\overline{Z}_t)_{t\in\pi})\Bigg]. \tag{5.23}$$

PROOF. We will use triangular inequality passing through $\overline{\widehat{Z}}_{t_i}$. Note that the term containing $\left\|\overline{\widehat{Z}}_{t_i} - \hat{z}_i(X^\pi_{t_i})\right\|_0^2$ is well-controlled by Lemma 5.12. By using (4.29) with Lemma 5.8 on (4.26), we get

$$\mathbb{E}\int_{t_i}^{t_{i+1}}\left\|Z_s - \overline{\widehat{Z}}_{t_i}\right\|_0^2 ds \leq C\mathbb{E}\int_{t_i}^{t_{i+1}}\left\|Z_s - \overline{Z}_{t_i}\right\|_0^2 ds + C\mathbb{E}\left(\mathbb{E}_i H_{i+1}^2 - (\mathbb{E}_i H_{i+1})^2\right) + Ch\mathbb{E}\int_{t_i}^{t_{i+1}}|\psi(\Theta_s)|^2 ds.$$

which implies, after summing over $i \in \{0, \dots N-1\}$,

$$\mathbb{E}\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\left\|Z_t - \overline{\widehat{Z}}_{t_i}\right\|_0^2 ds \leq C\sum_{i=0}^{N-1}\left(\mathbb{E}\left(H_{i+1}^2\right) - \mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right) + Ch + e(Z, (\overline{Z}_t)_{t\in\pi}). \tag{5.24}$$

The next step is to give a suitable bound for $\mathbb{E}\left(H_{i+1}^2\right) - \mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2$. Recall from (5.13) that $H_{i+1} = Y_{t_{i+1}} - \hat{u}_{i+1}(X^\pi_{t_{i+1}})$, then

$$\begin{aligned}
\sum_{i=0}^{N-1}\left(\mathbb{E}\left(H_{i+1}^2\right) - \mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right) &= \sum_{i=0}^{N-1}\mathbb{E}(H_{i+1}^2) - \sum_{i=0}^{N-1}\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \\
&= \mathbb{E}\left|Y_{t_N} - \hat{u}_N(X^\pi_{t_N})\right| + \sum_{i=0}^{N-2}\mathbb{E}(H_{i+1}^2) - \sum_{i=0}^{N-1}\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \\
&\leq \mathbb{E}\left|\phi(X_T) - \phi(X_T^\pi)\right|^2 + \mathbb{E}(H_0^2) + \sum_{i=1}^{N-1}\mathbb{E}(H_i^2) - \sum_{i=0}^{N-1}\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \\
&= \mathbb{E}\left|\phi(X_T) - \phi(X_T^\pi)\right|^2 + \sum_{i=0}^{N-1}\left(\mathbb{E}(H_i^2) - \mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right).
\end{aligned} \tag{5.25}$$

From (5.18) and (5.14) we have an lower and upper bound for $\mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2$. Indeed, first one has

$$(1-h)\mathbb{E}\left|Y_{t_i} - \widehat{u}_i(X^\pi_{t_i})\right|^2 \leq \mathbb{E}\left|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}\right|^2 + \left(\frac{1}{h} - 1\right)\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2. \tag{5.26}$$

Then, we have that for all $\gamma > 0$

$$\begin{aligned}
&(1-h)\ \mathbb{E}\left|Y_{t_i} - \hat{u}_i(X^\pi_{t_i})\right|^2 \\
&\leq \left(\frac{1}{h} - 1\right)\mathbb{E}\left|\hat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2 + (1+\gamma h)\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \\
&\quad + (1+\gamma h)\frac{C}{\gamma}\Bigg[\underbrace{h^2 + e_i(X, X^\pi_{t_i}) + e_i(Y, Y_{t_i}) + h\mathbb{E}|Y_{t_i} - \widehat{\mathcal{V}}_{t_i}|^2 + \mathbb{E}\int_{t_i}^{t_{i+1}}\left\|Z_s - \overline{\widehat{Z}}_{t_i}\right\|_0^2 ds}_{B_i}\Bigg].
\end{aligned}$$

Let us call the expression inside the squared brackets by B_i. Subtracting $(1-h)\mathbb{E}\left|\mathbb{E}_i H_{i+1}\right|^2$ and dividing by $(1-h)$,

$$\mathbb{E}(H_i^2) - \mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \leq \frac{1}{h}\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2 + \left(\frac{h+\gamma h}{1-h}\right)\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 + \frac{C}{\gamma}\frac{(1+\gamma h)}{(1-h)}B_i.$$

For $\gamma = 3C$ and sufficiently small h, we can force,

$$\frac{C}{\gamma}\frac{(1+\gamma h)}{(1-h)} \leq \frac{1}{2} \qquad \text{and} \qquad \frac{1}{1-h} \leq \frac{1}{2}.$$

Hence,

$$\mathbb{E}(H_i^2) - \mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \leq \frac{1}{h}\mathbb{E}\left|\widehat{u}_i(X^\pi_{t_i}) - \widehat{\mathcal{V}}_{t_i}\right|^2 + Ch\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 + \frac{1}{2}B_i.$$

Finally, note that,

$$\sum_{i=0}^{N-1}\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2 \leq \mathbb{E}\left|\phi(X_T) - \phi(X_T^\pi)\right|^2 + N\max_{i=0,\ldots,N-1}\mathbb{E}\left|Y_{t_i} - \hat{u}_i(X^\pi_{t_i})\right|^2. \tag{5.27}$$

Coming back to (4.37),

$$\sum_{i=0}^{N-1}\left(\mathbb{E}\left(H_{i+1}^2\right)-\mathbb{E}\left|\mathbb{E}_i(H_{i+1})\right|^2\right)\leq C\mathbb{E}\left|\phi(X_T)-\phi(X_T^\pi)\right|^2+N\sum_{i=0}^{N-1}\mathbb{E}\left|\widehat{u}_i(X_{t_i}^\pi)-\widehat{\mathcal{V}}_{t_i}\right|^2$$
$$+ChN\max_{i=0,\dots,N-1}\mathbb{E}\left|Y_{t_i}-\hat{u}_i(X_{t_i}^\pi)\right|^2+\frac{1}{2}\sum_{i=0}^{N-1}B_i.$$

Therefore, by plugging this bound in (4.36), noting that $|Y_{t_i}-\widehat{\mathcal{V}}_{t_i}|^2\leq 2|Y_{t_i}-\hat{u}_i(X_{t_i}^\pi)|^2+2|\hat{u}_i(X_{t_i}^\pi)-\widehat{\mathcal{V}}_{t_i}|^2$ and $hN=1$, we have,

$$\mathbb{E}\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\left\|Z_s-\overline{\widehat{Z}}_{t_i}\right\|_0^2ds$$
$$\leq C\Bigg[h+\mathbb{E}\left|\phi(X_T)-\phi(X_T^\pi)\right|^2+N\sum_{i=0}^{N-1}\mathbb{E}\left|\hat{u}_{t_i}(X_{t_i}^\pi)-\widehat{\mathcal{V}}_{t_i}\right|^2$$
$$+\max_{i=0,\dots,N-1}\mathbb{E}\left|Y_{t_i}-\hat{u}_i(X_{t_i}^\pi)\right|^2+e(X,X^\pi)+e(Y,(Y_t)_{t\in\pi})+e(Z,(\overline{Z}_t)_{t\in\pi})\Bigg].$$

Now, use Lemma 5.12 and Lemma 5.13 to get

$$\mathbb{E}\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\left\|Z_s-\overline{\widehat{Z}}_{t_i}\right\|_0^2ds\leq C\Bigg[h+\mathbb{E}\left|\phi(X_T)-\phi(X_T^\pi)\right|^2+N\varepsilon^{v,\eta}+\varepsilon^{z,\eta}$$
$$+e(X,X^\pi)+e(Y,(Y_t)_{t\in\pi})+e(Z,(\overline{Z}_t)_{t\in\pi})\Bigg].$$

Thus, it has been demonstrated. □

By combining Lemma 5.13 with Lemma 5.14 and using Assumptions 5.7, the proof of Theorem 5.9 is now complete. □

We finish this work with the following closing remark.

Remark 5.15 *Note that if the approximators are DeepOnets, then* $\varepsilon^{v,\eta},\varepsilon^{z,\eta}\to 0$ *as* $\eta\to\infty$. *See Remark 3.20.*

Bibliography

[AGM+16] S. Albeverio, L. Gawarecki, V. Mandrekar, B. Rüdiger, and B. Sarkar. Ito formula for mild solutions of spdes with gaussian and non-gaussian noise and applications to stability properties, 2016.

[AKF+17] Mahabal A, Sheth K, Gieseke F, Pai A, Djorgovski S, G, J Drake A, and Graham M. J. Deep-learnt classification of light curves. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, nov 2017.

[All07] Grégoire Allaire. *Numerical Analysis and Optimization An introduction to mathematical modeling and numerical simulation*. Oxford University Press, 2007.

[Alq19] Ali Mohammad Alqudah. Brain tumor classification using deep learning technique - a comparison between cropped, uncropped, and segmented lesion images with different sizes. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(6):3684–3691, dec 2019.

[App09] David Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2009.

[AR05] Sergio Albeverio and B. Rüdiger. Stochastic integrals and the lévy-ito decomposition theorem on separable banach spaces. *Stochastic Analysis and Applications - STOCHASTIC ANAL APPL*, 23:217–253, 01 2005.

[ATY+19] Md. Zahangir Alom, Tarek Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Nasrin, Mahmudul Hasan, Brian Essen, Abdul Awwal, and Vijayan Asari. A state-of-the-art survey on deep learning theory and architectures. *Electronics*, 8:292, 03 2019.

[Bar19] Dalya Baron. Machine learning in astronomy: a practical overview, 2019.

[BBG+21] Christian Beck, Sebastian Becker, Philipp Grohs, Nor Jaafari, and Arnulf Jentzen. Solving the kolmogorov PDE by means of deep learning. *Journal of Scientific Computing*, 88(3), jul 2021.

[BBJ+22] Victor Boussange, Sebastian Becker, Arnulf Jentzen, Benno Kuckuck, and Loïc Pellissier. Deep learning approximations for non-local nonlinear pdes with neumann boundary conditions, 2022.

[BBP97] Guy Barles, Rainer Buckdahn, and Etienne Pardoux. Backward stochastic differential equations and integral-partial differential equations. *Stochastics and Stochastic Reports*, 60(1-2):57–83, 1997.

[BE08] Bruno Bouchard and Romuald Elie. Discrete time approximation of decoupled forward-backward sde with jumps. *Stochastic Processes and their Applications*, 118(1):53–75, 2008.

[BEJ19] Christian Beck, Weinan E, and Arnulf Jentzen. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *Journal of Nonlinear Science*, 29(4):1563–1619, jan 2019.

[BGT20] Isabeau Birindelli, Giulio Galise, and Erwin Topp. Fractional truncated laplacians: representation formula, fundamental solutions and applications, 2020.

[BHH+20] Christian Beck, Fabian Hornung, Martin Hutzenthaler, Arnulf Jentzen, and Thomas Kruse. Overcoming the curse of dimensionality in the numerical approximation of allen–cahn partial differential equations via truncated full-history recursive multilevel picard approximations. *Journal of Numerical Mathematics*, 28(4):197–222, dec 2020.

[BHJ21] Christian Beck, Martin Hutzenthaler, and Arnulf Jentzen. On nonlinear feynman–kac formulas for viscosity solutions of semilinear parabolic partial differential equations. *Stochastics and Dynamics*, 21(08), jul 2021.

[BJMvW22] Sebastian Becker, Arnulf Jentzen, Marvin S. Müller, and Philippe von Wurstemberger. Learning the random variables in monte carlo simulations with stochastic gradient descent: Machine learning for parametric pdes and financial derivative pricing, 2022.

[BLT17] Guy Barles, Olivier Ley, and Erwin Topp. Lipschitz regularity for integro-differential equations with coercive hamiltonians and application to large time behavior. *Nonlinearity*, 30(2):703–734, jan 2017.

[Bog07] V.I. Bogachev. *Measure theory Vol. II*. Cambridge University Press, 2007.

[Bou19] Dimitri Bourilkov. Machine and deep learning applications in particle physics. *International Journal of Modern Physics A*, 34(35):1930019, dec 2019.

[BR11] Evelyn Buckwar and Martin G. Riedler. Runge–kutta methods for jump–diffusion differential equations. *Journal of Computational and Applied Mathematics*, 236(6):1155–1182, 2011.

[Bre11] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer New York, NY, 2011.

[BS73] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[BT04] Bruno Bouchard and Nizar Touzi. Discrete-time approximation and monte-carlo simulation of backward stochastic differential equations. *Stochastic Processes and their Applications*, 111(2):175–206, 2004.

[BWM04] David Benson, Stephen Wheatcraft, and Mark Meerschaert. Application of a fractional advection-dispersion equation. *Water Resources Research*, 36, 09 2004.

[Cas21] Javier Castro. Deep learning schemes for parabolic nonlocal integro-differential equations, 2021.

[Cas22] Javier Castro. The kolmogorov infinite dimensional equation in a hilbert space via deep learning methods, 2022.

[CC95] Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its applications to dynamic systems. *Neural Networks, IEEE Transactions on*, pages 911 – 917, 08 1995.

[CGMY02] Peter Carr, Hélyette Geman, Dilip B. Madan, and Marc Yor. The fine structure of asset returns: An empirical investigation. *The Journal of Business*, 75(2):305–332, 2002.

[CJKP18] Sonja Cox, Arnulf Jentzen, Ryan Kurniawan, and Primož Pušnik. On the mild itô formula in banach spaces. *Discrete &amp: Continuous Dynamical Systems - B*, 23(6):2217–2243, 2018.

[CJL19] Sonja Cox, Arnulf Jentzen, and Felix Lindner. Weak convergence rates for temporal numerical approximations of stochastic wave equations with multiplicative noise, 2019.

[Cra16] John Crank. *The Mathematics of Diffusion*. Oxford: Clarendon Press, 2016.

[CS07] Luis Caffarelli and Luis Silvestre. An extension problem related to the fractional laplacian. *Communications in Partial Differential Equations*, 32(8):1245–1260, aug 2007.

[CT04] Rama Cont and Peter Tankov. *Financial modelling with jump processes*. Chapman and Hall/CRC, August 2004. 2003.

[Dal66] Yu. Daleckij. Differential equations with functional derivatives and stochastic equations for generalized random processes. *Dokl. Akad. Nauk SSSR*, 166(5):1035–1038, 1966.

[DDG+20] Marta D'Elia, Qiang Du, Christian Glusa, Max Gunzburger, Xiaochuan Tian, and Zhi Zhou. Numerical methods for nonlocal and fractional models. *Acta Numerica*, 29:1–124, may 2020.

[Del13] Lukasz Delong. *Backward Stochastic Differential Equations with Jumps and Their Actuarial and Financial Applications*. Springer London, 2013.

[DPV12] Eleonora Di Nezza, Giampiero Palatucci, and Enrico Valdinoci. Hitchhiker's guide to the fractional sobolev spaces. *Bulletin des Sciences Mathématiques*, 136(5):521–573, 2012.

[DPZ92] Guiseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1992.

[DT17] Qiang Du and Xiaochuan Tian. Stability of nonlocal dirichlet integrals and implications for peridynamic correspondence material modeling, 2017.

[DT20] Gonzalo Dávila and Erwin Topp. The nonlocal inverse problem of donsker and varadhan, 2020.

[EGJS21] Dennis Elbrächter, Philipp Grohs, Arnulf Jentzen, and Christoph Schwab. DNN expression rate analysis of high-dimensional PDEs: Application to option pricing. *Constructive Approximation*, 55(1):3–71, may 2021.

[EHJ17] 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, nov 2017.

[EJRW22] Simon Eberle, Arnulf Jentzen, Adrian Riekert, and Georg Weiss. Normalized gradient flow optimization in the training of relu artificial neural networks, 2022.

[Eva10] Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010.

[FT02] Marco Fuhrman and Gianmario Tessitore. Nonlinear kolmogorov equations in infinite dimensional spaces: The backward stochastic differential equations approach and applications to optimal control. *The Annals of Probability*, 30(3):1397–1465, 2002.

[GDN09] Frank Proske Giulia Di Nunno, Bernt Øksendal. *Malliavin Calculus for Lévy Processes with Applications to Finance*. Springer Berlin, Heidelberg, 2009.

[GHJvW18] Philipp Grohs, Fabian Hornung, Arnulf Jentzen, and Philippe von Wurstemberger. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of black-scholes partial differential equations, 2018.

[GO08] Guy Gilboa and Stanley Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7:1005–1028, 01 2008.

[Gro67] Leonard Gross. Potential theory on hilbert space. *Journal of Functional Analysis*, 1(2):123–181, 1967.

[GS21a] Lukas Gonon and Christoph Schwab. Deep relu network expression rates for option prices in high-dimensional, exponential lévy models, 2021.

[GS21b] Lukas Gonon and Christoph Schwab. Deep relu neural networks overcome the curse of dimensionality for partial integrodifferential equations, 2021.

[HE16] Jiequn Han and Weinan E. Deep learning approximation for stochastic control problems, 2016.

[HJE18] Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

[HJK+20] Martin Hutzenthaler, Arnulf Jentzen, Thomas Kruse, Tuan Anh Nguyen, and Philippe von Wurstemberger. Overcoming the curse of dimensionality in the numerical approximation of semilinear parabolic partial differential equations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2244):20190630, dec 2020.

[HJKN20] Martin Hutzenthaler, Arnulf Jentzen, Thomas Kruse, and Tuan Anh Nguyen. A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *SN Partial Differential Equations and Applications*, 1(2), apr 2020.

[Hor91] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[HPBL21] Côme Huré, Huyên Pham, Achref Bachouch, and Nicolas Langrené. Deep neural networks algorithms for stochastic control problems on finite horizon: Convergence analysis. *SIAM Journal on Numerical Analysis*, 59(1):525–557, jan 2021.

[HPW19] Côme Huré, Huyên Pham, and Xavier Warin. Deep backward schemes for high-dimensional nonlinear pdes, 2019.

[KHT10] Arturo Kohatsu-Higa and Peter Tankov. Jump-adapted discretization schemes for lévy-driven sdes. *Stochastic Processes and their Applications*, 120(11):2258–2285, 2010.

[Kol38] A. N. Kolmogorov. On the analytic methods of probability theory. *Uspekhi Mat. Nauk*, 30(5):5–41, 1938.

[LJP+21] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, mar 2021.

[LKB+17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[LLPS93] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

[LMK21] Samuel Lanthaler, Siddhartha Mishra, and George Em Karniadakis. Error estimates for deeponets: A deep learning framework in infinite dimensions, 2021.

[LMT14] Antoine Lejay, Ernesto Mordecki, and Soledad Torres. Numerical approximation of backward stochastic differential equations with jumps, 09 2014.

[MBS17] Kevin Matzen, Kavita Bala, and Noah Snavely. Streetstyle: Exploring world-wide clothing styles from millions of photos, 2017.

[MNdH20] Martin Magill, Andrew M. Nagel, and Hendrick W. de Haan. Neural network solutions to differential equations in nonconvex domains: Solving the electric field in the slit-well microfluidic device. *Physical Review Research*, 2(3), jul 2020.

[MP43] W.S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[Mun00] James R. Munkres. *Topology*. Prentice Hall, Inc., Upper Saddle River, NJ, 2nd ed. edition, 2000.

[Nor97] J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[PP90] Etienne Pardoux and S.G. Peng. Adapted solution of a backward stochastic differential equation. *Systems & Control Letters*, 14:55–61, 01 1990.

[Pro04] Philip E. Protter. *Stochastic Integration and Differential Equations*. Springer Berlin, Heidelberg, 2 edition, 2004.

[PWG19] Huyen Pham, Xavier Warin, and Maximilien Germain. Neural networks-based backward scheme for fully nonlinear pdes, 2019.

[Ros58] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[Rü04] B. Rüdiger. Stochastic integration with respect to compensated poisson random measures on separable banach spaces. *Stochastics and Stochastic Reports*, 76(3):213–242, 2004.

[San91] I.W. Sandberg. Approximation theorems for discrete-time systems. *IEEE Transactions on Circuits and Systems*, 38(5):564–566, 1991.

[SS18] Justin Sirignano and Konstantinos Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, dec 2018.

[Sti18] P. R. Stinga. User's guide to the fractional laplacian and the method of semigroups, 2018.

[TMC+18] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Phys*, 14:447–450, 2018.

[TT14] Eitan Tadmor and Changhui Tan. Critical thresholds in flocking hydrodynamics with nonlocal alignment. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 372, 03 2014.

[VM13] Barbara Rüdigery Vidyadhar Mandrekar. *Stochastic Integration in Banach Spaces*. Probability Theory and Stochastic Modelling. Springer Cham, 2013.

[WL15] Michael Röckner Wei Liu. *Stochastic Partial Differential Equations: An Introduction*. Springer CHam, 2015.

[WR17] Haohan Wang and Bhiksha Raj. On the origin of deep learning, 2017.

[Zha04] Jianfeng Zhang. A numerical scheme for BSDEs. *The Annals of Applied Probability*, 14(1):459 – 488, 2004.