Deep learning schemes for parabolic nonlocal integro-differential equations

Javier Castro

Received: 16 April 2021 / Accepted: 2 October 2022
© The Author(s) 2022

Abstract
In this paper we consider the numerical approximation of nonlocal integro differential parabolic equations via neural networks. These equations appear in many recent applications, including finance, biology and others, and have been recently studied in great generality starting from the work of Caffarelli and Silvestre by Liu and Liu (Comm PDE 32(8):1245–1260, 2007). Based on the work by Hure, Pham and Warin by Hure et al. (Math Comp 89:1547–1579, 2020), we generalize their Euler scheme and consistency result for Backward Forward Stochastic Differential Equations to the nonlocal case. We rely on Lévy processes and a new neural network approximation of the nonlocal part to overcome the lack of a suitable good approximation of the nonlocal part of the solution.

Keywords Deep learning · Deep neural networks · Approximation · Nonlocal diffusion equations · Lévy processes · Stochastic differential equations

Mathematics Subject Classification 60H35 · 65M12 · 65C20

Contents
1 Introduction ................................................
1.1 Notation ................................................
1.2 Setting ................................................
1.3 Forward backward formulation of (1.2) ................................
1.4 Organization of this paper .......................................
2 Neural networks and approximation theorems ..............................
3 Discretization of the dynamics and the deep learning algorithm ....................
4 Preliminaries ................................................
4.1 Existence and uniqueness for the FBSDEJ ..............................

This article is part of the section “Computational Approaches” edited by Siddhartha Mishra.

J.C.’s work was funded in part by Chilean research grants FONDECYT 1191412, and CMM Conicyt PIA AFB170001. This article is part of the section “Computational Approaches” edited by Siddhartha Mishra.

✉ Javier Castro
jcastro@dim.uchile.cl

1 Departamento de Ingeniería Matemática, Universidad de Chile, Casilla, 170 Correo 3, Santiago, Chile

Published online: 24 October 2022
The goal of this article is to propose and analyse a deep learning scheme for PDEs of integral type which we refer to as PIDE models. The integral part of the considered equation is defined by a finite Lévy measure \( \lambda \) on \( \mathbb{R}^d \) (see Sect. 1.2).

A difficult problem in Applied Mathematics is to approximate solutions of Partial Differential Equations (PDEs) in large dimensions. In low dimensions such as 1, 2 or 3, classical methods such as finite differences or finite elements are commonly applied with satisfactory convergence orders (see e.g. Allaire [1, Chapters 2 and 6]). An important problem appears when we deal with high dimensional problems such as portfolio management, where each dimension represents the size of some financial derivative in the portfolio. More complications appear when the PDE is nonlocal, as presented in many applications. For finite difference methods, one needs to construct a mesh that, computationally speaking, has exponential cost on the dimension \( d \in \mathbb{N} \) of the considered PDE. This problem is known in the literature as the curse of dimensionality, and the most common attempt to solve this issue is via stochastic methods. Deep Learning (DL) methods have proven to be an efficient tool to handle this problem and to approximate solutions of high dimensional second order fully nonlinear PDEs. This is achieved by finding that the solution of the PDE, evaluated at a certain diffusion process, solves a Stochastic Differential Equation (SDE); then an Euler scheme together with DL is applied to solve the SDE, see [9, 31] for key developments.

Without being exhaustive, we present some of the current developments in this direction. First of all, Monte Carlo algorithms are an important approach to the resolution of this dimensional 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. The key developments in this area can be found in Han-Jentzen-E [29] and Beck-E-Jentzen [9]. On the other hand, Multilevel Picard method (MLP) is another approach and consist 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 [8, 32] for fundamental advances in this direction. On the other hand, the so-called Deep Galerkin method (DGM) is another DL approach used to solve quasilinear parabolic PDEs of the form \( L(u) = 0 \) with initial and boundary conditions. The cost function in this framework is defined in an intuitive way, it consists of the differences between the approximate solution \( \hat{u} \) evaluated at the initial time and spatial boundary, with the true initial and boundary conditions plus \( L(\hat{u}) \). These quantities are captured by an \( L^2 \)-type norm, which in high dimensions is minimized using Stochastic Gradient Descent (SGD) method. See [44] for the development of the DGM and [38] for an application.

In [31], the principal source of inspiration of this article, Hure, Pham, and Warin consider the framework introduced previously in [9] and present new approximation schemes for the solution of a parabolic nonlinear PDE and its gradient via Neural Networks. Via an intricate
use of intermediate numerical approximations for each term in their scheme, they prove the numerical consistency and high accuracy of the method, at least in the case of low dimensions.

In general, standard PDEs model situations where, in order to know the state of a system at a particular point, one needs information of the state in a arbitrarily small neighborhood of the point. On the contrary, PIDEs can model more general phenomena where long distance interactions and effects are not negligible, therefore, must be considered. An important example of PIDEs are those which involve fractional derivatives, such as the Fractional Laplacian. This operator has been extensively studied, from the PDE point of view, during the past ten years, starting from the fundamental work by Caffarelli and Silvestre [15]. See [21, 45] and references therein for nice introductions to this operator, one of the most relevant examples of integro-differential operators. More generally speaking, nonlocal equations are used in a wide range of scientific areas, see [10] for applications in advection dispersion equations, [25] for image processing, [23] for perodyinamic, [46] for hydrodynamics, and see [16, 17] for finances. For more theoretical results on nonlocal equations, see e.g. [6, 11, 18] and references therein. In [19], the authors give a complete introduction to nonlocal equations and then they develop nonlocal version of three numerical methods: finite difference, finite element and Spectral-Galerkin.

In [35], the authors present a discrete-time approximation of a BSDEJ (Backward SDE with Jumps) such that its solution converges weakly to a solution of the continuous time equation. They also use this method to approximate the solution to the correspondent PIDE. Very recently, we have learned about a rigorous and complete work by L. Gonon and C. Schwab [26, 27] where a proof that deep ReLu neural networks (NNs) are able to approximate expectation of certain type of functionals defined in a space of stochastic processes is given. In particular, viscosity solutions of a linear PIDE can be represented in such way by means of a variation of Feynman–Kac formula. Furthermore, relying on controlling the size of NNs that approximate the parameters for a certain fixed tolerance, they mathematically prove that neural networks, as considered in their setting, are capable of breaking the curse of dimensionality in such approximation problems. They follow a similar procedure and generalize the results shown on [24, 28]. Our result is in some sense different, we work with a nonlinear equation and do not ask for NNs approximation of the parameters but we do not show that our scheme can overcome the curse of dimensionality. As we mention before, we propose and provide error bounds for a deep learning scheme for nonlinear parabolic PIDEs based in the work of [31]. We emphasize that their methods differ from ours, and were made in an independent fashion.

We present here an extension and generalization of [31] to PIDEs, by adding nonlocal contribution to the PDE. Some important changes are needed in the algorithm, including the use of a third Neural Network to approximate nonlocal parts of the solution. Of particular utility will be the result shown in [12] to prove convergence of the proposed numerical scheme.

The basic idea of the Euler scheme presented in this article is based on that presented by Zhang in [49]. In that paper, the author gives a discrete time approximation of a BSDE (backward SDE) with no jump terms. That scheme involves the computation of conditional expectations and gives important bounds and results that were used in [31] to prove the convergence of a DL algorithm to solve a second order fully nonlinear PDE. In our case, nonlocal integral models require additional treatments. The work by Bouchard and Elie [12], very important for the work presented here, generalizes the properties given in [49] to the nonlocal setting by considering Lévy process. We will closely follow their approach to construct our numerical scheme.
1.1 Notation

For any $m \in \mathbb{N}$, $\mathbb{R}^m$ represents the finite dimensional Euclidean space with elements $x = (x_1, \ldots, x_m)$ and endowed with the usual norm $|x|^2 = \sum_{i=1}^m |x_i|^2$. 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 $x \cdot y = \sum_{i=1}^m x_i y_i$. For a general measure space $(E, \Sigma, \nu)$, $p \geq 1$ and $m \in \mathbb{N}$, $L^p(E, \Sigma, \nu; \mathbb{R}^m)$ represents the standard Lebesgue space of $p$-integrable functions from $E$ to $\mathbb{R}^m$ and norm

$$
\|f\|_{L^p(E, \Sigma, \nu; \mathbb{R}^m)} = \int_E |f(x)|^p \nu(dx).
$$

We write $L^p(E, \Sigma, \nu)$ when $m = 1$. Given a general probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random vector (or variable if $m = 1$) $X : \Omega \to \mathbb{R}^m$, for sake of simplicity and to avoid an overload of parenthesis we denote $\mathbb{E}[|X|^2] = \mathbb{E}(|X|^2)$. 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 $f = (f_1, \ldots, f_m)$. Along the paper we use several times that for $x_1, \ldots, x_k \in \mathbb{R}$ the following bound holds,

$$(x_1 + \cdots + x_k)^2 \leq k(x_1^2 + \cdots + x_k^2). \quad (1.1)$$

1.2 Setting

Let $d \geq 1$ and $T > 0$. Consider the following integro-differential PDE

$$
\begin{cases}
\mathcal{L}u(t, x) + f(t, x, u(t, x), \sigma(x)\nabla u(t, x), T[u](t, x)) = 0, & (t, x) \in [0, T] \times \mathbb{R}^d, \\
u(T, x) = g(x), & x \in \mathbb{R}^d.
\end{cases}
\quad (1.2)
$$

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 $u \in C^{1,2}([0, T] \times \mathbb{R}^d)$, as follows:

$$
\mathcal{L}u(t, x) = \partial_t u(t, x) + \nabla u(t, x) \cdot b(x) + \frac{1}{2} \text{Trace}(\sigma(x)\sigma(x)^T D^2 u(t, x))
\begin{aligned}
&+ \int_{\mathbb{R}^d} [u(t, x + \beta(x, y)) - u(t, x) - \nabla u(t, x) \cdot \beta(x, y)]\lambda(dy),
\end{aligned}
\quad (1.3)
$$

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

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

Also, $f : [0, T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$. We also assume the standard Lipschitz conditions on the functions in order to have a unique solution to (1.2) in the class $C^{1,2}$: there exists a
universal constant $K > 0$ such that

$$(C)$$

- (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 $\sup_{y \in \mathbb{R}^d} |\beta(0, y)| \leq K$.
- (Uniformly Lipschitz) $\sup_{y \in \mathbb{R}^d} |\beta(x, y) - \beta(x', y)| \leq K|x - x'|$, $\forall x, x' \in \mathbb{R}^d$.
- (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 $|f(t, x, y, z) - f(t', x', y', z')| \leq K (|t - t'|^{1/2} + |x - x'| + |y - y'| + |z - z'| + |w - w'|)$.
- (Invertibility) For each $y \in \mathbb{R}^d$, the map $\beta(x, y)$ 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 $x, \xi \in \mathbb{R}^d$, $a(x, \xi; y) \geq |\xi|^2 K^{-1}$ or $a(x, \xi; y) \leq -|\xi|^2 K^{-1}$.

The last condition is of technical type and it is needed to ensure the validity of certain approximation results (see Theorem 4.1). On the other hand, the nonlocal, integro-differential operator $I$ is defined as

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

The conditions stated in (1.4) are standard in the literature (see [5, 12, 35]) and are needed to ensure the existence and uniqueness (with satisfactory bounds mentioned below) of solutions to a FBSDEJ (forward BSDEJ) related to (1.2).

**Remark 1.1** In the literature (see [5, 20]) a Lipschitz condition imposed on $\beta$ is often written as

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

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 $\beta$ is Lipschitz with respect to its first variable in an integral sense. Our uniformly Lipschitz requirement on $\beta$ and $\lambda$ being a finite measure is enough to satisfy the said restriction.

### 1.3 Forward backward formulation of (1.2)

In the previous context consider the following stochastic setting for (1.2). Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, $\mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq T}$, be a stochastic basis satisfying the usual conditions: $\mathbb{F}$ is right continuous, and $\mathcal{F}_0$ is complete (contains all zero measure sets). The filtration $\mathbb{F}$ is generated by a $d$-dimensional Brownian motion (BM) $W = (W_t)_{0 \leq t \leq T}$ and a Poisson random measure $\mu$ on $\mathbb{R}_+ \times \mathbb{R}^d$ with intensity measure $\lambda$, these two random objects are assumed to be mutually independent.

Recall that $\lambda$ is a finite Lévy measure on $\mathbb{R}^d$. The compensated measure of $\mu$ is denoted as

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

(1.6)
and is such that for every measurable set $A$ satisfying $\lambda(A) < \infty$, $(\mu(t, A) := \mu([0, t], A))_t$ is a martingale. Given a time $t_i \in [0, T]$, the operator $\mathbb{E}_i$ will denote the conditional expectation with respect to $\mathcal{F}_{t_i}$:

$$\mathbb{E}_i (X) := \mathbb{E} \left( X \big| \mathcal{F}_{t_i} \right). \quad (1.7)$$

Recall the equation (1.2)–(1.3)–(1.5). As usual, $X_{r-}$ denotes the a.e. limit of $X_s$ as $s \uparrow r$. Let us consider the next forward and backward stochastic differential equations with jumps in terms of the unknown variables $(X, Y, Z, U)$:

$$X_t = x + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s + \int_0^t \beta(X_s-, y)\mu(ds, dy), \quad (1.8)$$

$$Y_t = g(X_T) + \int_t^T f(\Theta_s)ds - \int_t^T Z_s \cdot dW_s - \int_t^T U_s(y)\mu(ds, dy), \quad (1.9)$$

$$\Gamma_t = \int_{\mathbb{R}^d} U_t(y)\lambda(dy), \quad (1.10)$$

where $\Theta_s = (s, X_s, Y_s, Z_s, \Gamma_s)$ for $0 \leq s \leq T$ and $x \in \mathbb{R}^d$. Note that $(Z_t)_0 \leq t \leq T$ is a vector valued process.

By applying Itô's lemma (see [20, Thm 2.3.4]) to the solution $X_t$ in (1.8) and a $C^{1,2}([0, T] \times \mathbb{R}^d)$ solution $u$ of PIDE (1.2) as $Y_t$ in (1.9), we obtain the compact stochastic formulation of (1.2):

$$u(t, X_t) = u(0, X_0) - \int_0^t f(s, X_{s-}, u(s, X_{s-}), \sigma(X_{s-})\nabla u(s, X_{s-}), T[u](s, X_{s-}))ds,$$

$$+ \int_0^t [\sigma(X_{s-})\nabla u(s, X_{s-})] \cdot dW_s + \int_0^t \int_{\mathbb{R}^d} [u(s, X_{s-}) + \beta(X_{s-}, y)) - u(s, X_{s-})]\mu(ds, dy), \quad (1.11)$$

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

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

must satisfy (1.11) in some proper metric. An important statement here is that the conditions (1.4) ensure the existence of a viscosity solution $u \in C([0, T] \times \mathbb{R}^d)$ with at most polynomial growth such that $u(t, X_t) = Y_t$ (see [5, Thm 3.4]), and this is the reason why our scheme seek to approximate the solution to the FBSDEJ (1.8–1.9). In order to present our Neural Networks algorithm, we introduce them in Sect. 2.

### 1.4 Organization of this paper

The rest of this work is organized as follows. In Sect. 2 we give a concrete definition of the NNs that we will be using together with the approximation results needed in this paper. In Sect. 3 we introduce the discretization of the stochastic system that allows us to train our NNs. In Sect. 4 we state all the preliminary results and definitions needed for the proof of our main result. Finally, in Sect. 5 we state and prove the main result of this paper.
2 Neural networks and approximation theorems

Neural Networks (NNs) are not recent. In [41, 43], published in 1943 and 1958 respectively, the authors introduce the concept of a 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 [30, 36], using the Stone-Weierstrass theorem. These papers state that under suitable conditions on the approximated functions, measured in some mathematical terms, NNs have a very good performance. See [2, 48] for a review on the origin and state of the art survey of DL, respectively.

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 \( f : X \rightarrow Y \), 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 \( D = \{ (x_i, f(x_i)) : x_i \in A \} \) 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 \( \{ f_\theta : \theta \in \Xi \} \) where we can search for a good approximation of \( f \), with \( \Xi \subset \mathbb{R}^k \) for some \( k \in \mathbb{N} \). Finally, how good the approximation is, will be measured by a cost function \( L(\cdot; D) : \Xi \rightarrow \mathbb{R} \) and therefore, intuitively, we take \( f_{\theta^*} \) as the chosen approximation where \( \theta^* \) minimizes \( L(\cdot; D) \) over \( \Xi \).

The complexity and generality of the main 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 [7] for a review of ML in astronomy and [39] for a concrete use of Convolutional Neural Networks (CNN) to classify light curves. See [13] for a review of ML on experimental high energy physics and [47] for an application of NN on quantum state tomography. In [40], the authors use DL to find patterns in fashion and style trends by space and time using data from Instagram. In [3] the authors train a CNN to classify brain tumors into Glioma, Meningioma, and Pituitary Tumor reaching high levels of accuracy. See [37] for a survey on the use of DL in medical science where CNN are the most common type of DL structure.

To fix ideas, in this paper we focus on a simpler setting, where the input and output variables belong to multidimensional real spaces \( \mathbb{R}^d \) and \( \mathbb{R}^m \) respectively with \( d, m \in \mathbb{N} \). In order to define the family of candidates we need \( L + 1 \in \mathbb{N} \) layers with \( l_i \in \mathbb{N} \) neurons each for \( i \in \{ 0, ..., L \} \) where \( l_0 = d \) and \( l_L = m \), weight matrices \( \{ W_i \in \mathbb{R}^{l_i \times l_{i-1}} \}_{i=1}^L \), bias vectors \( \{ b_i \in \mathbb{R}^{l_i} \}_{i=1}^L \), and an activation function \( \phi : \mathbb{R} \rightarrow \mathbb{R} \) which is a way to break the linearity (see Definition 2.1). The resulting function will take values from the input space \( \mathbb{R}^{l_0} \) to the output space \( \mathbb{R}^{l_L} \).

**Remark 2.1** The first and last layer are called input and output layer respectively, the others \( L - 1 \) are often called hidden layers.

**Definition 2.1** Given \( L \in \mathbb{N} \) and \( l_0, (l_i, W_i, b_i)_{i=1}^L \) as above, consider the parameter \( \theta = (W_i, b_i)_{i=1}^L \) which can be seen as an element of \( \mathbb{R}^k \) with \( k = \sum_{i=1}^L (l_i l_{i-1} + l_i) \) and a function \( \phi : \mathbb{R} \rightarrow \mathbb{R} \). We define the neural network \( f(\cdot; \theta, \phi) : \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{l_L} \) as the following...
composition,

\[ f(x; \theta, \phi) = (A_L \circ \phi \circ A_{L-1} \circ \cdots \circ A_2 \circ \phi \circ A_1) (x), \]

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 \( \phi \) is applied component-wise. We denote \( f(\cdot; \theta, \phi) = f(\cdot; \theta) \) when the activation function is fixed and no confusion can arise.

In the following, the activation function will be fixed as well as the input and output dimensions, this is because those parameters are given by the mapping that we are trying to approximate, and the amount of layers \( L \) will also be fixed. The range of neural networks that we can reach varying the remaining parameters, which are the size and values of the weight matrices and bias vectors, will be called the set of neural networks and denoted by \( \mathcal{N}_{\phi,L,l_0,l_L} \). The following definition materializes the previous explanation.

**Definition 2.2** The set of Neural Networks associated to \((L, l_0 = d, l_L = m) \subset \mathbb{N}\) and the function \( \phi : \mathbb{R} \to \mathbb{R} \) is defined by,

\[ \mathcal{N}_{\phi,L,d,m} = \bigcup_{\kappa \in \mathbb{N}} \mathcal{N}_{\phi,L,d,m,\kappa} \]

where,

\[ \mathcal{N}_{\phi,L,d,m,\kappa} = \left\{ f(\cdot; \theta, \phi) \mid \theta = (W_i, b_i)_{i=1}^L, l_0 = d, l_L = m, W_i \in \mathbb{R}^{l_i \times l_{i-1}}, b_i \in \mathbb{R}^{l_i}, l_i \in \mathbb{N}, \right. \]

\[ i \in \{1, ..., L\}, \kappa = \sum_{i=1}^L (l_i l_{i-1} + l_i) \}\]

In this paper \( m \) will be typically \( d \), the space dimension of the PIDE (1.2), or 1. Using functional analysis arguments, K. Hornik proves in [30] that NNs are able to approximate functions in \( L^2 \) spaces for any given tolerance. The space of NNs used in his work, to which we refer as \( \mathcal{H} \), is slightly different from ours, indeed,

\[ \mathcal{H} = \mathcal{N}_{\phi,2,d,1} \cap \left\{ f(\cdot; \theta, \phi) \in \mathcal{N}_{\phi,2,d,1} \mid \theta = (W_1, b_1, W_2, 0) \in \mathbb{R}^{nd+n+n+1}, n \in \mathbb{N} \right\}. \]

Note that in this space the free parameter \( \kappa \) 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 \( f \in \mathcal{H} \) takes the following form

\[ f(x; \theta, \phi) = W_2 \cdot \phi(W_1 x + b_1), \]

for \((W_1, b_1, W_2, 0) \in \mathbb{R}^{nd+n+n+1} \) and \( n \in \mathbb{N} \). Hornik proves the following important result.

**Theorem 2.1 ([30], Theorem 1)** If \( \phi : \mathbb{R} \to \mathbb{R} \) is bounded and non-constant, then \( \mathcal{H} \) is dense in \( L^2(\mathbb{R}^d, B(\mathbb{R}^d), \mu; \mathbb{R}) \) for every finite measure \( \mu \) in \( \mathbb{R}^d \).

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

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

for \( h = (h_1, ..., h_m) \) and \( h_i \) a scalar function for \( i \in \{1, ..., m\} \). We also need to approximate the derivative \( \nabla u \) of the solution \( u \) to PIDE (1.2), the following proposition proves density of NNs in the space of square integrable vector valued functions.
Lemma 2.1 Let $m \in \mathbb{N}$ with $m \geq 1$. If the activation function $\phi$ is bounded and non-constant, then $N_{\phi, 2, d, m}$ is dense in $L^2(\mathbb{R}^d, B(\mathbb{R}^d), \mu; \mathbb{R}^m)$ for every finite measure $\mu$ on $\mathbb{R}^d$.

**Proof** Given $\varepsilon > 0$ and a function $h = (h_1, \ldots, h_m) \in L^2(\mathbb{R}^d, B(\mathbb{R}^d), \mu; \mathbb{R}^m)$ we need to find $f(\cdot; \theta, \phi) = (f_1, \ldots, f_m) \in N_{\phi, 2, d, m}$ such that

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

First, observe that $\mathcal{H} \subset N_{\phi, 2, d, 1}$ which implies, by using Theorem 2.1, that $N_{\phi, 2, d, 1}$ is also dense in $L^2(\mathbb{R}^d, B(\mathbb{R}^d), \mu; \mathbb{R})$ and therefore for every $i \in \{1, \ldots, m\}$ we can find $f_i(\cdot; \theta^i, \phi)$ with $\theta^i = (W^i_1, b^i_1, W^i_2, b^i_2)$ and $\kappa^i = n^i + n^i + n^i + 1$, depending on $\varepsilon$, such that

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

Consider $f \in N_{\phi, 2, d, m}$ defined by $\tilde{\theta} = (\tilde{W}_1, \tilde{b}_1, \tilde{W}_2, \tilde{b}_2)$ with

$$\tilde{W}_1 = \begin{pmatrix} W^1_1 \\ \vdots \\ W^m_1 \end{pmatrix} \in \mathbb{R}^{(\sum_{i=1}^m n^i) \times d}, \quad \tilde{b}_1 = \begin{pmatrix} b^1_1 \\ \vdots \\ b^m_1 \end{pmatrix} \in \mathbb{R}^{\sum_{i=1}^m n^i},$$

$$\tilde{W}_2 = \begin{pmatrix} W^{1,T}_2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & W^{m,T}_2 \end{pmatrix} \in \mathbb{R}^{m \times \sum_{i=1}^m n^i}, \quad \tilde{b}_2 = \begin{pmatrix} b^1_2 \\ \vdots \\ b^m_2 \end{pmatrix} \in \mathbb{R}^{m},$$

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

$$f(x; \tilde{\theta}, \phi) = \tilde{W}_2 \phi(\tilde{W}_1 x + \tilde{b}_1) + \tilde{b}_2 = \begin{pmatrix} W^{1,T}_2 & \phi(W^1_1 x + b^1_1) + b^1_2 \\ \vdots \\ W^{m,T}_2 & \phi(W^m_1 x + b^m_1) + b^m_2 \end{pmatrix} \begin{pmatrix} f_1(x; \theta^1, \phi) \\ \vdots \\ f_m(x; \theta^m, \phi) \end{pmatrix}.$$ 

Therefore,

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

This ends the proof. 

Lemma 2.1 allows us to state that if we take some function $h : \mathbb{R}^d \to \mathbb{R}^m$ in $L^2(\mathbb{R}^d, B(\mathbb{R}^d), \mu; \mathbb{R}^m)$, then the quantity

$$\inf_{\theta \in \mathbb{R}^d} \int_{\mathbb{R}^d} |f(x; \theta, \phi) - h(x)|^2 \mu(dx) \quad (2.1)$$

can be made arbitrarily small by possibly making $\kappa$ growing sufficiently large, whenever $\mu$ is a finite measure on $\mathbb{R}^d$ and the activation function that defines the NN is bounded and non-constant. The following lemma states the $L^2(\Omega, \mathcal{F}, \mathbb{P})$ integrability of a random variable constructed from a NN, it assumes an activation function with linear growth which is more general that the ones presented here.

**Lemma 2.2** Let $m \in \mathbb{N}$ with $m \geq 1$, $X \in L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^d)$ and $f(\cdot; \theta, \phi) \in N_{\phi, 2, d, m}$. Assume that $|\phi(x)| \leq C(1 + |x|)$ for $x \in \mathbb{R}$ and some positive constant $C$, then $f(X; \theta, \phi) \in L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^m)$. 

\(\square\)
Proof Let $f$ be represented by $\theta = (W_1, b_1, W_2, b_2) \in \mathbb{R}^{nd+n+mn+m}$ then,

$$
\mathbb{E}|f(X; \theta, \phi)|^2 = \mathbb{E}|W_2 \phi(W_1 X + b_1) + b_2|^2 = \sum_{i=1}^{m} \mathbb{E}|W_{2,i} \cdot \phi(W_1 X + b_1) + b_{2,i}|^2.
$$

Therefore, without loss of generality we can assume $f \in \mathcal{N}_{\phi,2,d,1}$ and take $\theta = (W_1, b_1, W_2, 0) \in \mathbb{R}^{nd+n+mn+1}$. By using the growth condition on $\phi$ and that for $a, b, c \in \mathbb{R}$ $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$

$$
\mathbb{E}|f(X; \theta, \phi)|^2 = \mathbb{E}|W_2 \cdot \phi(W_1 X + b_1)|^2
$$

\[
\leq \mathbb{E} \left( \sum_{i=1}^{n} |W_{2,i}|^2 \right)^2 \leq \mathbb{E} \left( \sum_{i=1}^{n} |W_{2,i}|^2 \right)^2 \leq 3C^2 \left[ \left( \sum_{i=1}^{n} |W_{2,i}|^2 \right)^2 + \left( \sum_{j=1}^{d} \mathbb{E} |W_{1,i,j}|^2 \right)^2 \right] < \infty.
\]

Note that the first two terms in the last expression are deterministic and finite, then, by using Cauchy-Schwarz twice on the third term we get,

$$
\mathbb{E} \left( \sum_{i=1}^{n} \sum_{j=1}^{d} \mathbb{E} |W_{1,i,j}||X_j|^2 \right)^2 \leq \sum_{i=1}^{n} |W_{2,i}|^2 \sum_{i=1}^{n} \sum_{j=1}^{d} |W_{1,i,j}|^2 \mathbb{E}|X|^2 < \infty.
$$

This finishes the proof. \( \square \)

### 3 Discretization of the dynamics and the deep learning algorithm

Fix a constant step partition of the interval $[0, T]$, defined as $\pi = \{ \frac{iT}{N} \}_{i \in \{0, \ldots, N\}}$, $t_i = \frac{iT}{N}$, 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 $\mu$ from (1.6). Let

$$
M_t := \mu((0, t], \mathbb{R}^d) \quad \text{and} \quad \Delta M_t := \mu((t_i, t_{i+1}], \mathbb{R}^d) = \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \mu(ds, dy). \quad (3.1)
$$

It is well-known that an Euler scheme for the first equation in (1.8) obeys the form

$$
X^\pi_0 = x, \quad (3.2)
$$

$$
X^\pi_{t_{i+1}} = X^\pi_{t_i} + b(X^\pi_{t_i})h + \sigma(X^\pi_{t_i})\Delta W_i + \int_{\mathbb{R}^d} \beta(X_{t_i}, y)\mu((t_i, t_{i+1}], dy). \quad (3.3)
$$
Note that this scheme neglects the left limits that appears on the original equation, although, it satisfies the next error bound (see [20, Thm. 5.1.1], [12] or [26]),

$$\max_{i=1,...,N} \mathbb{E} \left( \sup_{t \in [t_i, t_{i+1}]} |X_t - X_t^\pi|^2 \right) = O(h). \quad (3.4)$$

Under suitable conditions, mostly Lipschitz and linear growth assumptions, it can be proved that the constant behind $O(h)$ in (3.4) does not depend exponentially on $d$, see Lemma 4.3 in [26]. Adapting the argument of [31] to the non-local case, and in view of (1.11), we propose the following modified Euler scheme: for $i = 0, 1, \ldots, N$,

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

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

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

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

**Remark 3.1** Note that the nonlocal term in (1.2) forces us to define $F_i$ in such a way that its fifth argument must be a function $\psi$ 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 3.2** 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 $\tilde{\mu} ((t_i, t_{i+1}), dy)$. 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 $\tilde{\mu}$ behaves on the time interval $(t_i, t_{i+1})$.

Recall Theorem 2.1 and let $\phi : \mathbb{R} \to \mathbb{R}$ be a bounded and non-constant activation function. From now on we will be using NNs with a single hidden layer parameterized by $\theta \in \mathbb{X}$, where $\mathbb{X} = \mathbb{R}^\kappa$ 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,

$$\mathcal{U}_i (\cdot; \theta) : \mathbb{R}^d \to \mathbb{R} \quad (3.5)$$

$$\mathcal{Z}_i (\cdot; \theta) : \mathbb{R}^d \to \mathbb{R}^d \quad (3.6)$$

$$\mathcal{G}_i (\cdot, \circ; \theta) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \quad (3.7)$$

with $\mathcal{U}_i (\cdot; \theta) \in \mathcal{N}_{\phi, 2, d, 1, \kappa}$, $\mathcal{Z}_i (\cdot; \theta) \in \mathcal{N}_{\phi, 2, d, d, \kappa}$ and $\mathcal{G}_i (\cdot, \circ; \theta) \in \mathcal{N}_{\phi, 2, d+d, 1, \kappa}$ 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. Note that we dropped Let also

$$\langle \mathcal{G} \rangle_t (x; \theta) = \int_{\mathbb{R}^d} G_i(x, y; \theta) \lambda(dy). \quad (3.8)$$
We propose an extension of the DBDP1 algorithm presented on [31]. The main idea of the algorithm is that the NNs, evaluated on $X^\pi_{ti+1}$, are good approximations of the processes solving the FBSDEJ. Let $L_i$ be a cost function defined for $\theta \in \Xi$ as

$$L_i(\theta) = \mathbb{E}\left[|\hat{U}_{ti+1}(X^\pi_{ti+1}) - F_i(t_i, X^\pi_{ti}, U_i(X^\pi_{ti}, \theta), Z_i(X^\pi_{ti}, \theta), G_i(X^\pi_{ti}, \cdot; \theta), h, \Delta W_t)|^2 \right].$$

(3.9)

**Algorithm 1: DBDP1 PIDE extension**

Start with $\hat{U}_N(\cdot) = g(\cdot);

for $i \in \{N-1, \ldots, 1\}$ do

- Given $\hat{U}_{i+1}$;
- Compute $\theta^* = \text{argmin } L_i(\theta)$;
- Update $(\hat{U}_i, \hat{Z}_i, \hat{G}_i) = (U_i(\cdot; \theta^*), Z_i(\cdot; \theta^*), G_i(\cdot, \cdot; \theta^*))$;

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 [31], one has to use a Monte Carlo approximation together with Stochastic Gradient Descent (SGD). See also Remark 3.1.

### 4 Preliminaries

For a general measure space $(E, \Sigma, m)$, $p \geq 1$ and $m \in \mathbb{N}$, recall the definition and norm for the Lebesgue space $L^p(E, \Sigma, m)$ introduced in Sect. 1.1. For $s, t \in [0, T]$ such that $s \leq t$, we define some spaces of stochastic processes:

- $S^2_{[s,t]}(\mathbb{R})$ denotes the space of adapted càdlàg processes $Y : \Omega \times [s, t] \to \mathbb{R}$ such that

$$\|Y\|_{S^2_{[s,t]}}^2 := \mathbb{E}\left(\sup_{r \in [s,t]} |Y_r|^2 \right) < \infty.$$  

- $L^2_{W,[s,t]}(\mathbb{R}^d)$ denotes the space of predictable processes $Z : \Omega \times [s, t] \to \mathbb{R}^d$ such that

$$\|Z\|_{L^2_{W,[s,t]}}^2 := \mathbb{E}\left(\int_s^t \|Z_r\|^2 dr \right) < \infty.$$  

- $L^2_{\mu,[s,t]}(\mathbb{R})$ denotes the space of $\sigma(\mathcal{P}_{[s,t]} \times \mathcal{B}(\mathbb{R}^d))$-measurable processes $U : \Omega \times [s, t] \times \mathbb{R}^d \to \mathbb{R}$ with $\mathcal{P}_{[s,t]}$ denoting the predictable sigma algebra on $\Omega \times [s, t]$. These processes are such that

$$\|U\|_{L^2_{\mu,[s,t]}}^2 := \mathbb{E}\left(\int_s^t \int_{\mathbb{R}^d} |U_r(y)|^2 \lambda(dy) dr \right) < \infty.$$  

Whenever $[s, t] = [0, T]$, we avoid mentioning the interval of time, and denote $\mathcal{B}^2 = S^2 \times L^2_W(\mathbb{R}^d) \times L^2_{\mu}(\mathbb{R})$. In the following, $C > 0$ will denote a constant that may change from one line to another. Also, the notation $a \lesssim b$ means that there exists $C > 0$ such that $a \leq Cb$.

Springer
4.1 Existence and uniqueness for the FBSDEJ

In order to estimate errors we need a solution to compare, the following lemmas present well-known results concerning the existence and uniqueness of a solution to the decoupled system (1.8–1.9). We only check that our hypotheses match those of [4] and [5], these results are the same as those given in [12] and [20, Section 4.1].

**Lemma 4.1** There exists a unique solution \( X \in S^2 \) to (1.8) such that.

\[
\mathbb{E} \left( \sup_{s \leq u \leq t} |X_u - X_s|^2 \right) \leq |t - s| \left( 1 + \mathbb{E} |X_s|^2 \right).
\] (4.1)

**Proof** Recall Remark 1.1. Observe that conditions \((\text{C})\), particularly those imposed on \( \beta \), imply that

\[
\int_{\mathbb{R}^d} |\beta(x, y) - \beta(x', y)|^2 \lambda(dy) \leq K^2 \lambda(\mathbb{R}^d) |x - x'|^2.
\]

This, together with the rest of conditions \((\text{C})\) are enough to fulfill the Lipschitz and growth hypotheses needed on \([4, \text{Section 6.2}]\) to ensure the existence and uniqueness of a solution \( X \in S^2 \) to the FSDEJ (1.8). Estimate (4.1) follows by considering the process \((X_u - X_s)_{u \in [s, t]}\) and using Doob’s maximal inequality \([42, \text{Theorem 20, Section 1}]\) and Gronwall inequality.

\( \square \)

**Lemma 4.2** There exists a solution \((Y, Z, U) \in B^2\) to (1.9).

**Proof** We apply Theorem 2.1 of [5] with \( k = 1, Q = g(X_T) \) and a nonlinearity \( \tilde{f} : \Omega \times [0, T] \times \mathbb{R}^d \times L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda) \to \mathbb{R} \) defined as

\[
\tilde{f}(\omega, t, y, z, w) = f \left( t, X_t(\omega), y, z, \int_{\mathbb{R}^d} w(x) \lambda(dx) \right),
\]

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 4.1 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 \( w, w' \in L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)\),

\[
|\tilde{f}(\omega, t, y, z, w) - \tilde{f}(\omega, t, y', z', w')|
\]

\[
= \left| f \left( t, X_t(\omega), y, z, \int_{\mathbb{R}^d} w(x) \lambda(dx) \right) - f \left( t, X_t(\omega), y', z', \int_{\mathbb{R}^d} w'(x) \lambda(dx) \right) \right|
\]

\[
\leq K \left( |y - y'| + |x - x'| + \left| \int_{\mathbb{R}^d} (w - w') \lambda(dy) \right| \right)
\]

\[
\leq K \left( |y - y'| + |x - x'| + \lambda(\mathbb{R}^d)^{1/2} \| w - w' \|_{L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda)} \right),
\]

this proves the Lipschitz condition on \( \tilde{f} \). Using the previous bound is clear that,

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

These computations allow us to directly apply Theorem 2.1 of [5] this finishes the proof. \( \square \)
Combining previous lemmas we get that there exist a unique solution \((X, Y, Z, U)\) to the system (1.8–1.9) in the space \(S^2 \times B^2\) this implies,

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

(4.2)

### 4.2 Useful results from stochastic calculus

The 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 \(\mu\), which allows us to state the representation property. See the end of Section 2.4 in [20] where it is stated that when the filtration is generated by a Brownian Motion and an independent jump process the required representation holds.

**Lemma 4.3** (Martingale Representation Theorem) For any square integrable martingale \(M\) there exists \((Z, U) \in L^2_W(\mathbb{R}^d) \times L^2_{\mu}(\mathbb{R})\) such that for \(t \in [0, T]\)

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

We will need the next property involving conditional expectation, Itô isometry and that \(W\) is independent of \(\overline{\pi}\).

**Lemma 4.4** (Conditional Ito isometry) For \(V^1, V^2 \in L^2_\mu(\mathbb{R})\) and \(H, K \in L^2_W(\mathbb{R}^d),\)

\[
\mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} H_r \cdot dW_r \int_{t_i}^{t_{i+1}} K_r \cdot dW_r \right) = \mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} H_r \cdot K_r dr \right),
\]

(4.3)

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

\[
\mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} V^1(s, z) \lambda(dz) ds \right) = \mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} H_r \cdot dW_r \right) = 0.
\]

**Proof** Follows from the classical Itô isometry and using the definition of conditional expectation.

**Lemma 4.5** (Conditional Fubini) Let \(H \in L^2_\mu(\mathbb{R}^d)\) be a \(\mathbb{F}\)-adapted process and \(t > 0\), then

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

**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

\[
\int_A \left( \int_{\mathbb{R}^d} \mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} H(s, y) ds \right) \lambda(dy) \right) d\mathbb{P}(\omega) = \int_A \left( \int_{\mathbb{R}^d} \int_{t_i}^{t_{i+1}} H(s, y) ds \lambda(dy) \right) d\mathbb{P}(\omega).
\]
Note that because of $H \in L^2(\mu(\mathbb{R}^d))$,
\[
\int_{\Omega} \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} |H(s, y)(\omega)|^2 \lambda(dy) ds d\mathbb{P}(\omega) < \infty;
\]
this means that $H$ can be seen as an element of $L^2(\Omega \times [t_i, t_{i+1}] \times \mathbb{R}^d) \subset L^1(\Omega \times [t_i, t_{i+1}] \times \mathbb{R}^d)$, both spaces endowed with the correspondent finite product measure. Then we can use classical Fubini theorem:
\[
\int_{A} \left( \int_{\mathbb{R}^d} \int_{t_i}^{t_{i+1}} H(s, y)(\omega) ds d\mathbb{P}(\omega) \right) \lambda(dy) = \int_{A} \left( \int_{\mathbb{R}^d} \left( \int_{t_i}^{t_{i+1}} H(s, y)(\omega) ds \right) d\mathbb{P}(\omega) \right) \lambda(dy) = \int_{A} \left( \int_{\mathbb{R}^d} \left( \int_{t_i}^{t_{i+1}} H(s, y)(\omega) ds \right) \lambda(dy) \right) d\mathbb{P}(\omega).
\]
This finishes the proof. \(\Box\)

### 4.3 Measuring the error

We first introduce the conditional expectations of the averaged processes
\[
\overline{Z}_t = \frac{1}{h} \mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} Z_t dt \right), \quad \overline{\Gamma}_t = \frac{1}{h} \mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} \Gamma_t dt \right).
\] (4.4)

These quantities allow us to define the $L^2$-regularity of the solutions $(Z, \Gamma)$ (see [12] and [31]) as follows
\[
\varepsilon^Z(h) := \mathbb{E} \left( \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} |Z_t - \overline{Z}_t|^2 dt \right), \quad \varepsilon^\Gamma(h) := \mathbb{E} \left( \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} |\Gamma_t - \overline{\Gamma}_t|^2 dt \right). \] (4.5)

Both quantities can be made arbitrarily small as it is shown on [12] and presented in the following theorem.

**Theorem 4.1** Under assumptions (C), there exists a constant $C > 0$ such that
\[
\varepsilon^\Gamma(h) \leq Ch \quad \text{and} \quad \varepsilon^Z(h) \leq Ch.
\]

**Proof** See [12, Theorem 2.1 (i)] for the bound on $\varepsilon^\Gamma(h)$ and [12, Theorem 2.1 (ii)] for the bound on $\varepsilon^Z(h)$. Note that in the cited reference this result is presented, using our notation, as follows,
\[
\|\Gamma - \overline{\Gamma}\|^2_{L^2_W} \leq CN^{-1} \quad \text{and} \quad \|Z - \overline{Z}\|^2_{L^2_W} \leq CN^{-1}.
\]
Where $\Gamma_t = \Gamma_{t_i}$ for $t \in [t_i, t_{i+1})$ and $\overline{Z}_t = \overline{Z}_{t_i}$ for $t \in [t_i, t_{i+1})$. \(\Box\)
We introduce a somehow auxiliary scheme that at the same time depends on the main one. Let \( i \in \{0, \ldots, N-1\} \), as stated in Sect. 3. We follow the procedure taken in [31], with key modifications. Let us use the ideas of [12] to define \( \mathcal{F} \)-adapted discrete processes

\[
\widehat{V}_{ti} = \mathbb{E}_t \left( \hat{U}_{i+1}(X_{ti+1}^\pi) + f \left( t_i, X_{ti}^\pi, \widehat{Z}_{ti}, \widehat{\Gamma}_{ti} \right) \right) h, \tag{4.6}
\]

\[
\widehat{Z}_{ti} = \frac{1}{h} \mathbb{E}_t \left( \hat{U}_{i+1}(X_{ti+1}^\pi) \Delta W_i \right), \tag{4.7}
\]

\[
\widehat{\Gamma}_{ti} = \frac{1}{h} \mathbb{E}_t \left( \hat{U}_{i+1}(X_{ti+1}^\pi) \Delta M_i \right), \tag{4.8}
\]

where \( \widehat{V}_{ti} \) is well-defined for sufficiently small \( h \) by Lemma 4.6 and the variables \( \widehat{Z}_{ti}, \widehat{\Gamma}_{ti} \) are defined below.

**Lemma 4.6** The process \( \widehat{V}_{ti} \) is well-defined.

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

\[
\psi(\xi)(\omega) = \mathbb{E}_t \left( \hat{U}_{i+1}(X_{ti+1}^\pi)(\omega) + f \left( t_i, X_{ti}^\pi(\omega), \xi(\omega), \widehat{Z}_{ti}(\omega), \widehat{\Gamma}_{ti}(\omega) \right) \right) h.
\]

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 2.2. Let \( \xi, \bar{\xi} \in L^2 \), then \( \mathbb{P} \) a.s \( |\psi(\xi) - \psi(\bar{\xi})| \leq h|\xi - \bar{\xi}| \), therefore

\[
\|\psi(\xi) - \psi(\bar{\xi})\|_{L^2(\Omega, \mathcal{F}, \mathbb{P})} \leq h\|\xi - \bar{\xi}\|_{L^2(\Omega, \mathcal{F}, \mathbb{P})}
\]

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. \( \square \)

For fixed \( i \in \{0, \ldots, N\} \), let \( N_t \) be a process defined as \( N_t := \mathbb{E} \left( \hat{U}_{i+1}(X_{ti+1}^\pi) | \mathcal{F}_t \right) \) for \( t \in [t_i, t_{i+1}] \). Using Lemma 2.2, it is not difficult to see that \( N_t \) is a square integrable martingale and therefore, by Martingale Representation Theorem (see Lemma 4.3), there exist \( (\hat{Z}, \hat{U}) \in L^2_{\mu} \times L^2_{W} \) such that

\[
N_t = N_{t_i} + \int_{t_i}^t \hat{Z}_s \cdot dW_s + \int_{t_i}^t \int_{\mathbb{R}^d} \hat{U}_s(y)\eta(ds, dy).
\]

By taking \( t = t_{i+1} \) and using (1.7),

\[
\hat{U}_{i+1}(X_{ti+1}^\pi) = \mathbb{E}_t \left( \hat{U}_{i+1}(X_{ti+1}^\pi) \right) + \int_{t_i}^{t_{i+1}} \hat{Z}_s \cdot dW_s + \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \hat{U}_s(y)\eta(ds, dy).
\]

By multiplying by \( \Delta W_i \) and \( \Delta M_i \), then taking \( \mathbb{E}_t \) and using Itô isometry,

\[
\widehat{Z}_{ti} = \frac{1}{h} \mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} \hat{Z}_s ds \right),
\]

\[
\widehat{\Gamma}_{ti} = \frac{1}{h} \mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \hat{U}_s(y)\lambda(dy)ds \right).
\]

Let

\[
\widehat{U}_{ti}(y) := \frac{1}{h} \mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} \hat{U}_s(y)ds \right) \tag{4.9}
\]

By Lemma 4.5 one can see that

\[
\widehat{\Gamma}_{ti} = \frac{1}{h} \mathbb{E}_t \left( \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \hat{U}_s(y)\lambda(dy)ds \right) = \int_{\mathbb{R}^d} \widehat{U}_{ti}(y)\lambda(dy). \tag{4.10}
\]
The last equality can be seen as an analogous to (1.10) and makes sense with the notation $\tilde{\Gamma}_{t_i} = \langle \tilde{U}_{t_i} \rangle$. Also, we can establish the following useful bound:

$$\mathbb{E} \left| \tilde{\Gamma}_{t_i} - \langle G \rangle_i (X_{t_i}^{\pi}; \theta) \right|^2 \lesssim \mathbb{E} \left( \left\| \tilde{U}_{t_i} (\cdot) - G_i (X_{t_i}^{\pi}, \cdot, \theta) \right\|_{L^2(\lambda)}^2 \right).$$

Indeed, from (4.10) and (3.8), Hölder inequality and the fact that $\Theta_1$ of these functions is ensured by the properties of $V$ and therefore, applying the Theorem 2.1, these quantities can be made arbitrarily small as $E$ increases.

Following [31], we can find deterministic functions $v_i, z_i, \gamma_i$ such that $v_i (X_{t_i}^{\pi}) = \tilde{V}_{t_i}, z_i (X_{t_i}^{\pi}) = \tilde{Z}_{t_i}$ and $\gamma_i (y, X_{t_i}^{\pi}) = \tilde{U}_{t_i} (y)$ for $y \in \mathbb{R}^d$. The correspondent $L^2$-integrability of these functions is ensured by the properties of $\tilde{V}_{t_i}, \tilde{Z}_{t_i}$ and $\tilde{U}_{t_i}$. With the previous setup, the natural extension of the terms to estimate the error of the scheme shown on [31] must be

$$\mathcal{E}_i^v = \inf_{\xi \in \mathbb{R}^d} \mathbb{E} \left| v_i (X_{t_i}^{\pi}) - U_i (X_{t_i}^{\pi}; \xi) \right|^2, \quad \mathcal{E}_i^z = \inf_{\xi \in \mathbb{R}^d} \mathbb{E} \left| z_i (X_{t_i}^{\pi}) - Z_i (X_{t_i}^{\pi}; \xi) \right|^2,$$

$$\mathcal{E}_i^{\gamma} = \inf_{\xi \in \mathbb{R}^d} \mathbb{E} \left( \int_{\mathbb{R}^d} \left| \gamma_i (y, X_{t_i}^{\pi}) - G_i (X_{t_i}^{\pi}, y; \xi) \right|^2 \lambda(dy) \right). \quad (4.11)$$

The expected values can be written as a integral with respect a probability measure in $\mathbb{R}^d$ and therefore, applying the Theorem 2.1, these quantities can be made arbitrarily small as $\kappa$ increases.

The following results will be useful in the proof of the main result. In Section 2.5 of [12], it is explained that the results presented there still hold for a time-dependent non-linearity.

**Proposition 4.1** ([12], Proposition 2.1) *There exists a constant $C > 0$ independent of the step $h$ such that*

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

We will also need the following result.

**Lemma 4.7** *Consider $(X, Y, Z, U) \in S^2 \times B^2$ the solution to (1.8–1.9), $\Gamma$ defined as in (1.10) 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 (1.1) 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 (1.1) and the Lipschitz bound on $f$,

$$|f(\Theta_s)|^2 \leq 2 \left[ K^2 5(|X_s|^2 + |Y_s|^2 + |Z_s|^2 + |\Gamma_s|^2) + |f(s, 0, 0, 0, 0)|^2 \right].$$
Then, integrating on $\Omega \times [0, T]$ with respect to $d\mathbb{P} \times ds$, using Hölder inequality and bound (4.2),
\[
\mathbb{E}\left(\int_0^T |f(\Theta_s)|^2 ds\right) \leq 10K^2T\mathbb{E}\left(\sup_{s \in [0,T]} |X_s|^2 + \sup_{s \in [0,T]} |Y_s|^2\right) + 10K^2\left(\mathbb{E}\left(\int_0^T |Z_s|^2 ds + \lambda(\mathbb{R}^d)\mathbb{E}\left(\int_0^T \int_{\mathbb{R}^d} |U_s(y)|^2 \lambda(dy) ds\right)\right) + 2T \sup_{s \in [0,T]} |f(s, 0, 0, 0)|^2\right) < \infty
\]
this finishes the proof. \hfill \Box

5 Main result

As stated previously, the proof of our main result, Theorem 5.1, is deeply inspired in the case without jumps considered in [31]. 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 the useful bound (1.1); for $x_1, \ldots, x_k \in \mathbb{R}$ the following holds,
\[
(x_1 + \cdots + x_k)^2 \leq k(x_1^2 + \cdots + x_k^2).
\]

**Theorem 5.1** Under (C), there exists a constant $C > 0$ independent of the partition such that for sufficiently small $h$,
\[
\max_{i=0,\ldots,N-1} \mathbb{E} \left| Y_{t_i} - \widehat{U}_{t_i}(X_{t_i}^\pi) \right|^2 + \sum_{i=0}^{N-1} \mathbb{E} \left( \int_{t_i}^{t_{i+1}} \left[ |Z_t - \hat{Z}_t(X_{t_i}^\pi)|^2 + |\Gamma_t - (\hat{\mathcal{G}})_t(X_{t_i}^\pi)|^2 \right] dt \right) \leq C \left[ h + \sum_{i=0}^{N-1} (N \mathcal{E}^u + \mathcal{E}^z_i + \mathcal{E}^y_i) + \varepsilon^z(h) + \varepsilon^y(h) + \mathbb{E} \left| g(X_T) - g(X_T^\pi) \right|^2 \right],
\]
with $\mathcal{E}^u_i$, $\mathcal{E}^z_i$ and $\mathcal{E}^y_i$ given in (4.11), and $\varepsilon^z(h)$ and $\varepsilon^y(h)$ defined in (4.5).

**Proof Step 1:** Recall $\widehat{V}_{t_i}$ introduced in (4.6). The purpose of this part is to obtain a suitable bound of the term $\mathbb{E} \left| Y_{t_i} - \widehat{V}_{t_i} \right|^2$ in terms of more tractable terms. We have \hfill \Box

**Lemma 5.1** There exists $C > 0$ fixed such that for any $0 < h < 1$ sufficiently small, one has
\[
\mathbb{E} \left| Y_{t_i} - \widehat{V}_{t_i} \right|^2 \leq Ch^2 + C \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds \right) + C \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - Z_{t_i}|^2 ds \right) + C \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\Gamma_s - \Gamma_{t_i}|^2 ds \right) + Ch \mathbb{E} \left( \int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr \right) + C(1 + Ch) \mathbb{E} \left| Y_{t_{i+1}} - \widehat{U}_{t_{i+1}}(X_{t_{i+1}}^\pi) \right|^2,
\]
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 (1.9) between \( t_i \) and \( t_{i+1} \), we obtain

\[
\Delta Y_i = Y_{t_{i+1}} - Y_{t_i} = - \int_{t_i}^{t_{i+1}} f(\Theta_s) ds + \int_{t_i}^{t_{i+1}} Z_s \cdot dW_s + \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} U_s(y) \overline{\mu}(ds, dy). \tag{5.2}
\]

Using the definition of \( \widetilde{\mathcal{V}}_{t_i} \) in (4.6),

\[
Y_{t_i} - \widetilde{\mathcal{V}}_{t_i} = Y_{t_{i+1}} - \Delta Y_i - \widetilde{\mathcal{V}}_{t_i} = Y_{t_{i+1}} + \int_{t_i}^{t_{i+1}} [f(\Theta_s) - f(\Theta_{t_i})] ds - \int_{t_i}^{t_{i+1}} Z_s \cdot dW_s - \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} U_s(y) \overline{\mu}(ds, dy) - E_i(\widetilde{\mathcal{V}}_{t_{i+1}}(X_{t_{i+1}}^\pi)).
\]

Here \( \Theta_{t_i} = (t_i, X_{t_i}^\pi, \widetilde{\mathcal{V}}_{t_i}, \overline{Z}_{t_i}, \overline{\Gamma}_{t_i}) \). Then, by applying the conditional expectation for time \( t_i \) given by \( E_i \) and using that, in this case, the stochastic integrals are martingales,

\[
Y_{t_i} - \widetilde{\mathcal{V}}_{t_i} = E_i(Y_{t_{i+1}} - \widetilde{\mathcal{V}}_{t_{i+1}}(X_{t_{i+1}}^\pi)) + E_i\left( \int_{t_i}^{t_{i+1}} [f(\Theta_s) - f(\Theta_{t_i})] ds \right) = a + b.
\]

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

\[
E \left| Y_{t_i} - \widetilde{\mathcal{V}}_{t_i} \right|^2 \leq (1 + \gamma h)E \left[ E_i\left( Y_{t_{i+1}} - \widetilde{\mathcal{V}}_{t_{i+1}}(X_{t_{i+1}}^\pi) \right) \right]^2 + \left( 1 + \frac{1}{\gamma h} \right) E \left[ E_i\left( \int_{t_i}^{t_{i+1}} [f(\Theta_s) - f(\Theta_{t_i})] ds \right) \right]^2. \tag{5.3}
\]

With no lose of generality, because we are looking for bounds, we can replace \([f(\Theta_s) - f(\Theta_{t_i})] \) by \([f(\Theta_s) - f(\Theta_{t_i})] \). Also, we can drop the \( E_i \) due to the law of total expectation. The Lipschitz condition on \( f \) in (1.4) allows us to give a bound in terms of the difference between \( \Theta_s \) and \( \Theta_{t_i} \). Indeed, for a fixed constant \( K > 0 \),

\[
|f(\Theta_s) - f(\Theta_{t_i})| \leq K \left( |s - t_i|^{1/2} + |X_s - X_{t_i}^\pi| + |Y_s - \widetilde{\mathcal{V}}_{t_i}| + |Z_s - \overline{Z}_{t_i}| + |\Gamma_s - \overline{\Gamma}_{t_i}| \right).
\]

Therefore, we have the bound

\[
E \left( \int_{t_i}^{t_{i+1}} |f(\Theta_s) - f(\Theta_{t_i})| ds \right)^2 \leq C h^2 + E \left( \int_{t_i}^{t_{i+1}} |X_s - X_{t_i}^\pi|^2 ds \right) + E \left( \int_{t_i}^{t_{i+1}} |Y_s - \widetilde{\mathcal{V}}_{t_i}|^2 ds \right) + E \left( \int_{t_i}^{t_{i+1}} |Z_s - \overline{Z}_{t_i}|^2 ds \right) + E \left( \int_{t_i}^{t_{i+1}} |\Gamma_s - \overline{\Gamma}_{t_i}|^2 ds \right),
\]

where the Lipschitz constant \( K \) was absorbed by \( C \). Using now the triangle inequality \(|Y_s - \widetilde{\mathcal{V}}_{t_i}|^2 \leq 2|Y_s - Y_{t_i}|^2 + 2|Y_{t_i} - \widetilde{\mathcal{V}}_{t_i}|^2 \), and the approximation error of the \( X \) scheme (3.4), we
find
\[
\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |f(\Theta_s) - f(\hat{\Theta}_{t_i})| ds \right)^2 \\
\leq C h \left[ h^2 + 2\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds \right) + 2h \mathbb{E} |Y_{t_i} - \hat{\nabla}_{t_i}|^2 \\
+ \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - \hat{Z}_{t_i}|^2 ds \right) + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\hat{\Gamma}_s - \hat{\Gamma}_{t_i}|^2 ds \right) \right].
\] (5.4)

and therefore, replacing in (5.3),
\[
\mathbb{E} |Y_{t_i} - \hat{\nabla}_{t_i}|^2 \\
\leq (1 + \gamma h) \mathbb{E} \left| Y_{t_{i+1}} - \hat{\nabla}_{t+1}(X_{t+1}^{\pi_t}) \right|^2 \\
+ (1 + \gamma h) C \gamma \left[ h^2 + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds \right) + h \mathbb{E} |Y_{t_i} - \hat{\nabla}_{t_i}|^2 \\
+ \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - \hat{Z}_{t_i}|^2 ds \right) + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\hat{\Gamma}_s - \hat{\Gamma}_{t_i}|^2 ds \right) \right].
\] (5.5)

Recall \( Z_{t_i} \) and \( \hat{\Gamma}_{t_i} \) introduced in (4.4). Now, we are going to prove the following
\[
\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - \hat{Z}_{t_i}|^2 ds \right) = \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - \hat{Z}_{t_i}|^2 ds \right) + h \mathbb{E} |Z_{t_i} - \hat{Z}_{t_i}|^2.
\] (5.6)
\[
\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\hat{\Gamma}_s - \hat{\Gamma}_{t_i}|^2 ds \right) = \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\hat{\Gamma}_s - \hat{\Gamma}_{t_i}|^2 ds \right) + h \mathbb{E} |\hat{\Gamma}_{t_i} - \hat{\Gamma}_{t_i}|^2.
\] (5.7)

Let us prove the latter, the former is analogous. Recall that the \( \Gamma \) components represents the nonlocal part and therefore is one dimensional.
\[
|\Gamma_t - \hat{\Gamma}_{t_i}|^2 = |(\Gamma_t - \hat{\Gamma}_{t_i}) + (\hat{\Gamma}_{t_i} - \hat{\Gamma}_{t_i})|^2 \\
= (\Gamma_t - \hat{\Gamma}_{t_i})^2 + (\hat{\Gamma}_{t_i} - \hat{\Gamma}_{t_i})^2 + 2(\Gamma_t - \hat{\Gamma}_{t_i})(\hat{\Gamma}_{t_i} - \hat{\Gamma}_{t_i}).
\]

It is sufficient to establish that the double product is 0 when integrating and taking expectation. Recall that \( \hat{\Gamma}_{t_i} \) from (4.4) is a \( \mathcal{F}_{t_i} \) measurable random variable. Then,
\[
\int_{t_i}^{t_{i+1}} (\Gamma_t - \hat{\Gamma}_{t_i})(\hat{\Gamma}_{t_i} - \hat{\Gamma}_{t_i}) dt = \left( \int_{t_i}^{t_{i+1}} (\Gamma_t - \hat{\Gamma}_{t_i}) dt \right) (\hat{\Gamma}_{t_i} - \hat{\Gamma}_{t_i}) \\
= \left[ \int_{t_i}^{t_{i+1}} \Gamma_t dt - \mathbb{E}_i \left( \int_{t_i}^{t_{i+1}} \Gamma_t dt \right) \right] (\hat{\Gamma}_{t_i} - \hat{\Gamma}_{t_i}).
\]

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 (5.6) and (5.7) are
proven. By multiplying (5.2) by $\Delta W_i$ and taking $\mathbb{E}_i$,

$$
\mathbb{E}_i (\Delta W_i Y_{t_i+1}) + \mathbb{E}_i \left( \int_{t_i}^{t_{i+1}} f(\Theta_{r_i})dr \right) \\
= \mathbb{E}_i \left( \int_{t_i}^{t_{i+1}} dW_r \int_{t_i}^{t_{i+1}} Z_r dW_r \right) \\
+ \mathbb{E}_i \left( \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} U_r(y)\tilde{\mu}(dy, dr) \int_{t_i}^{t_{i+1}} dW_r \right) \\
= \mathbb{E}_i \left( \int_{t_i}^{t_{i+1}} Z_r dr \right) = h\overline{Z}_{t_i},
$$

where we have used Lemma 4.3. Then, subtracting $h\overline{Z}_{t_i} = \mathbb{E}_i (\widehat{\mu}_{t_i+1}(X_{t_i+1}^\pi) \Delta W_i),$

$$
h(\overline{Z}_{t_i} - \overline{Z}_{t_i}) = \mathbb{E}_i \left[ \Delta W_i (Y_{t_i+1} - \widehat{\mu}_{t_i+1}(X_{t_i+1}^\pi)) \right] + \mathbb{E}_i \left( \Delta W_i \int_{t_i}^{t_{i+1}} h(\Theta_{r_i})dr \right).
$$

By multiplying (5.2) by $\Delta M_i$ and taking $\mathbb{E}_i$,

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

Then, subtracting $h\overline{\Gamma}_{t_i} = \mathbb{E}_i \left( \widehat{\mu}_{t_i+1}(X_{t_i+1}^\pi) \Delta M_i \right),$

$$
h(\overline{\Gamma}_{t_i} - \overline{\Gamma}_{t_i}) = \mathbb{E}_i \left[ \Delta M_i \left( Y_{t_i+1} - \widehat{\mu}_{t_i+1}(X_{t_i+1}^\pi) \right) \right] + \mathbb{E}_i \left( \Delta M_i \int_{t_i}^{t_{i+1}} f(\Theta_{r_i})dr \right).
$$

Summarizing, one has

$$
h(\overline{Z}_{t_i} - \overline{Z}_{t_i}) = \mathbb{E}_i \left[ \Delta W_i \left( Y_{t_i+1} - \widehat{\mu}_{t_i+1}(X_{t_i+1}^\pi) \right) - \mathbb{E}_i \left[ Y_{t_i+1} - \widehat{\mu}_{t_i+1}(X_{t_i+1}^\pi) \right] \right] \\
+ \mathbb{E}_i \left[ \Delta W_i \int_{t_i}^{t_{i+1}} f(\Theta_{r_i})dr \right];
$$

$$
h(\overline{\Gamma}_{t_i} - \overline{\Gamma}_{t_i}) = \mathbb{E}_i \left[ \Delta M_i \left( Y_{t_i+1} - \widehat{\mu}_{t_i+1}(X_{t_i+1}^\pi) \right) - \mathbb{E}_i \left[ Y_{t_i+1} - \widehat{\mu}_{t_i+1}(X_{t_i+1}^\pi) \right] \right] \\
+ \mathbb{E}_i \left[ \Delta M_i \int_{t_i}^{t_{i+1}} f(\Theta_{r_i})dr \right].
$$

For the sake of brevity, define now

$$
H_i := Y_{t_i} - \widehat{\mu}_{t_i}(X_{t_i}^\pi); \quad (5.8)
$$
note that it depends on $i$. By the properties related with Itô isometry, from the previous identities we have

$$
\mathbb{E} \left( h^2 \left| Z_{t_i} - \hat{Z}_{t_i} \right|^2 \right) \leq 2dh \left( \mathbb{E}(H_{t_i+1}^2) - \mathbb{E}(H_{t_i+1}^2) \right) + 2dh^2 \mathbb{E} \left[ \int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr \right]; \quad (5.9)
$$

$$
\mathbb{E} \left( h^2 \left| \Gamma_{t_i} - \hat{\Gamma}_{t_i} \right|^2 \right) \leq 2\lambda(\mathbb{R}^d) h \left( \mathbb{E}(H_{t_i+1}^2) - \mathbb{E}(H_{t_i+1}^2) \right) + 2\lambda(\mathbb{R}^d) h^2 \mathbb{E} \left[ \int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr \right]. \quad (5.10)
$$

**Remark 5.1** Note that in the previous bound is important the finiteness of the Levy measure $\lambda$. 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 (5.4). Using (5.6) and (5.7),

$$
\mathbb{E} \left| Y_{t_i} - \hat{V}_{t_i} \right|^2 \leq (1 + \gamma h) \mathbb{E} \left| \varepsilon_t(H_{t_i+1}) \right|^2 + (1 + \gamma h) \frac{C}{\gamma} \left[ h^2 + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds \right) + h \mathbb{E} \left| Y_{t_i} - \hat{V}_{t_i} \right|^2 \\
+ \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - Z_{t_i}|^2 ds \right) + h \mathbb{E} \left| Z_{t_i} - \hat{Z}_{t_i} \right|^2 \\
+ \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\Gamma_s - \hat{\Gamma}_{t_i}|^2 ds \right) + h \mathbb{E} \left| \Gamma_{t_i} - \hat{\Gamma}_{t_i} \right|^2 \right].
$$

Now use (5.9) and (5.10) to find that

$$
\mathbb{E} \left| Y_{t_i} - \hat{V}_{t_i} \right|^2 \leq (1 + \gamma h) \mathbb{E} \left| \varepsilon_t(H_{t_i+1}) \right|^2 + (1 + \gamma h) \frac{C}{\gamma} \left[ h^2 + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds \right) + h \mathbb{E} \left| Y_{t_i} - \hat{V}_{t_i} \right|^2 \\
+ \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - Z_{t_i}|^2 ds \right) + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\Gamma_s - \hat{\Gamma}_{t_i}|^2 ds \right) \\
+ 2d \left( \mathbb{E}(H_{t_i+1}^2) - \mathbb{E}(H_{t_i+1}^2) \right) + 2dh \mathbb{E} \left[ \int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr \right] \\
+ 2\lambda(\mathbb{R}^d) \left( \mathbb{E}(H_{t_i+1}^2) - \mathbb{E}(H_{t_i+1}^2) \right) + 2\lambda(\mathbb{R}^d) dh \mathbb{E} \left[ \int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr \right] \right].
$$

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

$$
(1 + \gamma h) \mathbb{E} \left| \varepsilon_t(H_{t_i+1}) \right|^2 + Dh^2 + D \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 \right) \\
+ Dh \mathbb{E} \left| Y_{t_i} - \hat{V}_{t_i} \right|^2 + D \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - Z_{t_i}|^2 ds \right) \\
+ D \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\Gamma_s - \hat{\Gamma}_{t_i}|^2 ds \right) + (1 + \gamma h) C 2dh \mathbb{E}(H_{t_i+1}^2) + 2dh \mathbb{E} \left[ \int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr \right] \\
+ (1 + \gamma h) C 2\lambda(\mathbb{R}^d) \mathbb{E}(H_{t_i+1}^2) + 2\lambda(\mathbb{R}^d) Dh \mathbb{E} \left[ \int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr \right] \\
- 2(1 + \gamma h) \mathbb{E} \left| \varepsilon_t(H_{t_i+1}) \right|^2.
$$
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 $\mathbb{E} \left| Y_{t_i} - \hat{Y}_{t_i} \right|^2$ on the left side

\[
(1 - Ch)\mathbb{E} \left| Y_{t_i} - \hat{Y}_{t_i} \right|^2 \\
\leq Ch^2 + C\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds \right) + C\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - \bar{Z}_{t_i}|^2 ds \right) \\
+ C\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\Gamma_s - \bar{\Gamma}_{t_i}|^2 ds \right) + C(1 + Ch)\mathbb{E} \left( H_{i+1}^2 \right) + Ch\mathbb{E} \left( \int_{t_i}^{t_{i+1}} f(\Theta_r)^2 dr \right).
\]

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

\[
\mathbb{E} \left| Y_{t_i} - \hat{Y}_{t_i} \right|^2 \leq Ch^2 + C\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds \right) + C\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - \bar{Z}_{t_i}|^2 ds \right) \\
+ C\mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\Gamma_s - \bar{\Gamma}_{t_i}|^2 ds \right) + C(1 + Ch)\mathbb{E} \left( H_{i+1}^2 \right),
\]

Finally, by recalling that $H_{i+1} = Y_{t_{i+1}} - \hat{U}_{t_{i+1}}(X^\pi_{t_{i+1}})$, we have established (5.1).

**Step 2:** The last term in (5.1),

\[
C(1 + Ch)\mathbb{E} \left| Y_{t_{i+1}} - \hat{U}_{t_{i+1}}(X^\pi_{t_{i+1}}) \right|^2,
\]

was left without a control in previous step. Here in what follows we provide a control on this term. Recall the error terms $\epsilon^Z(h)$ and $\epsilon^\Gamma(h)$ introduced in (4.5). The purpose of this section is to show the following estimate:

**Lemma 5.2** There exists a constant $C > 0$ such that,

\[
\max_{i \in \{0, \ldots, N-1\}} \mathbb{E} \left| Y_{t_i} - \hat{U}_i(X^\pi_{t_i}) \right|^2 \\
\leq C \left[ N \sum_{i=0}^{N-1} \mathbb{E} \left| \hat{U}_i(X^\pi_{t_i}) - \hat{Y}_{t_i} \right|^2 + h + \epsilon^Z(h) + \epsilon^\Gamma(h) + \mathbb{E} \left| g(X_T) - g(X^\pi_T) \right|^2 \right].
\]

(5.11)

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

**Proof of Lemma 5.2** We have that $(a + b)^2 \geq (1 - h)a^2 + (1 - \frac{1}{h})b^2$ and

\[
\mathbb{E} \left| Y_{t_i} - \hat{Y}_{t_i} \right|^2 = \mathbb{E} \left| (Y_{t_i} - \hat{U}_i(X^\pi_{t_i})) + (\hat{U}_i(X^\pi_{t_i}) - \hat{Y}_{t_i}) \right|^2 \\
\geq (1 - h)\mathbb{E} \left| Y_{t_i} - \hat{U}_i(X^\pi_{t_i}) \right|^2 + \left( 1 - \frac{1}{h} \right)\mathbb{E} \left| \hat{U}_i(X^\pi_{t_i}) - \hat{Y}_{t_i} \right|^2.
\]

(5.12)
Therefore, we have an upper (5.1) and lower bound for $\mathbb{E} \left| Y_{t_i} - \tilde{\mathcal{Y}}_{t_i} \right|^2$. By connecting these bounds,

\[
(1 - h) \mathbb{E} \left| Y_{t_i} - \tilde{\mathcal{Y}}_{t_i} (X_{t_i}^\pi) \right|^2 + \left(1 - \frac{1}{h}\right) \mathbb{E} \left| \tilde{\mathcal{Y}}_{t_i} (X_{t_i}^\pi) - \mathcal{Y}_{t_i} \right|^2 \\
\leq Ch^2 + C \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_{t_i}|^2 ds \right) + C \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - Z_{t_i}|^2 ds \right) \\
+ C \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\Gamma_s - \Gamma_{t_i}|^2 ds \right) + Ch \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |f(\Theta_s)|^2 ds \right) + C(1 + Ch) \mathbb{E} \left( H_{i+1}^2 \right).
\]

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

\[
\mathbb{E} \left| Y_{t_i} - \tilde{\mathcal{Y}}_{t_i} (X_{t_i}^\pi) \right|^2 \\
\leq N C \sum_{k=i}^{N-1} \mathbb{E} \left| \tilde{\mathcal{Y}}_{k} (X_{t_k}^\pi) - \mathcal{Y}_{t_k} \right|^2 + C(N - i) h^2 \\
+ C \sum_{k=i}^{N-1} \left[ \mathbb{E} \left( \int_{t_k}^{t_{k+1}} |Y_s - Y_{t_k}|^2 ds \right) + \mathbb{E} \left( \int_{t_k}^{t_{k+1}} |Z_s - Z_{t_k}|^2 ds \right) \\
+ \mathbb{E} \left( \int_{t_k}^{t_{k+1}} |\Gamma_s - \Gamma_{t_k}|^2 ds \right) \right] \\
+ Ch \sum_{k=i}^{N-1} \mathbb{E} \left( \int_{t_k}^{t_{k+1}} |f(\Theta_s)|^2 ds \right) + C \mathbb{E} \left| Y_{t_N} - g(X_{t_N}^\pi) \right|^2
\]

\[
\leq N C \sum_{k=0}^{N-1} \mathbb{E} \left| \tilde{\mathcal{Y}}_{k} (X_{t_k}^\pi) - \mathcal{Y}_{t_k} \right|^2 + CN h^2 \\
+ C \sum_{k=0}^{N-1} \left[ \mathbb{E} \left( \int_{t_k}^{t_{k+1}} |Y_s - Y_{t_k}|^2 ds \right) + \mathbb{E} \left( \int_{t_k}^{t_{k+1}} |Z_s - Z_{t_k}|^2 ds \right) \\
+ \mathbb{E} \left( \int_{t_k}^{t_{k+1}} |\Gamma_s - \Gamma_{t_k}|^2 ds \right) \right] \\
+ Ch \sum_{k=0}^{N-1} \mathbb{E} \left( \int_{t_k}^{t_{k+1}} |f(\Theta_s)|^2 ds \right) + C \mathbb{E} \left| Y_{t_N} - g(X_{t_N}^\pi) \right|^2.
\]
Applying maximum on $i \in \{0, \ldots, N-1\}$, recalling (4.5) and the bounds from Lemmas (4.7) and (4.1),
\[
\max_{i \in \{0, \ldots, N-1\}} \mathbb{E} \left| Y_{t_i} - \widehat{U}_i(X^\pi_{t_i}) \right|^2 \\
\leq C \left[ N \sum_{i=0}^{N-1} \mathbb{E} \left| \widehat{U}_i(X^\pi_{t_i}) - \widehat{V}_i \right|^2 + O(h) + \varepsilon^2(h) + \varepsilon^\Gamma(h) + \mathbb{E} \left| g(X_T) - g(X^\pi_T) \right|^2 \right].
\]
This is nothing that (5.11).

**Remark 5.2** 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 $N^\delta$, which is better. However, at some point of the proof the value $\delta = 1$ is necessary.

**Step 3:** Estimate (5.11) contains some uncontrolled terms on its RHS. Here the purpose is to bound the term
\[
\sum_{i=0}^{N-1} \mathbb{E} \left| \widehat{U}_i(X^\pi_{t_i}) - \widehat{V}_i \right|^2,
\]
in terms of more tractable terms. In this step we will prove

**Lemma 5.3** There exists $C > 0$ such that,
\[
\max_{i \in \{0, \ldots, N-1\}} \mathbb{E} \left| Y_{t_i} - \widehat{U}_i(X^\pi_{t_i}) \right|^2 \\
\leq C \left[ h + \sum_{i=0}^{N-1} (N\varepsilon^\nu_i + \varepsilon^z_i + \varepsilon^\pi_i) + \varepsilon^2(h) + \varepsilon^\Gamma(h) + \mathbb{E} \left| g(X_T) - g(X^\pi_T) \right|^2 \right],
\]
with $\varepsilon^\nu_i$, $\varepsilon^z_i$ and $\varepsilon^\pi_i$ defined in (4.11).

In what follows, we will prove 5.13.

**Proof** Fix $i \in \{0, \ldots, N-1\}$. Recall the martingale $(N_t)_{t \in [t_i, t_{i+1}]}$ and take $t = t_{i+1}$,
\[
\widehat{U}_{i+1}(X^\pi_{t_{i+1}}) = \mathbb{E}_i \left( \widehat{U}_{i+1}(X^\pi_{t_{i+1}}) \right) + \int_{t_i}^{t_{i+1}} \hat{Z}_s \cdot dW_s + \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \hat{U}_s(y)\bar{\mu}(ds, dy).
\]
Now we replace the definition of $\widehat{V}_i$,
\[
\widehat{U}_{i+1}(X^\pi_{t_{i+1}}) = \widehat{V}_{i+1}(X^\pi_{t_{i+1}}) - \int_{t_i}^{t_{i+1}} \hat{Z}_s \cdot dW_s + \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \hat{U}_s(y)\bar{\mu}(ds, dy).
\]
In what follows recall the value of $F$ in the loss function $L_i(\theta)$ (3.9) evaluated at the point
\[
(t_i, X^\pi_{t_i}, \mathcal{U}_i(X^\pi_{t_i}; \theta), Z_i(X^\pi_{t_i}; \theta), G_i(X^\pi_{t_i}; \cdot; \theta), h, \Delta W_i),
\]
and that $(\mathcal{G})_i(X_{t_i}; \theta)$ is given in (3.8):
\[
F(t_i, X^\pi_{t_i}, \mathcal{U}_i(X^\pi_{t_i}; \theta), Z_i(X^\pi_{t_i}; \theta), G_i(X^\pi_{t_i}; \cdot; \theta), h, \Delta W_i) \\
= \mathcal{U}_i(X^\pi_{t_i}; \theta) - hf(t_i, X^\pi_{t_i}, \mathcal{U}_i(X^\pi_{t_i}; \theta), Z_i(X^\pi_{t_i}; \theta), (\mathcal{G})_i(X_{t_i}; \theta)) + Z_i(X^\pi_{t_i}; \theta) \cdot \Delta W_i \\
+ \int_{\mathbb{R}^d} G_i(X^\pi_{t_i}; y; \theta)\bar{\mu}\left((t_i, t_{i+1}], dy\right).
\]
Now fix a parameter \( \theta \) and replace (5.14) on \( L_i(\theta) \):

\[
\mathbb{E} \left| \tilde{U}_{t+1}(X^\pi_{t+1}) - F(t_i, X^\pi_{t_i}, U_i(X^\pi_{t_i}; \theta), Z_i(X^\pi_{t_i}; \theta), G_i(X^\pi_{t_i}; \cdot; \theta), \Delta t_i, \Delta W_i) \right|^2 \\
= \mathbb{E} \left| \tilde{V}_{t_i} - f(t_i, X^\pi_{t_i}, \tilde{V}_{t_i}, \tilde{Z}_{t_i}, \tilde{\Gamma}_{t_i})h + \int_{t_i}^{t_{i+1}} \tilde{Z}_s \cdot dW_s \\
+ \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \tilde{U}_s(y)\overline{\mu}(ds, dy) - U_i(X^\pi_{t_i}; \theta) \\
+ h f(t_i, X^\pi_{t_i}, U_i(X^\pi_{t_i}; \theta), Z_i(X^\pi_{t_i}; \theta), \langle G \rangle_i(X^\pi_{t_i}; \theta)) - Z_i(X^\pi_{t_i}; \theta) \cdot \Delta W_i \\
- \int_{\mathbb{R}^d} G_i(X^\pi_{t_i}, y; \theta)\overline{\mu}(ds, dy) \right|^2 \\
= \mathbb{E} \left| \tilde{V}_{t_i} - U_i(X^\pi_{t_i}; \theta) \\
+ h \left( f(t_i, X^\pi_{t_i}, U_i(X^\pi_{t_i}; \theta), Z_i(X^\pi_{t_i}; \theta), \langle G \rangle_i(X^\pi_{t_i}; \theta)) - f(t_i, X^\pi_{t_i}, \tilde{V}_{t_i}, \tilde{Z}_{t_i}, \tilde{\Gamma}_{t_i}) \right) \right|^2 \\
+ \left[ \int_{t_i}^{t_{i+1}} \tilde{Z}_s \cdot dW_s - \int_{t_i}^{t_{i+1}} Z_i(X^\pi_{t_i}; \theta) \cdot dW_s \\
+ \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \tilde{U}(s, y)\overline{\mu}(ds, dy) - \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} G_i(X^\pi_{t_i}, y; \theta)\overline{\mu}(ds, dy) \right]^2 \\
= \mathbb{E} \left| a + b \right|^2.
\]

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

\[
\mathbb{E}(b^2) = \mathbb{E} \left( \int_{t_i}^{t_{i+1}} [\tilde{Z}_s - Z_i(X^\pi_{t_i}; \theta)]dW_s \right)^2 \\
+ \mathbb{E} \left( \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} [\tilde{U}(s, y) - G_i(X^\pi_{t_i}, y; \theta)]\overline{\mu}(ds, dy) \right)^2 ;
\]

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

\[
L_i(\theta) = \mathbb{E} \left( \tilde{V}_{t_i} - U_i(X^\pi_{t_i}; \theta) + h \left[ f(t_i, X^\pi_{t_i}, U_i(X^\pi_{t_i}; \theta), Z_i(X^\pi_{t_i}; \theta), \langle G \rangle_i(X^\pi_{t_i}; \theta)) - f(t_i, X^\pi_{t_i}, \tilde{V}_{t_i}, \tilde{Z}_{t_i}, \tilde{\Gamma}_{t_i}) \right] \right)^2 \\
+ \mathbb{E} \left( \int_{t_i}^{t_{i+1}} [\tilde{Z}_s - Z_i(X^\pi_{t_i}; \theta)]dW_s \right)^2 + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} [\tilde{U}(s, y) - G_i(X^\pi_{t_i}, y; \theta)]\overline{\mu}(ds, dy) \right)^2 .
\]

By the same arguments on equations (5.6) and (5.7),

\[
c_0 = \mathbb{E} \left( \int_{t_i}^{t_{i+1}} [\tilde{Z}_s - Z_i]^2 ds \right) + h^2 \mathbb{E} \left[ Z_{t_i} - Z_i(X^\pi_{t_i}; \theta) \right]^2 \\
+ \mathbb{E} \left( \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \tilde{U}(y) - \tilde{U}_{t_i}(y)^2 \lambda(dy) ds \right) \\
+ h \mathbb{E} \left( \int_{\mathbb{R}^d} (\tilde{U}_{t_i}(y) - G_i(X^\pi_{t_i}, y; \theta))^2 \lambda(dy) \right).
\]
With this decomposition of \( L_i(\theta) \), for optimization reasons, we can ignore the part that does not depend on the optimization parameter \( \theta \). Let

\[
\hat{L}_i(\theta) = \mathbb{E} \left( \hat{V}_i - \mathcal{U}_i(X^\pi_i; \theta) + h \left[ f(t_i, X^\pi_i, \mathcal{U}_i(X^\pi_i; \theta), Z_i(X^\pi_i; \theta), (G)_i(X^\pi_i; \theta)) - f(t_i, X^\pi_i, \hat{V}_i, \hat{Z}_i, \hat{\Gamma}_i) \right]^2 \right)
\]

\[+ h \mathbb{E} \left[ |Z_i(X^\pi_i; \theta) - \hat{Z}_i|^2 \right] + h \mathbb{E} \left( \int_{\mathbb{R}^d} (\mathcal{U}_i(y) - G_i(X^\pi_i, y; \theta))^2 \lambda(dy) \right).
\]

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

\[
\mathbb{E} \left( \hat{V}_i - \mathcal{U}_i(X^\pi_i; \theta) + h \left[ f(t_i, X^\pi_i, \mathcal{U}_i(X^\pi_i; \theta), Z_i(X^\pi_i; \theta), (G)_i(X^\pi_i; \theta)) - f(t_i, X^\pi_i, \hat{V}_i, \hat{Z}_i, \hat{\Gamma}_i) \right]^2 \right)
\]

\[\leq (1 + \gamma h) \mathbb{E} \left[ |\hat{V}_i - \mathcal{U}_i(X^\pi_i; \theta)|^2 \right] + \left( 1 + \frac{1}{\gamma h} \right) h^2 K^2 \mathbb{E} \left( |\hat{V}_i - \mathcal{U}_i(X^\pi_i; \theta)|^2 + |Z_i(X^\pi_i; \theta) - \hat{Z}_i|^2 + |(G)_i(X^\pi_i; \theta) - \hat{\Gamma}_i|^2 \right)
\]

\[\leq (1 + Ch) \mathbb{E} \left[ |\hat{V}_i - \mathcal{U}_i(X^\pi_i; \theta)|^2 \right] + C h \mathbb{E} \left[ |Z_i(X^\pi_i; \theta) - \hat{Z}_i|^2 + \mathbb{E} \left( \|\mathcal{U}_i(\cdot) - G_i(X^\pi_i, \cdot; \theta)\|^2_{L^2(\lambda)} \right) \right].
\]

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

\[
\hat{L}(\theta) \leq C \mathbb{E} \left[ |\hat{V}_i - \mathcal{U}_i(X^\pi_i; \theta)|^2 \right] + h \left( \mathbb{E} \left[ |Z_i(X^\pi_i; \theta) - \hat{Z}_i|^2 \right] \right)
\]

\[+ h \mathbb{E} \left( \|\mathcal{U}_i(\cdot) - G_i(X^\pi_i, \cdot; \theta)\|^2_{L^2(\lambda)} \right).
\]

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

\[
\mathbb{E} \left( \hat{V}_i - \mathcal{U}_i(X^\pi_i; \theta) + h \left[ f(t_i, X^\pi_i, \mathcal{U}_i(X^\pi_i; \theta), Z_i(X^\pi_i; \theta), (G)_i(X^\pi_i; \theta)) - f(t_i, X^\pi_i, \hat{V}_i, \hat{Z}_i, \hat{\Gamma}_i) \right]^2 \right)
\]

\[\geq (1 - Ch) \mathbb{E} \left[ |\hat{V}_i - \mathcal{U}_i(X^\pi_i; \theta)|^2 \right] - \frac{h}{2} \left( \mathbb{E} \left[ |Z_i(X^\pi_i; \theta) - \hat{Z}_i|^2 \right] + \mathbb{E} \left( |(G)_i(X^\pi_i; \theta) - \hat{\Gamma}_i|^2 \right) \right);
\]

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

\[
\hat{L}(\theta) \geq (1 - Ch) \mathbb{E} \left[ |\hat{V}_i - \mathcal{U}_i(X^\pi_i; \theta)|^2 \right]
\]

\[- \frac{h}{2} \left( \mathbb{E} \left[ |Z_i(X^\pi_i; \theta) - \hat{Z}_i|^2 \right] + \mathbb{E} \left( \int_{\mathbb{R}^d} (\mathcal{U}_i(y) - G_i(X^\pi_i, y; \theta))^2 \lambda(dy) \right) \right].
\]
Connecting this bounds using that $\hat{L}(\theta^*) \leq \hat{L}(\theta)$ yields that $\forall \theta$,

\[
(1 - Ch) \mathbb{E} \left| \bar{\nabla}_t \mathcal{U}_t (X_t^\pi; \theta^*) \right|^2 + \frac{h}{2} \mathbb{E} \left| \bar{Z}_t - Z_t (X_t, \theta^*) \right|^2 \\
+ \frac{h}{2} \mathbb{E} \left( \int_{\mathbb{R}^d} (\bar{U}_t (y) - G_t (X_t^\pi, y; \theta^*))^2 \lambda (dy) \right) \\
\leq C \mathbb{E} \left| \bar{\nabla}_t \mathcal{U}_t (X_t^\pi; \theta) \right|^2 \\
+ h \left( \mathbb{E} \left| \bar{Z}_t - Z_t (X_t, \theta) \right|^2 \right) + h \mathbb{E} \left( \left\| \bar{U}_t (\cdot) - G_t (X_t^\pi, \cdot; \theta) \right\|_{L^2 (\lambda)}^2 \right).
\]

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

\[
\mathbb{E} \left| \bar{\nabla}_t - \bar{\nabla}_t (X_t^\pi) \right|^2 + \frac{h}{2} \mathbb{E} \left| \bar{Z}_t - Z_t (X_t) \right|^2 + \frac{h}{2} \mathbb{E} \left( \int_{\mathbb{R}^d} (\bar{U}_t (y) - \bar{G}_t (X_t^\pi, y))^2 \lambda (dy) \right) \\
\leq C \left( \epsilon_i^2 + h \epsilon_i^z + h \epsilon_i^y \right).
\]

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

\[
\max_{i \in \{0, \ldots, N-1\}} \mathbb{E} \left| Y_t - \bar{\nabla}_t (X_t^\pi) \right|^2 \leq C \left[ h + \sum_{i=0}^{N-1} (N \epsilon_i^y + \epsilon_i^z + \epsilon_i^y) + \sum_{i=0}^{N-1} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_t|^2 ds \right) \\
+ \epsilon^Z (h) + \epsilon^T (h) + \mathbb{E} \left| g (X_T) - g (X_T^\pi) \right|^2 \right].
\]

Finally, using Proposition 4.1, one ends the proof of (5.13).

\[\square\]

**Step 4:** We are going to show some bounds for the terms involving the $\Gamma$ and $U$ components, the same bounds holds for $Z$ component and are shown in [31]. By using (5.10) on (5.7),

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

this implies, after using (4.5) and (4.7),

\[
\mathbb{E} \left( \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \left| \Gamma_t - \bar{\Gamma}_t \right|^2 dt \right) \leq \mathbb{E} \left( \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \left| \Gamma_t - \bar{\Gamma}_t \right|^2 dt \right) \\
+ C \sum_{i=0}^{N-1} \left( \mathbb{E} (H_{i+1}^2) - \mathbb{E} \left| \mathbb{E}_i (H_{i+1}) \right|^2 \right) + Ch \\
= \epsilon^T (h) + Ch + C \sum_{i=0}^{N-1} \left( \mathbb{E} (H_{i+1}^2) - \mathbb{E} \left| \mathbb{E}_i (H_{i+1}) \right|^2 \right).
\]
From [31] we get the analogous bound for the $Z$ component, therefore, putting this two together yields

$$
\mathbb{E} \left( \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \left( |Z_t - \bar{Z}_{t_j}|^2 + |\Gamma_t - \bar{\Gamma}_{t_j}|^2 \right) dt \right)
\leq \varepsilon^Z(h) + \varepsilon^\gamma(h) + C h + C \sum_{i=0}^{N-1} \left( \mathbb{E}(H_{i+1}^2) - \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2 \right). \tag{5.16}
$$

This tells us that the next mission in this proof is to give a suitable bound for $\mathbb{E}(H_{i+1}^2) - \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2$. Recall from (5.8) that $H_{i+1} = Y_{i+1} - \hat{U}_{i+1}(X_{t_{i+1}}^\pi)$, then

$$
\sum_{i=0}^{N-1} \left( \mathbb{E}(H_{i+1}^2) - \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2 \right) = \sum_{i=0}^{N-1} \mathbb{E}(H_{i+1}^2) - \sum_{i=0}^{N-1} \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2
$$

$$
= \mathbb{E} |Y_{i+1} - \bar{U}_N(X_{t_N})| + \sum_{i=0}^{N-2} \mathbb{E}(H_{i+1}^2) - \sum_{i=0}^{N-1} \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2
$$

$$
\leq \mathbb{E} |Y_{i+1} - \bar{U}_N(X_{t_N})| + \mathbb{E}(H_0^2) + \sum_{i=1}^{N-1} \mathbb{E}(H_i^2)
$$

$$
- \sum_{i=0}^{N-1} \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2
$$

$$
= \mathbb{E} |g(X_T) - g(X_T^\pi)|^2 + \sum_{i=0}^{N-1} \left( \mathbb{E}(H_i^2) - \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2 \right). \tag{5.17}
$$

From (5.12) and (5.4) we have an upper and lower bound on $\mathbb{E} |Y_t - \hat{V}_t|^2$. Indeed, first one has

$$
(1 - h) \mathbb{E} |Y_t - \hat{U}_t(X_{t}^\pi)|^2 \leq \mathbb{E} |Y_t - \hat{V}_t|^2 + \left( \frac{1}{h} - 1 \right) \mathbb{E} |\hat{U}_t(X_{t}^\pi) - \hat{V}_t|^2. \tag{5.18}
$$

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

$$
(1 - h) \mathbb{E} |Y_t - \hat{U}_t(X_{t}^\pi)|^2 \leq \left( \frac{1}{h} - 1 \right) \mathbb{E} |\hat{U}_t(X_{t}^\pi) - \hat{V}_t|^2 + (1 + \gamma h) \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2 + (1 + \gamma h) \frac{C}{\gamma}
$$

$$
\left[ h^2 + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Y_s - Y_t|^2 ds \right) + h \mathbb{E} |Y_n - \hat{V}_n|^2 + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |Z_s - \bar{Z}_n|^2 ds \right) + \mathbb{E} \left( \int_{t_i}^{t_{i+1}} |\Gamma_s - \bar{\Gamma}_n|^2 ds \right) \right].
$$

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

$$
\mathbb{E}(H_i^2) - \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2 \leq \frac{1}{h} \mathbb{E} |\hat{U}_t(X_{t}^\pi) - \hat{V}_t|^2 + \left( \frac{h + \gamma h}{1 - h} \right) \mathbb{E} |\mathbb{E}_i(H_{i+1})|^2 + \frac{C}{\gamma} \frac{(1 + \gamma h)}{(1 - h)} B_i.
$$

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

$$
\frac{C}{\gamma} \frac{(1 + \gamma h)}{(1 - h)} \leq \frac{1}{2} \quad \text{and} \quad \frac{1}{1 - h} \leq \frac{1}{2}.
$$
Hence,
\[ \mathbb{E}(H_i^2) - \mathbb{E}[\mathbb{E}_i(H_i+1)^2] \leq \frac{1}{h} \mathbb{E}[\hat{U}_i(X_i^\pi) - \hat{V}_i]^2 + C h \mathbb{E}[\mathbb{E}_i(H_i+1)^2] + \frac{1}{2} B_i. \]

Finally, note that,
\[ \sum_{i=0}^{N-1} \mathbb{E}[\mathbb{E}_i(H_i+1)]^2 \leq \mathbb{E}[|g(X_T) - g(X_T^\pi)|^2] + N \max_{i=0, \ldots, N-1} \mathbb{E}[|Y_i - \hat{U}_i(X_i^\pi)|^2]. \quad (5.19) \]

**Remark 5.3** Note that in equation (5.19) 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 \( \delta \) improvement mentioned on Remark 5.2 will not be of much help.

Coming back to (5.17),
\[ \sum_{i=0}^{N-1} \left( \mathbb{E}(H_i^2) - \mathbb{E}[\mathbb{E}_i(H_i+1)^2] \right) \leq 2 \mathbb{E}[|g(X_T) - g(X_T^\pi)|^2] + N \sum_{i=0}^{N-1} \mathbb{E}[|\hat{U}_i(X_i^\pi) - \hat{V}_i|^2] \]
\[ + C h N \max_{i=0, \ldots, N-1} \mathbb{E}[|Y_i - \hat{U}_i(X_i^\pi)|^2] + \frac{1}{2} \sum_{i=0}^{N-1} B_i. \]

Therefore, by plugging this bound in (5.16), noting that \(|Y_i - \hat{V}_i|^2 \leq 2|Y_i - \hat{U}_i(X_i^\pi)|^2 + 2|\hat{U}_i(X_i^\pi) - \hat{V}_i|^2, h N = 1, and using Lemma 4.1, we have for some \( C > 0, \)
\[ \mathbb{E}\left(\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} (|Z_t - \tilde{Z}_{ti}|^2 + |\Gamma_t - \tilde{\Gamma}_{ti}|^2) dt\right) \]
\[ \leq C \left[ \mathbb{E}[|g(X_T) - g(X_T^\pi)|^2] + \varepsilon^Z(h) + \varepsilon^\Gamma(h) + h \right. \]
\[ \left. + N \sum_{i=0}^{N-1} \mathbb{E}[|\hat{U}_i(X_i^\pi) - \hat{V}_i|^2] + \max_{i=0, \ldots, N-1} \mathbb{E}[|Y_i - \hat{U}_i(X_i^\pi)|^2] \right]. \]

Now, use (5.15) together with Lemma 5.3 to get
\[ \mathbb{E}\left(\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} (|Z_t - \tilde{Z}_{ti}|^2 + |\Gamma_t - \tilde{\Gamma}_{ti}|^2) dt\right) \]
\[ \leq C \left[ \mathbb{E}[|g(X_T) - g(X_T^\pi)|^2] + \varepsilon^Z(h) + \varepsilon^\Gamma(h) + h \right. \]
\[ \left. + \sum_{i=0}^{N-1} (N \varepsilon_i^w + \varepsilon_i^z + \varepsilon_i^\gamma) \right]. \]
Again, recalling (5.15) using the previous bound and,
\[
\sum_{i=0}^{N-1} E \left( \int_{t_i}^{t_{i+1}} \left[ |Z_t - \hat{Z}_{t_i}(X_{t_i}^{\pi})|^2 + |\Gamma_t - (\hat{G})_{t_i}(X_{t_i}^{\pi})|^2 \right] dt \right)
\]
\[
\leq \sum_{i=0}^{N-1} E \left( \int_{t_i}^{t_{i+1}} \left[ |Z_t - \hat{Z}_{t_i}(\cdot)|^2 + |\Gamma_t - (\hat{G})_{t_i}(\cdot)|^2 \right] dt \right)
\]
\[
+ \sum_{i=0}^{N-1} h E \left( \left[ \left| \hat{Z}_{t_i} - \hat{Z}_i(X_{t_i}^{\pi}) \right|^2 + \left\| \hat{G}_{t_i} (\cdot) - \hat{G}(X_{t_i}^{\pi}, \cdot) \right\|_{L^2(\lambda)}^2 \right] dt \right),
\]
we conclude that there exist \( C > 0 \), independent of the partition, such that for \( h \) sufficiently small,
\[
\max_{i=0, \ldots, N-1} \left| Y_{t_i} - \hat{U}_{t_i}(X_{t_i}^{\pi}) \right|^2 + \sum_{i=0}^{N-1} E \left( \int_{t_i}^{t_{i+1}} \left[ |Z_t - \hat{Z}_{t_i}(X_{t_i}^{\pi})|^2 + |\Gamma_t - (\hat{G})_{t_i}(X_{t_i}^{\pi})|^2 \right] dt \right)
\]
\[
\leq C \left[ h + \sum_{i=0}^{N-1} \left( N \xi_i^{v} + \xi_i^{z} + \xi_i^{Y} \right) + \varepsilon Z(h) + \varepsilon (h) \right] + E \left[ g(X_T) - g(X_{T}^{\pi}) \right]^2 \right].
\]

Thus it has been demonstrated.

We state some remarks from the proof.

\textbf{Remark 5.4} Note that the terms \( \xi_i^{v}, \xi_i^{z} \) and \( \xi_i^{Y} \), can be made arbitrarily small, in view of Lemma 2.1. The challenge here, and in almost every DL algorithm, is that we do not know how many units per layer, i.e., how large \( \kappa \) 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.

\textbf{Remark 5.5} The main difficulty of the adaptation of the proof given in [31], 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
\[
\int_{\mathbb{R}^d} \left[ u(t, X_{t_i}^{\pi} + \beta(X_{t_i}^{\pi}, y)) - u(t, X_{t_i}^{\pi}) \right] \lambda(dy),
\]
this seems intuitive because this will lead our third NN to approximate the nonlocal part of the PIDE and, therefore, receive one parameter: \( X_{t_i}^{\pi} \). But, we also need to approximate or been able to calculate the stochastic integral
\[
\int_{\mathbb{R}^d} \left[ u(t, X_{t_i}^{\pi} + \beta(X_{t_i}^{\pi}, y)) - u(t, X_{t_i}^{\pi}) \right] \hat{\mu} ((t_i, t_{i+1}], dy),
\]
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.

\textbf{Remark 5.6} The non local part of the PIDE (1.2) 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 (1.2). This addition results in the natural definition of analogous objects from [31] such as the \( \Gamma, \hat{\Gamma} \) components for the nonlocal case.
Remark 5.7 The result of the theorem states that the better we can approximate \( v_t, z_t, \gamma_t \) by NN architectures, the better we can approximate \( (Y_t, Z_t, \Gamma_t) \) by \( (\hat{U}_t(X^\pi_t), \hat{Z}_t(X^\pi_t), (\hat{G}_t)(X^\pi_t)) \).

Remark 5.8 Because of the finiteness of the measure \( \lambda \), the case of the Fractional Laplacian mentioned in the introduction is not contained in Theorem 5.1. We hope to extend our results to this case in a forthcoming result.

5.1 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 [31] 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 \( U \in L^1_{\mu}(\mathbb{R}^d) \) the definition of stochastic integral with respect to \( \mu \) ([4, 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 [4, Thm 2.3.10]). And therefore,

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

For simplicity assume that \( \lambda \) 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 \( \hat{U}_{i+1} \) 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 (3.9), 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 \( \{x_k^i : k \in I\}, \{x_k^{i+1} : k \in I\}, \{w_k : k \in I\} \) of \( X^\pi_t, X_{t+1}^\pi \) and \( \Delta W_i \) respectively. Then,

\[
L_i(\theta) \approx \frac{1}{M} \sum_{k \in I} \left( \frac{\hat{U}_{i+1}(x_k^{i+1}) - F(t_i, x_k^i, U_t(\cdot; \theta), Z_t(\cdot; \theta), G_t(\cdot; \theta), h, w_i)}{2} \right)^2.
\]

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 [14, 34]. Recall that \( F \) needs two different integrals of \( G_t(x_k^i, \cdot; \theta) \), to approximate these values let \( L \in \mathbb{N} \) and \( J = \{1, ..., L\} \) and consider, for every \( k \in I \), simulations \( \{y_k^l : l \in J\} \) 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,

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

\[
\int_{\mathbb{R}^d} G_i(x_k^i, y; \theta) \bar{\mu} ((t_i, t_i+1], dy) = \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} G_i(x_k^i, y; \theta) \mu(dt, dy) \]

\[
- \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} G_i(x_k^i, y; \theta) dt \lambda(dy)
\]

\[
\approx \sum_{t_i \leq s < t_i+1} G_i(x_k^i, \Delta P_s; \theta) \mathbb{I}_{\mathbb{R}^d}(\Delta P_s) - \frac{h}{L} \sum_{l \in J} G_i(x_k^i, y_l^k; \theta).
\]

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 \( \theta^* \) and define

\[
(\hat{U}_i, \hat{Z}_i, \hat{G}_i) = (U_i(\cdot; \theta^*), Z_i(\cdot; \theta^*), G_i(\cdot, \cdot; \theta^*)).
\]

**Remark 5.9** The nonlocal term in equation (1.2) 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 \( \lambda \) 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.

**Acknowledgements** We would like to thank Lukas Gonon and Christoph Schwab for let us know of their recent preprints [26, 27].

**Funding** Open Access funding enabled and organized by Projekt DEAL. Partial financial support was received from Chilean research grants FONDECYT 1191412, and CMM Conicyt PIA AFB170001.

**Data Availability Statement** Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

**Declarations**

**Conflicts of interest** The author declare he has no financial interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

**References**

1. Grégoire, A.: Numerical Analysis and Optimization An introduction to mathematical modeling and numerical simulation. Oxford University Press; Illustrated edition (July 19: 472 pages. ISBN- 10, 9780805839852 (2007)
2. Md, Z.A., Tarek, M.T., Chris, Y., Stefan, W., Paheding, S., Mst, S.N., Mahmudul, H., Brian, C.V.E., Abdul, A.S.A., Vijayan, K.A.: A state-of-the-art survey on deep learning theory and architectures. Electronics 8(3), 292 (2019). https://doi.org/10.3390/electronics8030292

3. Ali, M.A., Hiam, A., Isam, A.Q., Amin, A., Wafaa, A.S.: Brain tumor classification using deep learning technique—a comparison between cropped, uncropped, and segmented lesion images with different sizes. Int. J. Adv. Trends Comput. Sci. Eng. 8(6), 2 (2019)

4. David, A.: Lévy processes and stohastic calculus, Cambridge Studies In Advanced Mathematics, 2nd Edition, (April 1, 2009). ISBN-10: 0521738652

5. Barles, G., Buckdahn, R., Pardoux, E.: Backward stochastic differential equations and integral-partial differential equations. Stochas. Stocha. Rep. 60, 57–83 (1996)

6. Guy, B., Olivier, L., Erwin, T.: Lipschitz regularity for integro-differential equations with coercive hamiltonians and applications to large time behavior. Nonlinearity, Volume 30, Number 2 (2017), arXiv:1602.07806 [math.AP]

7. Dalya, B.: Machine learning in astronomy: a practical overview, arXiv:1904.07248v1 [astro-ph.IM] 15 Apr 2019

8. Christian, B., Fabian, H., Martin, H., Arnulf, J., Thomas, K.: Overcoming the curse of dimensionality in the numerical approximation of Allen-Cahn partial differential equations via truncated full-history recursive multilevel Picard approximations, Accepted in J. Numer. Math. arXiv:1907.06729 [math.NA], 2019

9. Christian, B., Weinan, E., Arnulf, J.: Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations, J. Nonlinear Sci. 29 (2019), 1563–1619, arXiv:1709.05963v1 [math.NA], 2017

10. Benson, D.A., Wheatcraft, S.W., Meerschaert, M.M.: Application of a fractional advection-dispersion equation. Water Resour. Res. 36(6), 1403–1412 (2000)

11. Isabeau, B., Giulio, G., Erwin, T.: Fractional truncated laplacians: representation formula. Fundam. Solut. Appl. arXiv:2010.02707 [math.AP], 2020

12. Bruno, B., Romuald, E.: Discrete time approximation of decoupled Forward-Backward SDE with jumps. Stochastic Processes and their Applications, Elsevier, 2008, 118 (1), pp. 53–75. ffhal00015486

13. Dimitri, B.: Machine and deep learning applications in particle physics. Int. J. Modern Phys. A 34(35):1930019 https://doi.org/10.1142/S0217751X19300199, arXiv:1912.08245v1 [physics.data-an]

14. Buckwar, E., Riedler, M.G.: Runge-Kutta methods for jump-diffusion differential equation. J. Comput. Appl. Math. 236, 1155–1182 (2011)

15. Luis, C., Luis, S.: An extension problem related to the fractional Laplacian. Comm. PDE Vol. 32, 2007 Issue 8 pp. 1245–1260

16. Carr, P., Geman, H., Madan, D.B., Yor, M.: The fine structure of asset returns: An empirical investigation. J. Bus. 75, 305–332 (2002)

17. Rama, C., Peter, T.: Financial modelling with jump processes. Chapman and Hall/CRC; 1st edition (December 30, 2003). ISBN-10: 1584884134, 552 pp

18. Gonzalo, D., Erwin, T.: The nonlocal inverse problem of Donsker And Varadhan, arXiv:2011.13295 [math.AP], 2020

19. D'Elia, M., Qiang, D., Glusa, C., Gunzburger, M., Xiaochnuan, T., Zhi, Z.: Numerical methods for nonlocal and fractional models. Acta Numer. 2, 1–124 (2020)

20. Łukasz, D.: Backward Stochastic differential equations with jumps and their actuarial and financial applications. EEA series, Springer-Verlag London, 2013. https://doi.org/10.1007/978-1-4471-5331-3, 288+X pp

21. Eleonora, D.N., Giampiero, P., Enrico, V.: Hitchhiker’s guide to the fractional Sobolev spaces. Bull. Sci. Math. 136(5), 521–573 (2012)

22. Di Nunno, G., Proske, B.,O.F.: Malliavin Calculus for Levy Processes with Applications to Finance. Springer-Verlag, Berlin Heidelberg (2009). https://doi.org/10.1007/978-3-540-78572-9, XIV+418 pp

23. Qiang, D., Xiaochnuan, T., Stability of nonlocal dirichlet integrals and implications for peridynamic correspondence material modeling, SIAM J. Appl. Math. (2018) Vol. 78, No. 3, pp. 1536–1552, arXiv:1710.05119 [physics.comp-ph]

24. Dennis, E., Philipp, G., Arnulf, J., Christoph, S.: DNN Expression Rate Analysis of High-dimensional PDEs: Application to Option Pricing, Tech. Report 2018-33. Seminar for Applied Mathematics, ETH Zürich, Switzerland, 2018

25. Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. Multisc. Model. Simul. 7, 1005–1028 (2008)

26. Lukas, G., Christoph, S.: Deep ReLU neural network approximation for stochastic differential equations with jumps, arXiv:2102.11707 (2021)
27. Lukas, G., Christoph, S.: Deep ReLU network expression rates for option prices in high-dimensional, exponential Lévy models, arXiv:2101.11897 (2021)
28. Philipp, G., Fabian, H., Arnulf, J., Philippe, von W.: A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations, To appear in Mem. Amer. Math. Soc.; arXiv:1809.02362 (2018), 124 pages
29. Han, J., Jentzen, A.: Solving high-dimensional partial differential equations using deep learning. Proc. Natl. Acad. Sci. 115, 8505–8510 (2018)
30. Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural Netw. 4, 251–257 (1991)
31. Hure, C., Pham, H., Warin, X.: Deep backward schemes for high-dimensional nonlinear PDE’s. Math. Comp. 89, 1547–1579 (2020)
32. Martin, H., Arnulf, J., Thomas, K., Tuan, A.N., Philippe, von W.: Overcoming the curse of dimensionality in the numerical approximation of semilinear parabolic partial differential equations, arXiv:1807.01212 [math.PR], 2018. Accepted in Proc. Roy. Soc. A
33. Benjamin, J., Sylvie, M., Wojbor, A.W.: Nonlinear SDEs driven by Lévy processes and related PDEs. ALEA Lat. Am. J. Probab. Math. Stat. 4 (2008), 1–29. arXiv:0707.2723, 2007
34. Arturo, K.-H., Peter, T.: Jump-adapted discretization schemes for Lévy-driven SDEs, Stochastic Processes and their Applications, Volume 120, Issue 11, 2010, Pages 2258-2285, ISSN 0304-4149, https://doi.org/10.1016/j.spa.2010.07.001
35. Antoine, L., Ernesto, M., Soledad, T.: Numerical approximation of backward stochastic differential equations with jumps, 2007. fiinria-00357992v2
36. Moshe, L.I., Vladimir, Y.L., Allan, P., Shimon, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Netw. 6, 861–867 (1993)
37. Geert, L., Thijs, K., Babak, E.B., Arnaud, A.A.S., Francesco, C., Mohsen, G., Jeroen, A.W.M., van der Laak, Bram, van G., Clara, I.S.: A survey on deep learning in medical image analysis. Med. Image Anal. 42, Pages 60–88 (2017), arXiv:1702.05747v2 [cs.CV] 4 Jun 2017
38. Martin, M., Andrew, M.N., Hendrick, W.H.: Neural network solutions to differential equations in non-convex domains: solving the electric field in the slit-well microfluidic device. Phys. Rev. Research 2, 033110 – Published 21 July 2020. arXiv:2004.12235v1 [physics.comp-ph], 2020
39. Mahabal, A., Sheth, K., Gieseke, F., Pai, A., Djorgovski, S. G., Drake, A. J., Graham, M. J., CSS/CRTS/PTF Teams: Deep-learnt classification of light curves, 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, 2017, pp. 1-8, https://doi.org/10.1109/SSCI.2017.8280984. arXiv:1709.06257v1 [astro-ph.IM]
40. Kevin, M., Kavita, B., Noah, S.: StreetStyle: Exploring world-wide clothing styles from millions of photos, arXiv:1706.01869v1 [cs.CV] 6 Jun 2017
41. Warreb, S.M., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133 (1943)
42. Philip, E.: Propter. Stochastic integration and differential equations. Springer, Berlin (2004)
43. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 6 (1958)
44. Justin, S., Konstantinos, S.: DGM: A deep learning algorithm for solving partial differential equations. J. Comput. Phys. 375, 15 December 2018, Pages 1339–1364, arXiv:1708.07469 [q-fin.MF], 2017
45. Pablo, R.S.: User’s guide to the fractional Laplacian and the method of semigroups, in: Fractional Differential Equations, Walter de Gruyter GmbH & Co KG, pp. 235–266, arXiv:1808.05334v2 [cond-mat.dis-nn], 2018
46. Tadmor, E., Tan, C.: Critical thresholds in flocking hydrodynamics with non-local alignment. Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci. 372, 20130401 (2014)
47. Giacomo, T., Guglielmo, M., Juan, C., Matthias, T., Roger, M., Giuseppe, C.: Neural-network quantum state tomography for many-body systems. Nat. Phys. 14, pages 447–450 (2018), arXiv:1703.05334v2 [cond-mat.dis-nn]
48. Haohan, W., Bhiksha, R.: On the origin of deep learning. arXiv:1702.07800v4 [cs.LG] 3 Mar 2017
49. Zhang, J.: A numerical scheme For BSDES. Ann. Appl. Prob. 14(1), 459–488 (2004)
50. Zhang, X.: Stochastic functional differential equations driven by levy processes and quasi-linear partial, integro-differential equation. Ann. Appl. Probab. 22(6), 2505–2538 (2012). arXiv:1106.3601 [math.PR]