Advances in Continuous and Discrete Models

RESEARCH Open Access

Universal approximation property of a continuous neural network based on a nonlinear diffusion equation

Hirotada Honda[1*]

*Correspondence:
honda.hirotada@iniad.org
[1]Faculty of Information and Networking for Innovation and Design, Toyo University, Akabanedai 1-7-11, Kita-Ku, Tokyo, Japan

Abstract

Recently, differential equation-based neural networks have been actively studied. This paper discusses the universal approximation property of a neural network that is based on a nonlinear partial differential equation (PDE) of the parabolic type.

Based on the assumption that the activation function is non-polynomial and Lipschitz continuous, and applying the theory of the difference method, we show that an arbitrary continuous function on any compact set can be approximated using the output of the network with arbitrary precision. Additionally, we present an estimate of the order of accuracy with respect to Δt and Δx.

Mathematics Subject Classification: 35Q93; 49J20; 68T07

Keywords: Universal approximation theorem; Neural ODE; Galerkin approximation; Padé approximation; Difference scheme

1 Introduction

Recently, neural networks have been applied in numerous fields, both in social and natural sciences. However, their performance remains a topic of active research. Since Rosenblatt's work [60], neural networks have been studied extensively. In fact, the set of functions realized by neural network models has been under discussion for some time.

Surprisingly, the transform mapping theorem, which is similar to the universal approximation property, was derived in an early research by Kolmogorov [41] and its simplified proof was provided by Sprecher [70]. However, the neural networks they considered differed slightly from existing conventional implementations. Later, in the 1980s, several studies were conducted on the universal approximation property of neural networks. On the one hand, these results greatly encouraged and facilitated research in neural networks. On the other hand, they found the universal approximation property of neural network models to be closely related with (almost simultaneous) controllability in the theory of optimal control. However, there are some differences between the two. When discussing the universal approximation property of a neural network, these works typically include the effect of the output layer, whose activation function may differ from that of the hidden layer. Arguments concerning these areas are introduced and discussed in detail in the

next section. Moreover, some recent studies have considered neural networks from the perspective of optimal transport [67, 68, 78]. These arguments have led to the application of the dynamical system theory to neural networks.

For example, E [83] regarded a neural network as a method of estimating the parameters of a dynamical system. In particular, he formulated ResNet [29] as an Euler scheme for an ordinary differential equation (ODE) and discussed its stability in the forward direction. They deduced certain conditions under which forward propagation operates stably in the sense that gradient explosion and vanishing problems do not occur when the eigenvalues of the system are considered. Additionally, they highlighted a close relationship (or even equivalence) between the adjoint equation and backpropagation and introduced a regularity method.

This dynamical systems-based approach toward neural networks became more popular after a study by Chen et al., which provided a framework for representing a neural network with an ODE solver. This framework was referred to as the neural ODE [10].

Thereafter, neural ODEs began to be widely used and implemented [10].

Meanwhile, some methods have been proposed based on ODEs and partial differential equations (PDEs) [28, 31]. Han and Li [28] formulated a neural network using an ODE, considering a cost function optimized using the Hamilton–Jacobi–Bellman (HJB) equation. In our previous study [31], we proposed a framework for a neural network in which we considered the initial-boundary value problem for PDEs.

A maximum principle-based approach was also provided in [44]. Notably, some recent works have actively discussed the application of differential equations to graph neural networks (GNNs) (see, for instance, [9]), along with the "expressive power" and "stability" of GNNs. Oono and Suzuki [56] discussed that the expressive power of a GNN decreases when the it has an excessively large numbers of layers. They also proposed a concept called "over-smoothing" in which the feature vectors of all nodes tend to reach an equivalent state. This has driven ongoing research on the diffusion process of GNN models [9], which is related to the topic of the present work. From this perspective, the authors of [9] worked on the application of a range of differential equations that are popular in classical physics; see, for instance, [62]. Of note, they also considered PDEs with a diffusion term, as used in works on image processing [40, 58]. The results of these studies motivate us to consider a PDE with a diffusion term here. This study is also motivated by ongoing research on optimal control theory, especially work on the ensemble controllability of stochastic processes in terms of the Fokker–Planck equation [2]. Although the drift term differs slightly from that used here, this highlights the necessity of the control of diffusion PDEs. Insights obtained in studies on machine learning literature might be helpful in this regard. Along these lines, we consider neural networks based on PDEs with a diffusion term in our study. Our motivations are twofold.

(i) Although neural ODEs perform well, their essential difference from classical neural networks is that the width of each layer does not change. This limitation can be overcome by PDE-based neural networks, which also consider the infinite limit of the width of the network. Because we aim to approximate a neural network with a continuous dynamical system, this advantage of PDE-based neural networks appears to be more natural.

(ii) Similar to that in the case of ODE-based control, some fruitful theories have also been provided on PDE-based control (or distributed control). A sophisticated

theoretical framework has been developed in considerable literature on diffusion equations. We can also understand the increasing freedom of such models by considering a range of forms and values of boundary conditions.

However, some uncertainties remain regarding the performance of these continuous neural networks. For example, the universal approximation property is an important aspect that all neural networks must exhibit.

Although various types of neural networks based on dynamical systems have been developed, some scope for further exploration remains in terms of their universal approximation property based on a PDE, particularly with a diffusion term.

In this paper, we first introduce the formulation of a PDE-based neural network and then show that it is well-defined under some natural setup conditions. Next, we prove the existence of a temporally global solution to the model. We also posit the existence of a vanishing diffusion limit. Finally, we show that our model possesses a universal approximation property with respect to the maximum norm.

The remainder of this paper is organized as follows. In Sect. 2, we define some notations that we use throughout this paper. In Sect. 3, we formulate the research problem and introduce some existence theorems. In Sect. 4, we present our main result, which is proven in Sect. 7. In Sect. 5, we compare our results to those of related works, referring to the history of arguments on the universal approximation property of neural networks. We also confirm the main contributions of the present work, and clarify our key theoretical and practical insights. Section 6 provides some preliminary statements to support the main results presented in Sect. 7. In Sect. 8, we discuss the learnability of our model as well as its performance based on some numerical experiments. Finally, our conclusions and some possible avenues for future research are presented in the final section.

2 Notations

In this section, we introduce some notations used for general analysis. First, let us define $I = (0,1)$ and $\partial I = \{0\} \cup \{1\}$. Let $\mathcal{G}$ denote an arbitrary region in $\mathbb{R}$. We denote the closure of $\mathcal{G}$ as $\overline{\mathcal{G}}$.

Hereafter, $C(\mathcal{G})$ denotes a set of continuous functions on $\mathcal{G}$. For $r \in \mathbb{N}$, a set of functions that are r-times continuously differentiable on $\mathbb{R}$ is denoted as $C^r(\mathbb{R})$. A set of infinitely differentiable functions with a compact support in $\mathcal{G}$ is denoted as $C_0^\infty(\mathcal{G})$. A set of Lipschitz continuous functions on $\mathbb{R}$ is denoted as $C^L(\mathbb{R})$. For $d \in \mathbb{N}$, we often denote a vector $\vec{u} = (u_1, u_2, \dots, u_d) \in \mathbb{R}^d$ as $[u_j]_j$. For two vectors $\vec{u}$ and $\vec{v} \in \mathbb{R}^p$ in general, we denote their inner product as $\vec{u} \cdot \vec{v}$. For a vector space X and an element $\vec{v} \in X$, we denote a set spanned by $\vec{v}$ as $\mathrm{Span}\langle \vec{v} \rangle$.

Let $\| \cdot \|_{L_p(\mathcal{G})}$ denote the usual L_p norm with $1 \le p \le +\infty$ on $\mathcal{G}$; i.e., for a function f in general, we define

$$\|f\|_{L_p(\mathcal{G})} \equiv \begin{cases} \left(\int_{\mathcal{G}} |f(x)|^p \, \mathrm{d}x\right)^{\frac{1}{p}} & (p \in [1, +\infty)), \\ \operatorname{ess\,sup}_{x \in \mathcal{G}} |f(x)| & (p = \infty). \end{cases}$$

We use a notation $(\cdot, \cdot)_{\mathcal{G}}$ to denote the inner product in $L_2(\mathcal{G})$ space:

$$(f, g)_{\mathcal{G}} \equiv \int_{\mathcal{G}} f(x) g(x) \, \mathrm{d}x.$$

In particular, when the region is clear, we simply denote it as $(\cdot,\cdot)$. The norm in $L_2(\mathcal{G})$ is often denoted as $|\cdot|$. We also use this notation to denote the norm in the Euclidean space, where a step function is regarded as a simple function in the L_2 space.

For $r \in \mathbb{N}$, we define Sobolev spaces $H^r(\mathcal{G})$, which are the spaces of functions $f(x), x \in \mathcal{G}$, equipped with the norm $\|f\|^2_{H^r(\mathcal{G})} \equiv \sum_{|\alpha|\le r} \|D^\alpha f\|^2_{L_2(\mathcal{G})}$. $H^{-r}(\mathcal{G})$ $(r > 0)$ is defined as the dual space of $H^r_0(\mathcal{G})$, which is the closure of $C^\infty_0(\mathcal{G})$ with respect to the norm of $H^r(\mathcal{G})$ (see, [49], §11.1 and §12.1).

For a Banach space $\mathcal{B}$ with the norm $\|\cdot\|_{\mathcal{B}}$, we denote the space of $\mathcal{B}$-valued measurable functions f on the interval (a,b) by $L_p((a,b);\mathcal{B})$, the norm of which is defined by

$$|f|_{L_p((a,b);\mathcal{B})} \equiv \begin{cases} (\int_a^b \|f(t)\|^p_{\mathcal{B}}\,\mathrm{d}t)^{\frac{1}{p}} & (p \in [1,+\infty)), \\ \operatorname{ess\,sup}_{a\le t\le b} \|f(t)\|_{\mathcal{B}} & (p = \infty). \end{cases}$$

Similarly, we often use notations like $C([a,b];\mathcal{B})$ to denote sets of $\mathcal{B}$-valued functions that are continuous with respect to time on the interval specified as the brackets. We also denote an adjoint space of $\mathcal{B}$ by $\mathcal{B}'$. For a Hilbert space H and its linear subspace $M \subset H$ in general, we denote the orthogonal complement of M as $M^\perp$. When the inner product of two elements v_1 and v_2 in H vanishes, we use the notation $v_1 \perp v_2$. Hereafter, we shall use the notations below:

$$Z *_x f(t,x) \equiv \int_I Z(t,x,y) f(t,y)\,\mathrm{d}y,$$

$$Z * f(t,x) \equiv \int_0^T \mathrm{d}\tau \int_I Z(t-\tau,x,y) f(\tau,y)\,\mathrm{d}y,$$

where $Z(t,x,y)$ denotes the fundamental solution to the initial boundary value problem of the heat equation with the vanishing Dirichlet condition. Given $T > 0$, we use the notation $\mathcal{H}_T \equiv (0,T) \times I \times I$. The other notations used in this paper are summarized in Table 4 in Appendix A.

3 Formulation: differential equation-based neural networks

Here, we formulate the continuous limit of a multi-layer neural network [32]. Because in the supervised learning, the input vector takes the form of a vector in a Euclidean space $\mathbb{R}^J$, we represent it as a simple function on a unit interval by partitioning it into J intervals. Given $T > 0$, we formulated the continuous version of a neural network as follows:

$$u_t - \nu u_{xx} = \phi\left(\int_I w_1(t,x,y) u(t,y)\,\mathrm{d}y\right) \quad \text{in } I_T \equiv I \times (0,T), \tag{3.1}$$

where $\nu > 0$, $\phi(\cdot)$ denotes the activation function, $T > 0$ corresponds to the depth of a classical neural network, and w_1 and w_0 are the weight parameters at the middle and output layers, respectively. Additionally, we impose the initial and boundary value conditions as follows:

$$u(0,x) = u_0(x) \quad \text{on } I, \qquad u = 1 \quad \text{on } \partial I\ \forall t \in (0,T). \tag{3.2}$$

We employ a non-vanishing Dirichlet condition in (3.2), with which we can easily assure the existence of a solution in (6.12) in the proof of Lemma 4. We shall comment on this

issue later again. In (3.2), given the input data $\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_J)^\top \in \mathbb{R}^J$, the initial data is a simple function of the form: $u_0(x) = \sum_{j=1}^J \xi_j \chi_{I_j}$, with χ_{I_j} as the indicator functions of $I_j \equiv ((j-1)/J, j/J]$ $(j = 1, 2, \dots, J)$. Because we usually deal with a finite dimensional input, we translate it into this finite dimensional vector, and the corresponding simple function on a unit interval I. This is a different formulation from that of Liu and Markowich [52], in which they employed a region of the same dimension as the input feature. In that model, they computed the multiple integral of the input over the d-dimensional space in each layer. In the case of a two-dimensional CNN, their formulation coincides with the functionality of the convolution layer. In higher dimensions, however, it is different from how the multi-layer neural network works in the usual supervised learning.

By taking $v \equiv u - 1$, we can transform problem (3.1)–(3.2) as below.

$$\begin{cases} v_t - v v_{xx} = \phi\left(\int_I w_1(t,x,y) v(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y\right) & \text{in } I_T, \\ v(0,x) = u_0(x) - 1 \equiv \tilde{u}_0 \quad \text{on } I, \\ v = 0 \quad \text{on } \partial I \ \forall t \in (0,T). \end{cases} \tag{3.3}$$

The following result was obtained for problem (3.3).

Theorem 1 *Let $T > 0$ be arbitrary, and the following be assumed*:

(i) $u_0 \in L_2(I)$,

(ii) $\phi \in C^L(\mathbb{R})$,

(iii) $w_1 \in L_2(\mathcal{H}_T)$.

Then, there exists a constant $T_{u_0} \in (0, T]$ that depends on $\|u_0\|_{L_2(I)}$ such that problem (3.3) has a unique solution $v \in C([0, T_{u_0}]; L_2(I))$ on the interval $[0, T_{u_0})$. In addition, this solution satisfies

$$\|v\|_{C([0,T_{u_0}];L_2(I))} \le c\big(\|u_0\|_{L_2(I)}\big),$$

where $c(\|u_0\|_{L_2(I)})$ is a positive constant that depends monotonically increasingly on $\|u_0\|_{L_2(I)}$.

We prove this theorem in Appendix B, in which we use the notation $A = -v\frac{\partial^2}{\partial x^2}$ and define a sesquilinear form $\sigma(\cdot,\cdot) : H_0^1(I) \times H_0^1(I) \to \mathbb{R}$ [21] by

$$(Au, v) = \sigma(u, v) \quad \big(u, v \in H_0^1(I)\big). \tag{3.4}$$

Remark 1 The solution v mentioned in Theorem 1 also belongs to the space [64]

$$L_2\big((0, T_{u_0}); H^1(I)\big) \cap H^1\big((0, T_{u_0}); H^{-1}(I)\big),$$

and satisfies the same estimates as the one in the theorem with the norm of these spaces. The proof of this fact is contained in the proof of Theorem 1 in Appendix B.

Remark 2 In our proof above, we do not require $\phi(\cdot)$ to satisfy $\phi(0) = 0$, nor the linear growth, as required in [52].

Next, we show the existence of a temporally global solution.

Theorem 2 *Let $T > 0$ be an arbitrary positive number and assume that in addition to the assumptions* (i), (ii) *of Theorem* 1, $w_1 \in L_2(\mathcal{H}_\infty)$ *is satisfied. Then, there exists a temporally global solution $v \in C([0,T]; L_2(I))$ to problem* (3.3), *which satisfies*

$$\sup_{t\in[0,T]} \left|v(t)\right| \leq \chi\left(\|u_0\|_{L_2(I)}\right),$$

where $\chi(\cdot)$ is a monotonically increasing function.

Remark 3 As in Theorem 1, the solution v mentioned in Theorem 2 also belongs to the space

$$L_2\big(\big(0,T;H^1(I)\big) \cap H^1\big((0,T);H^{-1}(I)\big),$$

and satisfies the same estimates as the one in the theorem with the norm of these spaces.

The proof of Theorem 2 is given in Appendix B as well. Note that the estimate above does not depend on the diffusion coefficient $\nu > 0$. Thus, under the assumptions of Theorem 2, we can let ν tend to zero, to assert the corollary below [47].

Corollary 1 *Under the assumptions of Theorem* 2, *if we denote the solution to* (3.3) *by $v^{(\nu)}$, then we can take a sequence $\{v^{(\nu_m)}\}_{m=1}^{\infty} \subset L_2(I_T)$ satisfying the following*:

$$\left\|v^{(\nu_m)} - v^{(0)}\right\|_{L_2(I_T)} \to 0, \quad m \to +\infty,$$

where $v^{(0)} \in L_2(I_T)$ is a solution to the hyperbolic equation

$$\begin{cases} v_t^{(0)} - \phi\big(\int_I w_1(t,x,y)v^{(0)}(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y\big) = 0 \quad \text{in } I_T, \\ v^{(0)}(0,x) = \tilde{u}_0 \quad \text{on } I. \end{cases}$$

In our previous studies [31, 32], we set several cost functions that corresponded to specific tasks and demonstrated the presence of optimal controls, and we used the gradient descent algorithm to find the sub-optimal control. For example, in [34] in which we discussed the multiclass classification problem, the cost function is given by

$$\begin{aligned} J[\vec{w}] = -\int \mathrm{d}P(\vec{X},\vec{t}_{(\vec{X})}) \Bigg[& \sum_{k=1}^{K-1} t_{(\vec{X}),k} \ln\left\{\phi_0^{(k)}\left(\int_I w_0^{(k)}(y)u(T,y;w_1,u_{0(\vec{X})})\,\mathrm{d}y\right)\right\} \\ & + t_{(\vec{X}),K} \ln\left\{1 - \sum_{k'=1}^{K-1} \phi_0^{(k')}\left(\int_I w_0^{(k')}(y)u(T,y;w_1,u_{0(\vec{X})})\,\mathrm{d}y\right)\right\}\Bigg] \\ & + \frac{\gamma_1}{2}\sum_{k=1}^{K-1}\left\|w_0^{(k)}\right\|_{H^1(I)}^2 + \frac{\gamma_2}{2}\|w_1\|_{L_2(\mathcal{H}_T)}^2, \end{aligned}$$

where $\phi_0(\cdot)$ is an activation function of the output layer, $P(\vec{X},\vec{t}_{(\vec{X})})$ is the probability distribution of $(\vec{X},\vec{t}_{(\vec{X})})$, and $t_{(\vec{X}),k} \in \{0,1\}$ satisfies $\sum_{k=1}^{K} t_{(\vec{X}),k} = 1\ \forall \vec{X} \in \mathbb{R}^J$. However, because we

consider the feed-forward network hereafter, we do not consider $\phi_0(\cdot)$ in the present paper. Instead, we discuss the universal approximation property of this neural network with the output layer of the linear unit.

Hereafter, we frequently represent the solution to (3.3) (or equivalently, (3.1)–(3.2)) as $u(t,x;w_1,\vec{\xi})$, to clarify its dependency on w_1 and $\vec{\xi}$. Thus, we regard the solution $u(T,x;w_1,\vec{\xi})$ as a function on K by identifying u_0 with $\vec{\xi}=(\xi_1,\dots,\xi_J)^\top$.

4 Main result

In this section, we show the universal approximation property of the partial differential equation-based neural network prescribed in Sect. 3. We discuss the universal approximation property of our neural network model based on a nonlinear partial differential equation [32]. As is the case with the previous works, we restrict ourselves to an arbitrary compact set $K\subset\mathbb{R}^J$. Our main result is

Theorem 3 *Let $T>0$ be given and $\phi\in C^L(\mathbb{R})$ be a non-polynomial function. Then, for an arbitrary compact set $K\subset\mathbb{R}^J$, $F\in C(K)$, and $\varepsilon>0$, there exist $w_0\in L_2(I)$, $w_1\in L_2(\mathcal{H}_T)$ such that*

$$\sup_{\vec{\xi}\in K}\left|F(\vec{\xi})-\int_I w_0(x)u(T,x;w_1,\vec{\xi})\,\mathrm{d}x\right|<\varepsilon,$$

where $u(T,x;w_1,\vec{\xi})$ is the value of a solution to (3.1)–(3.2) at time T that corresponds to the initial input value $\vec{\xi}$.

We will prove Theorem 3 in Sect. 8.

Remark 4 In this paper, we only consider the scalar-valued function $F:K\to\mathbb{R}$ as Leshno [43] did. This does not lose generality, because if we can approximate this function, then we can approximate an arbitrary continuous map $F:K\to\mathbb{R}^n$ by concatenating the network in parallel, as is done in [14], as long as $J\geq n$ holds. We also point out that our PDE-based neural network is defined only on one-dimensional Euclidean space. This is because in many supervised learning, the input data is a vector with independent attributes, which can be associated with a simple function on I as we did above. This is the similar approach with [72]. When we consider GNN, however, this assumption does not hold, which is one of our future works.

Remark 5 The controls w_0 and w_1 depend on T and ν. Therefore, we cannot assure at this moment that the same conclusion holds with $\nu=0$. The discussion concerning this vanishing diffusion limit is our future work.

5 Comparison with existing works

Before going on to the proofs of our results, we discuss here the difference and novelty of our result in comparison to the existing related works. Actually, numerous contributions to the literature have studied the universal approximation property of neural networks. As an early work, Lippmann [51] postulated the formation of a range of surfaces for classification tasks to classify the points in a topological space using a neural network with two hidden layers.

This conjecture was rigorously proven by Funahashi [22], who stated that an arbitrary continuous function on a compact subset K in $\mathbb{R}^n$ could be approximated with a neural network with a single hidden layer that contained a sigmoid activation function.

Funahashi [23] also hypothesized that any L_2 function could be approximated by a three-layer neural network with a finite number of units in the hidden layer. Four-layer networks have also been conjectured to outperform three-layer networks. These considerations are related to the study of the generalization performance of neural networks [6].

Irie and Miyake [36] derived the integral representation of three-layer neural networks based on the Fourier integral theorem under the continuity of the hidden layer.

Around the same time, Cybenko [11] first discussed the universal approximation property of sigmoidal functions. They showed that a set of the functions of the form $\sum_j w_j \sigma(\vec{y}^\top \vec{x} + b)$ with some constants w_j and b and vectors $\vec{x}$ and $\vec{y}$ in a compact space K is dense in $C(K)$. Their discussion did not assume the activation function to be monotonic.

Hornik, Stinchcombe, and White [35] proposed general measurable functions by making use of the Stone-Weierstrass Theorem and the cosine squasher proposed by Gallant and White. Their results can be regarded as similar to those of Funahashi [22].

Leshno [43] obtained more general results by using the fact that the set of functions spanned by the so-called ridge functions, i.e., those of the form $f(\vec{w}\top\vec{x} + \theta)$, is dense both in $C(\mathbb{R}^n)$ and $L_p(\mu)$, where μ is an arbitrary finite measure on $\mathbb{R}^n$. Recently, Yun [85] proved the approximation property of a neural network constructed using a parametric sigmoidal function.

Some Bayesian perspectives on neural networks, even with an infinite number of nodes, have been discussed (see, for example, [54, 84]). Their key insight is that as the number of nodes tends to infinity, the output can be regarded as a set of Gaussian processes.

However, all these studies considered only general neural networks. Many works have also considered the universal approximation property of neural networks based on continuous dynamical systems.

Haber and Ruthotto [27] proposed a formulation of a neural network in a supervised learning framework as a dynamical system. There, they clarified the necessary condition for the stability of an equilibrium point as well as the stability of Euler method as a discrete approximation of the continuous solution. They also pointed out the close relationship between backpropagation and the adjoint method in optimal control theory. Q. Li et al. [45] discussed the approximation property of an ODE-based neural network, and gave the sufficient condition under which the set of the realizations of an ODE-based neural network can approximate an arbitrary continuous map $f : \mathbb{R}^n \to \mathbb{R}^m$ $(n \geq 2)$ on any compact set with respect to L_p $(p \in [1, +\infty))$ norms.

Along these lines, Aizawa and Kimura [1] recently presented the universal approximation property of neural ODEs [10] and ResNet using the result of Leshno [43]. However, their method is restricted to linear models.

Esteve [18] presented the recent works concerning the approximation property of neural ODEs and, moreover, presented that the optimal value of the loss function is estimated above to the order of T^{-1} under the Tikhonov regularization (which they called an empirical risk minimization). Roughly contemporaneously, Teshima [76] also investigated the universal approximation property of neural ODEs. In their proof, they made use of their previous work [75] with a relatively slight modification. They also discussed the relationship between their result and a preceding work by Zhang et al. [88], which showed

a counterexample that cannot be approximated by a neural ODE. Zhang et al [88] also presented the universal approximation property of an augmented neural ODE [17].

Recently, a survey by DeVore et al. [14] thoroughly presented the existing results on the approximation property of neural networks. The power of Rectified Linear Unit (ReLU) networks was among the most important results introduced here, as they can contain all piecewise-continuous functions on an arbitrary compact set.

From the perspective of a practical application, Laakmann and Petersen [42] applied a neural network to the numerical computation of a transport equation.

Studies in the field of optimal control have also considered the universal approximation property of continuous neural networks.

Balet and Zuazua [61] proved the simultaneous controllability [48] of a flow map of an ODE. This means that given an arbitrary finite input in a Euclidean space, the flow map can lead to an arbitrary set of classification labels.

By making use of this property, they also showed that an arbitrary simple function, and consequently an L_2 map $f : \mathbb{R}^d \to \mathbb{R}^d$ can be approximated with arbitrary precision with respect to the L_2 norm. They also discussed the relationship of the universal approximation property and simultaneous control [48, 53]. However, their method is not applicable here for two reasons. First, their method of rotating the coordinate does not suffice because we consider equations with a diffusion term. Second, because we aim at an approximation with respect to the maximum norm, their method is not applicable, as it divides the region into two sections, in one of which the function is allowed to be discontinuous. From the perspective of optimal control theory, the universal approximation property corresponds to approximate ensemble controllability [53]. Thus, our arguments here can also be regarded as describing this property of a specific type of control via a nonlinear diffusion equation. We prove this property of our model using some results from studies on machine learning.

The relationship between the optimal control of neural network and optimal transport models has been pointed out as well (see, for instance, [50]). For example, Sontag and Sussmann [69] discussed the controllability of temporally continuous recurrent neural networks. Balet et al. [61] above also argued this point and studied a nonlinear transport equation, which they called a neural transport equation (NTE), as given below.

$$\begin{aligned} &\partial_t \varrho + \nabla \cdot \left[\left(W(t)\sigma\left(A(t)x + b(t)\right)\right)\varrho \right] = 0, \\ &\varrho(0) = \varrho^0. \end{aligned}$$

They proved a method to approximate a target measure in the form of a finite combination of Dirac measures by the solution of an NTE at $t = T$ with arbitrary precision in the sense of 1-Wasserstein distance.

In [45], the authors theoretically considered the formulation of an ODE-based neural network and proved its universal approximation property. They first observed that the earlier discussion concerning the universal approximation property of neural ODEs [88] relied on a stronger assumption under which the right-hand side of an ODE already possesses the universal approximation property. They showed that any arbitrary continuous function $f : \mathbb{R}^n \to \mathbb{R}^n$ on any compact set in a Euclidean space can be approximated in an L_p norm with arbitrary precision. They also pointed out that the set of realizations of an ODE is uniformly approximated by that of ResNet. They derived their main results based

on another work by one of the authors [66]. However, in their formulation, they distinguished the one-dimensional and multi-dimensional input cases. In contrast, our proof in the present work need not distinguish these cases, because we start from Leshno's result [43].

Regarding ResNet, Tabuada [73] gave some conditions on activation functions under which the universal approximation property of a map $f; \mathbb{R}^n \to \mathbb{R}^n$ with L_∞ norm is assured. They used the technique of ensemble controllability and deduced the quadratic differential equation that should be satisfied for each activation function. Research is also being actively conducted on the controllability of systems driven by linear or nonlinear partial differential equations [4].

For example, Fernández-Cara et al. [20] studied the null controllability of a heat equation with a spatially nonlocal term, which is roughly similar to the setup considered in the present work. Because their model was linear, they considered its adjoint equation and applied the Fourier-series representation of the equation and the compactness-uniqueness argument. They also stated that the approximate controllability, which is equivalent with the universal approximation property in the terminology of our PDE-based neural network, holds under the analyticity assumption of the kernel operator of the nonlocal term. However, our setup employs a nonlinear activation function, which is essentially different from this work; they listed the nonlinear case as an open problem.

As another example, [26] discussed the controllability of a nonlinear heat equation with a distributed control in an unbounded domain in $\mathbb{R}^n$. In this formulation, however, the control term is not included in the nonlinear term, which differs from our framework.

In fact, the application of diffusion equations in image processing has been discussed in prior works [82, 86]. Along these lines, Ruthotto and Haber [63] proposed parabolic and hyperbolic CNNs models that respectively included spatial and temporal second-order derivative terms. They also considered the application of neural networks in this field as an extension of prior applications of PDEs.

Some other works have also addressed the control of parabolic PDEs [4], including linear and nonlinear heat equations. In this regard, the present work provides a link between the insights in the literature on neural networks and research on controllability in optimal control theory.

Regarding the PDE-based neural network, Liu and Markowich [52] proposed a hyperbolic nonlinear integro-differential form without a diffusion term. However, they considered only the mathematical well-posedness of the form, and did not mention the universal approximation property. An earlier work [46] also proposed PDE-based neural networks for the transport and HJB equations, one of which used a diffusion term as in the present work. Neither of these works, however, mentioned the universal approximation properties of the models. Li and Shi [46] also proposed adding an extra constraint in cases with a diffusion term. In contrast, the present work shows that the universal approximation property is satisfied even without a trick of this nature.

At the end of this section, we list some recent notable results. Ivan et al. [59] proposed a framework to train neural ODEs using the Lyapunov function, which avoids the traditional backpropagation and achieves a faster computation. Moreover, a link between turnpike theory and optimal control has been considered in relation to neural ODEs [24]. Geshkovski and Zuazua [24] computed some examples of the turnpike property of a neu-

ral ODE using the MNIST dataset. They also mentioned that related results have been reported in the form of specific setups as in [18] and [19].

Based on the aforementioned arguments, the main contributions of this study are summarized as follows.

(i) Motivated by the application of diffusion equations in image processing and considering GNNs with a diffusion term, we formulate PDE-based neural networks with a diffusion term and rigorously clarify the conditions under which the existence of the solution is assured.
(ii) We describe the universal approximation property of our model in the sense of the maximum norm.

Our key findings are summarized as follows.

(i) We show that some insights from studies on machine learning can be applied to the theory of the optimal control of PDEs.
(ii) Even though Leshno's result (Lemma 1 below) is a useful tool to prove our result for continuous neural networks, some additional formulations are required to discuss the convergence of the temporal and spatial unit when they tend to 0.
(iii) Because our model contains a diffusion term, our method differs from those presented in prior works, although it is based on that reported by Leshno [43]. More concretely, our proof uses estimates of the approximation of the discretized diffusion equation that were not considered in previous studies.

In subsequent sections, we prove the results presented above.

6 Preliminary results

Before proving Theorem 3, we prepare some auxiliary results in this section. We first cite the following lemma ([43], Theorem 1).

Lemma 1 *Let f be a measurable function on $\mathbb{R}$ with a certain $J \in \mathbb{N}$. Then,* $\operatorname{Span}\langle f_{\mathbf{w},\theta}(x)\rangle$ $(\mathbf{w} \in \mathbb{R}^J, \theta \in \mathbb{R})$ *is fundamental in $C(\mathbb{R}^J)$ if and only if f is not a polynomial, where $f_{\mathbf{w},\theta}(x) = f(\mathbf{w} \cdot \mathbf{x} + \theta)$.*

Owing to this lemma, for an arbitrary $\vec{\xi} \in K \subset \mathbb{R}^J$ and $\varepsilon > 0$, by taking a suitable $M \in \mathbb{N}$, $\{\sigma_{0(m)}\}_{m=1}^M \subset \mathbb{R}$, $\{\vec{\sigma}_1^{(m)}\}_{m=1}^M \subset \mathbb{R}^J$, and $\{\theta^{(m)}\}_{m=1}^M \subset \mathbb{R}$, we can obtain

$$\sup_{\vec{\xi}\in K}\left|F(\vec{\xi}) - \sum_{m=1}^{M} \sigma_{0(m)}\phi\left(\vec{\sigma}_1^{(m)\top}\vec{\xi} + \theta^{(m)}\right)\right| < \varepsilon. \tag{6.1}$$

Based on Lemma 1, we construct an approximating solution of (3.3) that agrees with the approximation stated in (6.1). Next, by taking the temporal and spatial meshes finer enough, we show that the solution of (3.3) itself can approximate the target continuous function. In applying these steps, we make use of the estimate on the approximation accuracy of Galerkin approximation.

Let us consider the approximate problem of (3.3). Regarding the spatial variable, we employ the Galerkin approximation [77]. For this purpose, let $\mathfrak{S}_h \equiv \{S_h\}_h$ be a family of finite-dimensional subspaces of $H_0^1(I)$ with parameter $h < 1$ that tends to 0 [77]. In the sequel, we set an integer L and take $h = \frac{1}{L}$ (i.e., we divide I into L equipartitions). It is also

assumed that

$$\inf_{g\in S_h}\left(\|v-g\|_{L_2(I)}+h\|v-g\|_{H^1(I)}\right)\le Ch^s\|v\|_{H^s(I)}\quad(1\le s\le r)$$

holds (r is a positive value, for example, $r=2$ [3]). We also define an approximation operator $A_h:\mathfrak{S}_h\to\mathfrak{S}_h$ by using the sesqui-linear form $\sigma(\cdot,\cdot)$ in (3.4) as follows:

$$(A_h\phi_h,\psi_h)=\sigma(\phi_h,\psi_h).$$

Thus, A_h is the operator associated with the restriction of $\sigma(\cdot,\cdot)$ on $\mathfrak{S}_h\times\mathfrak{S}_h$ [21]. We further define an operator P_h indicating a projection of $u\in L_2(I)$ onto $\mathfrak{S}_h$ with respect to the L_2 inner product [21].

Then, we divide the time interval $(0,T]$ into N intervals $\{[(n-1)k,nk)\}_{n=1}^N$, with $Nk=T$. By using a notation $U_h^{(n)}=[U_{h(l)}^{(n)}]_l$, we consider the discretized scheme of (3.3) on $t\in(0,T]$.

$$\begin{cases}\widetilde{U}_h^{(n)}=P_hu(nk,\cdot;0,\vec{\xi})\quad(n=0,1,2,\dots,N-2),\\ \widetilde{U}_h^{(N-1)}=\overline{P}_hu((N-1)k,\cdot;0,\vec{\xi}),\\ \widetilde{U}_h^{(N)}=r(kA_h)\widetilde{U}_h^{(N-1)}\\ \qquad\quad+kr(kA_h)P_h\phi\left(h\sum_{l=2}^{L-1}w_{\cdot,l}^{(N-1)}\widetilde{U}_{h(l)}^{(N-1)}+h\sum_{l=1}^{L}w_{\cdot,l}^{(N-1)}\right),\end{cases}\tag{6.2}$$

where $r(kA_h)$ denotes the Padé approximation [3] of the semigroup

$$e^{-tA_h}\approx r(kA_h)\equiv(I_L+kA_h)^{-1}.$$

Here, I_L is an L-dimensional identity matrix. We also use a notation $\|\cdot\|$ hereafter to denote a Euclidean norm; note that this is equivalent with L_2 norm as long as we consider the piecewise $L_2(I)$ functions.

Moreover, we introduce notations $\overline{P}_hf=[(\overline{P}_hf)_l]_l$, with

$$(\overline{P}_hf)_l\equiv\frac{1}{h}\int_{I_l}f(x)\,\mathrm{d}x\quad(l=1,2,\dots,L),$$

for $f\in L_1(I)$ in general, where $I_l\equiv[\frac{l-1}{L},\frac{l}{L})$ $(l=0,1,2,\dots,L)$, and $\overline{P}_h:L_2(I)\to\mathfrak{S}_h$, the projection onto the finite dimensional space $\mathfrak{S}_h$ for each $h=\frac{1}{L}$. Because these are projection operators, note that the inequalities $\|\overline{P}_hf\|\le\|f\|$ and $\|P_hf\|\le\|f\|$ always hold. The value $\|\overline{P}_hf\|$ is computed by regarding it as a simple function on I, and then taking the usual norm of $L_2(I)$.

Remark 6 The operator $\overline{P}_h$ has often been used in the literature on the discrete approximation of operators ([79, 81]). It is known that this is equivalent with the operation

$$\widetilde{P}_hf\equiv\left[f(lh)\right]_l\in\mathbb{R}^L,$$

in the sense that the following equality holds [81].

$$\lim_{h\to0}\|\overline{P}_hf-\widetilde{P}_hf\|=0.$$

In (6.2), we utilized the vanishing Dirichlet condition of (3.3). From theory, $u((N-1)k,x)$ can be represented as follows [37].

$$\begin{aligned} u\big((N-1)k,x;0,\vec{\xi}\big) &= Z *_x \tilde{u}_0 + Z * \phi(0) \\ &= \sum_{j=1}^{\infty} \lambda_j \left(\eta_j, \sum_{q=1}^{J} \xi_q \chi_{I_q} \right) e^{-\lambda_j (N-1)k} \eta_j(x) + c\big((N-1)k,x\big), \end{aligned}$$

where λ_j and η_j are the eigenvalues and eigenvectors of an operator A, respectively, and

$$c(t,x) = -Z *_x 1 + Z * \phi(0).$$

Hereafter, we often use a notation $u((N-1)k,\cdot)$ to denote $u((N-1)k,\cdot;0,\vec{\xi})$. Thus, we have

$$\begin{aligned} \widetilde{U}_{h(l)}^{(N-1)} &= \overline{P}_h u\big((N-1)k,\cdot\big) \\ &= \sum_{j=1}^{\infty} \lambda_j \left(\eta_j, \sum_{q=1}^{J} \xi_q \chi_{I_q} \right) e^{-\lambda_j (N-1)k} (\overline{P}_h \eta_j)_l + \big(\overline{P}_h c\big((N-1)k,\cdot\big)\big)_l \\ &\equiv \vec{c}^{(l)\top} \vec{\xi} + \big(\overline{P}_h c\big((N-1)k,\cdot\big)\big)_l \quad (l = 1,2,\dots,L). \end{aligned} \tag{6.3}$$

Hereafter, we use the notation $\vec{\sigma}_0 = [\sigma_{0(m)}]_{m=1}^{M} \in \mathbb{R}^M$. We prepare a lemma.

Lemma 2 *With $L = aM$, where an integer a has a sufficiently large value, there exist $\vec{w}'_{0(h,k)} \in \mathbb{R}^L$, $\{\vec{\theta}_{0(h)}^{(p)}\}_{p=1}^{L} \subset \mathbb{R}^J$, and $\{\theta_{1(h)}^{(p)}\}_{p=1}^{L} \subset \mathbb{R}$ such that*

$$\begin{aligned} &\vec{w}'^{\top}_{0(h,k)} \big[r(kA_h) \widetilde{U}_h^{(N-1)} + kr(kA_h) \big[\phi\big(h\vec{\theta}_{0(h)}^{(p)\top} \vec{\xi} + \theta_{1(h)}^{(p)} \big) \big]_p \big] \\ &\quad = \vec{\sigma}_0^{\top} \big[\phi\big(\vec{\sigma}_1^{(m)\top} \vec{\xi} + \theta^{(m)} \big) \big]_m. \end{aligned}$$

Remark 7 The left-hand side of the equality in Lemma 2 is the inner product of the vectors in $\mathbb{R}^L$, whereas the right-hand side is that of the vectors in $\mathbb{R}^M$.

Proof First, we introduce disjoint subsets of $\{1,2,\dots,L\}$:

$$D_{(m)} \equiv \big\{ (m-1)a+1, (m-1)a+2, \dots, ma \big\} \quad (m = 1,2,\dots,M).$$

It is obvious that $\{1,2,\dots,L\} = \bigcup_{m=1}^{M} D_{(m)}$. Then, we take $\vec{\theta}_{0(h)}^{(p)}$ and $\theta_{1(h)}^{(p)}$ so that $h\vec{\theta}_{0(h)}^{(p)} = \vec{\sigma}_1^{(m)}$ $(p \in D_{(m)})$ and $\theta_{1(h)}^{(p)} = \theta^{(m)}$ $(p \in D_{(m)})$, respectively. Let us take $\vec{w}'_{0(h,k)}$ so that the followings are satisfied.

$$\vec{w}'^{\top}_{0(h,k)} r(kA_h) \widetilde{U}_h^{(N-1)} = 0, \tag{6.4}$$

$$k\vec{w}'^{\top}_{0(h,k)} r(kA_h) \boldsymbol{B}_h = \vec{\sigma}_0^{\top}, \tag{6.5}$$

where $\boldsymbol{B}_h = h[\vec{e}_1, \vec{e}_2, \dots, \vec{e}_M]$ is an $L \times M$ matrix with $\vec{e}_j = [H(l \in D_{(j)})]_l \in \mathbb{R}^L$, $H(\cdot)$ being a function that returns unity if the statement in the bracket is true, and returns 0 otherwise. (6.4) means that $\vec{w}'_{0(h,k)}$ should belong to a subspace in $\mathbb{R}^{L-2}$, which is denoted as $\mathcal{G}_h$

hereafter. Therefore, we rewrite (6.5) as follows:

$$\boldsymbol{B}_h^\top r(kA_h)^\top|_{\mathcal{G}_h}\vec{w}'_{0(h,k)} = \vec{\sigma}_0/k, \tag{6.6}$$

where $\boldsymbol{B}_h^\top r(kA_h)^\top|_{\mathcal{G}_h}$ denotes the restriction of $\boldsymbol{B}_h^\top r(kA_h)^\top$ onto the space $\mathcal{G}_h \subset \mathbb{R}^{L-2}$.

Based on proposition 8.14, which was presented in [87], (6.6) can have a solution if and only if $\vec{\sigma}_0 \perp N(r(kA_h)\boldsymbol{B}_h|_{\mathcal{G}_h})$, where $N(\cdot)$ denotes the kernel of the operator in its argument.

Conversely, we can show that $r(kA_h)$ is of full rank. In fact, if we consider that A_h is positive definite, all the eigenvalues of A_h are positive. Moreover, because A_h is self-adjoint, we observe that it is diagonalizable [12], and so is $r(kA_h)$. Thus, $r(kA_h)\vec{\nu} = 0$ means $\vec{\nu} = \vec{0}$. However, it is apparent that $\boldsymbol{B}_h$ is suborthogonal in the sense that its column vectors are orthogonal to each other. Thus, $N(r(kA_h)\boldsymbol{B}_h|_{\mathcal{G}_h}) = \{0\}$, which yields the desired result. □

Remark 8 From construction, the solution $\vec{w}'_{0(h,k)}$ to (6.6) depends on h and k. As above, (6.6) has at least one solution. If we denote this solution by $\breve{w}'_{0(h,k)}$, then a set of solutions for (6.6) can be denoted as $\breve{w}'_{0(h,k)} + N(\boldsymbol{B}_h^\top r(kA_h)^\top)$.

Now, we define:

$$w_0(x) = w'_{0(h,k)(l)} \quad \text{on } I_l \ (l = 1, 2, \ldots, L),$$

where $w'_{0(h,k)(l)}$ is the l-th component of the vector $\vec{w}'_{0(h,k)}$. We assert that, for a certain $R > 0$, we have at least one solution stated in Lemma 2, in a certain ball with radius R in $L_2(I)$. In the sequel, we use the following notations:

$$\mathcal{G}_\infty^{(k)} \equiv \big\{f \in L_2(I) | f \perp \operatorname{Span}\big\langle r(kA)u\big((N-1)k, \cdot\big)\big\rangle\big\},$$
$$\mathcal{G}_h^{(k)} \equiv \big\{f \in L_2(I) | f \perp \operatorname{Span}\big\langle r(kA_h)\overline{P}_h u\big((N-1)k, \cdot\big)\big\rangle\big\}.$$

Now, we state

Lemma 3 *For a certain $R > 0$, k and $h_1 > 0$, we have a solution $\vec{w}'_{0(h,k)}$ to (6.4) and (6.5) that satisfies*

$$\big\|\vec{w}'_{0(h,k)}\big\| \le R$$

for $\forall h \in (0, h_1]$.

Proof First, we take a small $k > 0$, $h_1 = \frac{1}{L_1} > 0$, and $\vec{w}'_{0(h_1,k)} \in \mathbb{R}^{L_1}$, which satisfy

$$\begin{cases} k\boldsymbol{B}_{h_1}^\top r(kA_{h_1})^\top \vec{w}'_{0(h_1,k)} = \vec{\sigma}_0, \\ \vec{w}'^\top_{0(h_1,k)} r(kA_{h_1})\widetilde{U}_{h_1}^{(N-1)} = 0. \end{cases} \tag{6.7}$$

Note that the existence of such $\vec{w}'_{0(h_1,k)}$ is guaranteed by Lemma 2. Moreover, the solution $\vec{w}'_{0(h_1,k)}$ to (6.7) belongs to the intersection of $\mathcal{G}_{h_1}^{(k)}$ and a set represented as $\breve{w}_{0(h_1,k)} + N(\boldsymbol{B}_{h_1}^\top)$, where $\breve{w}'_{0(h_1,k)}$ is the solution to the problem in $\mathfrak{S}_{h_1}$:

$$r(kA_{h_1})^\top \breve{w}'_{0(h_1,k)} = \overline{P}_{h_1}\widetilde{\sigma}_0/k, \tag{6.8}$$

with $\frac{1}{h_1} = L_1 = a_1 M$. Here, $\widetilde{\sigma}_0$ is a notation used when we regard $\vec{\sigma}_0$ as an element in $L_2(I)$. Hereafter, we often regard $\mathcal{G}_{h_1}^{(k)}$ as a subset of $L_2(I)$. Note that we can easily obtain the solution of (6.8) if we recall the definition of $r(kA_{h_1})$. We denote one such solution as $\check{w}'_{0(h_1,k)} \in \mathcal{G}_{h_1}^{(k)}$ again:

$$\begin{cases} r(kA_{h_1})^{\top} \check{w}'_{0(h_1,k)} = \overline{P}_{h_1} \widetilde{\sigma}_0 / k, \\ \check{w}'^{\top}_{0(h_1,k)} r(kA_{h_1}) \widetilde{U}_{h_1}^{(N-1)} = 0. \end{cases}$$

For $h > 0$, we define a map $G_h^{(k)} : L_2(I) \to \mathfrak{S}_h$ as follows:

$$G_h^{(k)}\big[\check{w}'_{0(h,k)}\big] = r(kA_h) \check{P}_{\mathcal{G}_h} \check{w}'^{\top}_{0(h,k)} - \overline{P}_h \widetilde{\sigma}_0 / k,$$

where $\check{P}_{\mathcal{G}_h^{(k)}} : L_2(I) \to \mathcal{G}_h^{(k)}$ is a projection onto $\mathcal{G}_h^{(k)}$ with respect to the L_2 inner product.

We will below that if the norm of $\check{w}'_{0(h,k)}$ is large enough, even if we take the projection above, the norm of $\check{w}'_{0(h,k)}$ after the projection is large enough as well.

Next, we define

$$S_R^{(k)} \equiv \big\{ f \in L_2(I) \,|\, \|f\| = R, f \in \mathcal{G}_{\infty}^{(k)} \big\} \quad \big(R > \|\overline{P}_h \widetilde{\sigma}_0 / k\| \big).$$

Because the dimension of $(\mathcal{G}_{\infty}^{(k)})^{\perp}$ is of unity, it holds that $S_R^{(k)} \neq \emptyset$. We also take a small $\varepsilon_1 > 0$ and sufficiently small $\tilde{h}$ so that

$$\big| \big(f, r(kA_{\tilde{h}}) \overline{P}_{\tilde{h}} u\big((N-1)k, \cdot\big) \big) \big| < \varepsilon_1 \quad \forall f \in S_R^{(k)}.$$

This is possible if we note

$$\begin{aligned} &\big| \big| \big(f, r(kA_h) \overline{P}_h u\big((N-1)k, \cdot\big) \big) \big| - \big| \big(f, r(kA) u\big((N-1)k, \cdot\big) \big) \big| \big| \\ &\quad \le \|f\| \big\| r(kA_h) \overline{P}_h u\big((N-1)k, \cdot\big) - r(kA) u\big((N-1)k, \cdot\big) \big\|, \end{aligned}$$

and the relationship that holds with $v \in L_2(I)$ [3]:

$$\big\| r(kA_h) v - r(kA) v \big\| \le c\big(\gamma(h) + h^r + k \big) \|v\|, \tag{6.9}$$

with r being the one stated right after (6.1), where $\gamma(h)$ tends to zero as h does. Moreover, let $R > 0$ have a sufficiently large value so that the following holds (R should be redefined, if necessary):

$$\big\| r(kA) v_0 - \overline{P}_h \widetilde{\sigma}_0 / k \big\| > \delta_0 \quad \forall v_0 \in S_R^{(k)}. \tag{6.10}$$

In order to show that this is possible, we can demonstrate the continuity of the resolvent $r(kA) = (I_d + kA)^{-1}$ with respect to k, where I_d is an identity operator. We can prove this by using the resolvent equation [39] and the boundedness of $r(kA)$, as presented by Fujita and Mizutani [21]. Thus, for $\varepsilon_1 > 0$ above, if we take a sufficiently small k, we have

$$\big\| r(kA) v_0 - v_0 \big\| \le \frac{\varepsilon_1}{2},$$

for $v_0 \in S_R^{(k)}$. This yields

$$\left\| r(kA)v_0 \right\| \geq R - \varepsilon_1,$$

with an arbitrary $\varepsilon_1 > 0$. Therefore, if we take R sufficiently large, we arrive at (6.10) and consequently,

$$G_h^{(k)}[v_0] \neq 0 \quad \forall h \in \left(0, \min\{h_1, \tilde{h}\}\right),$$

for this v_0. Now, for an arbitrary $h_2 = \frac{1}{a_2 M}$ with $a_2 > a_1$, we define a homotopy mapping $H : L_2(I) \times [0,1] \to \mathfrak{S}_{h_2}$:

$$H(f,s) \equiv sD_{a_2,a_1} G_{h_1}^{(k)} f + (1-s) G_{h_2}^{(k)} f,$$

where $s \in [0,1]$ and D_{a_2,a_1} is a $a_2 M \times a_1 M$ matrix whose components are either 0 or 1. That is, this matrix is used to translate the image of $G_{h_1}^{(k)}$ as an element of $\mathbb{R}^{a_2 M}$. In virtue of the arguments presented above, we have

$$H(f,s) \neq \vec{\sigma}_0 \quad \forall f \in S_R^{(k)}, s \in [0,1].$$

Then, we have

$$H(f,0) = G_{h_2}^{(k)} f, \qquad H(f,1) = D_{a_2,a_1} G_{h_1}^{(k)} f,$$

and $H(f,s)$ is a compact operator for each s because its range has a finite dimension. Owing to the result of degree theory [87], we can conclude that the equation

$$G_{h_2}^{(k)} f = 0$$

has a solution. Consequently,

$$\boldsymbol{B}_{h_2}^{\top} r(kA_{h_2})^{\top} \check{P}_{\mathcal{G}_{h_2}^{(k)} f} = \boldsymbol{B}_{h_2}^{\top} \overline{P}_{h_2} \widetilde{\sigma}_0 / k$$

has a solution $\forall h_2 \in (0, h_1]$ that satisfies $\|f\| \leq R$. If we take $\vec{w}'_{0(h_2,k)} = \check{P}_{\mathcal{G}_{h_2}^{(k)} f}$, this is the desired solution. □

By using this, we assert the following lemma.

Lemma 4 *Let h and k be sufficiently small positive numbers. Then, for an arbitrary $\vec{\xi} \in K \subset \mathbb{R}^J$ and $\varepsilon > 0$, there exists an array $\boldsymbol{W} = [w_{p,l}^{(N-1)}]_{p,l=1,2,\dots,L}$, with which $\widetilde{U}_h^{(N)}$ defined in (6.2) satisfies*

$$\left| F(\vec{\xi}) - \vec{w}'^{\top}_{0(h,k)} \widetilde{U}_h^{(N)} \right| < \frac{\varepsilon}{2}. \tag{6.11}$$

Proof In fact, based on (6.4) and (6.5), we consider the following equations for $\boldsymbol{W} = [w_{p,l}^{(N-1)}]_{p,l=1,2,\dots,L}$.

$$\begin{cases} h\sum_{l=2}^{L-1} w_{p,l}^{(N-1)} \vec{c}^{(l)} \cdot \vec{\xi} = h\vec{\theta}_{0(h)}^{(p)} \cdot \vec{\xi}, \\ h\sum_{l=2}^{L-1} w_{p,l}^{(N-1)} \overline{P}_h c((N-1)k, l) + h\sum_{l=1}^{L} w_{p,l}^{(N-1)} = \theta_{1(h)}^{(p)} \\ \quad (p = 1, 2, \dots, L). \end{cases} \tag{6.12}$$

For each fixed p ($1 \le p \le L$), this can be written as an equation for $\vec{w}_p \equiv [w_{p,l}^{(N-1)}]_l$ as shown below:

$$h\boldsymbol{T}_p \vec{w}_p = \breve{\boldsymbol{\theta}}_{p(h)} \equiv \left(h\vec{\theta}_{0(h)}^{(p)} \cdot \vec{\xi}, \theta_{1(h)}^{(p)}\right)^\top \in \mathbb{R}^2, \tag{6.13}$$

where

$$\boldsymbol{T}_p \equiv \begin{bmatrix} 0 & \vec{c}^{(2)} \cdot \vec{\xi} & \cdots & \vec{c}^{(L-1)} \cdot \vec{\xi} & 0 \\ 1 & 1 + (\overline{P}_h c((N-1)k, \cdot))_2 & \cdots & 1 + (\overline{P}_h c((N-1)k, \cdot))_{L-1} & 1 \end{bmatrix} \in \mathbb{R}^{2\times L}.$$

Thanks to the same argument as in the proof of Lemma 2, we shall show that $N(\boldsymbol{T}_p^\top) = \{0\}$. In fact, if we recall that $\boldsymbol{T}_p$ is a linear map from $\mathbb{R}^2$ to $\mathbb{R}^L$, if $\boldsymbol{T}_p^\top \vec{q} = \vec{0}$ holds with $\vec{q} = (q_1, q_2)^\top$, then, it can be easily observed that $q_2 = 0$ (actually, adding a non-vanishing Dirichlet boundary condition in (3.2) works here). Regarding q_1, if $q_1 \neq 0$, all the following equalities should hold:

$$\vec{c}^{(l)} \cdot \vec{\xi} = 0 \quad (l = 2, 3, \dots, L-1).$$

However, from (6.3), this means that

$$\left[\overline{P}_h Z * u_0(T, \cdot)\right]_l = 0 \quad (l = 2, 3, \dots, L-1), \tag{6.14}$$

which does not hold if we take L sufficiently large. In fact, for an arbitrary $\varepsilon' > 0$, if we take $h > 0$ small enough, we obtain

$$\left\| Z * u_0(T, \cdot) - \overline{P}_h\left(Z * u_0(T, \cdot)\right) \right\| < \frac{\varepsilon'}{2},$$

and thus, we have

$$\begin{aligned} \left\| Z * u_0(T, \cdot) \right\| &\le \left\| Z * u_0(T, \cdot) - \overline{P}_h\left(Z * u_0(T, \cdot)\right) \right\| + \left\| \overline{P}_h\left(Z * u_0(T, \cdot)\right) \right\| \\ &\le \frac{\varepsilon'}{2} + \left\| \overline{P}_h\left(Z * u_0(T, \cdot)\right) \right\|. \end{aligned}$$

But (6.14) implies that if we take h sufficiently small, then we can attain

$$\left\| \overline{P}_h\left(Z * u_0(T, \cdot)\right) \right\| < \frac{\varepsilon'}{2}.$$

Thus, we have $\|Z * u_0(T,\cdot)\| < \varepsilon'$. Because ε' is arbitrary, we have $Z * u_0(T,\cdot) = 0$. If we recall (6.3), this implies

$$\sum_{j=1}^{\infty} \lambda_j \left(\eta_j, \sum_{q=1}^{J} \xi_q \chi_{I_q} \right) e^{-\lambda_j (N-1)k} \eta_j(x) = 0,$$

from which we obtain $\lambda_j(\eta_j, \sum_{q=1}^{J} \xi_q \chi_{I_q}) e^{-\lambda_j (N-1)k} = 0$, and consequently, $(\eta_j, \sum_{q=1}^{J} \xi_q \chi_{I_q}) = 0 \ \forall j = 1, 2, \dots$. This means $u_0 \equiv 0$, a contradiction. Thus, we can conclude that $\vec{q} = \vec{0}$, which means that $N(\boldsymbol{T}_p^{\top}) = \{\vec{0}\}$. Thus, (6.13) and consequently (6.12) has a solution. This means that

$$\begin{aligned} &\vec{w}_{0(h,k)}^{\prime\top} \left[r(kA_h) \widetilde{U}_h^{(N-1)} + kr(kA_h) \left[\phi \left(h \sum_{l=2}^{L-1} w_{p,l}^{(N-1)} \vec{c}^{(l)} \cdot \vec{\xi} \right.\right.\right. \\ &\qquad \left.\left.\left. + h \sum_{l=2}^{L-1} w_{p,l}^{(N-1)} \overline{P}_h c\big((N-1)k, l\big) + h \sum_{l=1}^{L} w_{p,l}^{(N-1)} \right) \right]_p \right] \\ &\quad = \vec{\sigma}_0^{\top} \big[\phi \big(\vec{\sigma}_1^{(m)\top} \vec{\xi} + \theta^{(m)} \big) \big]_m, \end{aligned}$$

holds with $\boldsymbol{W} = [w_{p,l}^{(N-1)}]_{p,l=1,2,\dots,L}$ prescribed. Moreover, recalling (6.3), this is rewritten as

$$\begin{aligned} &\vec{w}_{0(h,k)}^{\prime\top} \left[r(kA_h) \widetilde{U}_{h(l)}^{(N-1)} + kr(kA_h) \left[\phi \left(h \sum_{l=2}^{L-1} w_{p,l}^{(N-1)} \widetilde{U}_h^{(N-1)} + h \sum_{l=1}^{L} w_{p,l}^{(N-1)} \right) \right]_p \right] \\ &\quad = \vec{\sigma}_0^{\top} \big[\phi \big(\vec{\sigma}_1^{(m)\top} \vec{\xi} + \theta^{(m)} \big) \big]_m. \end{aligned}$$

If we further recall (6.2), this implies that

$$\vec{w}_{0(h,k)}^{\prime\top} \big[\widetilde{U}_h^{(N)} + kr(kA_h)[P_h \phi_1 - \phi_1]_p \big] = \vec{\sigma}_0^{\top} \big[\phi \big(\vec{\sigma}_1^{(m)\top} \vec{\xi} + \theta^{(m)} \big) \big]_m,$$

where $\phi_1 = \phi(h \sum_{l=2}^{L-1} w_{p,l}^{(N-1)} \widetilde{U}_{h(l)}^{(N-1)} + h \sum_{l=1}^{L} w_{p,l}^{(N-1)})$.

Note that $\|\vec{w}_{0(h,k)}^{\prime}\|$ is bounded with respect to (h,k) thanks to the proof of Lemma 3. Thus, if we take k small enough, we can bound the second term of the left-hand side above small enough. This, together with Lemma 1, yields the desired statement. □

By using the solution of (6.13), we construct a function $\bar{w}_1(t,x,y)$

$$\bar{w}_1(t,x,y) = \begin{cases} w_{p,l}^{(N-1)} & \text{on } ((N-1)k, Nk] \times I_p \times I_l \ (p,l = 1,2,\dots,L), \\ 0 & \text{on } (0,(N-1)k]. \end{cases} \tag{6.15}$$

By noting that $\bar{w}_1 \in L_2(\mathcal{H}_T)$, we set $\bar{u}(t,x) \equiv u(t,x;\bar{w}_1,\vec{\xi})$, which solves (3.3) with $w_1 = \bar{w}_1$. This neural network with $\bar{w}_1$ and $\bar{u}$ can be regarded as a forward neural network with (6.2), which is a kind of RBF network with a Gaussian kernel [7, 84]; in our case, however, we use the fundamental solution with the Dirichlet condition. We have a similar result to Lemma 3 for $\bar{w}_1$ as well.

Corollary 2 *For a certain $R' > 0$, k and $h_2 > 0$, we have a solution $\vec{w}_p$ to (6.13) that satisfies*

$$\|\vec{w}_p\| \le R' \quad (p = 1, 2, \dots, L),$$

for $\forall h \in (0, h_2]$. Thus, for $\bar{w}_1$ defined in (6.15), we have

$$\left\|\bar{w}_1(t, \cdot, \cdot)\right\|_{L_2(I\times I)} \le R', \quad t \in \left((N-1)k, T\right].$$

Remark 9 Regarding the mapping degree of a map between two spaces with different dimensions, one can refer to, for instance, the work of [55].

7 Proof of Theorem 3

Now, we present Lemma 5 below, which is crucial for the proof of the main theorem. It assures that we can make $\widetilde{U}_h^{(N)}$ and $\bar{u}(T)$ be sufficiently close if we take h and k small enough while maintaining some relationship. In its proof, we insert another variable $V_h^{(N)}$, with which we can prove the lemma by using the estimate of the fundamental solution of heat equation.

After proving Lemma 5, we can easily prove Theorem 3, by using Lemmas 3 and 4 as well. First, we note that we have the estimate of $\phi(\cdot)$ just we did in the proof of Theorem 2 right above (B.18) in Appendix B. Combining it with the *a-priori* estimate there, we can estimate the left-hand side from above in the form

$$\left\|\phi\left(\int_I w_1(\cdot, \cdot, y)\check{\nu}(\cdot, y)\,\mathrm{d}y + \int_I w_1(\cdot, \cdot, y)\,\mathrm{d}y\right)\right\|_{L_2(I_T)} \le c\big(|u_0|\big), \tag{7.1}$$

where $c(|u_0|)$ is a positive constant that depends on $|u_0|$.

Lemma 5 *$\{\widetilde{U}_h^{(n)}\}_n$ defined in (6.2) satisfies*:

$$\left\|\widetilde{U}_h^{(N)} - \bar{u}(T)\right\| \le d(k, h),$$

where $d(k, h)$ is a constant that is independent of w_1, and tends to 0 when k and h tend to 0 satisfying $h^2 = o(\log(\frac{T}{k})^{-1})$.

Proof The overview of the proof of this lemma goes as follows. Because $\bar{u}$ is a continuous variable, while $\widetilde{U}_h^{(N)}$ is a discretized one, we shall insert another variable $V_h^{(N)}$, and estimate both $\|V_h^{(N)} - \bar{u}(T)\|$ and $\|V_h^{(N)} - \widetilde{U}_h^{(N)}\|$, to obtain the desired result. For the first one, we shall estimate the accuracy of the discrete approximation of a continuous solution by its discretization. For the latter one, we shall make use of the property of Padé approximation.

On the one hand, based on Duhamel's principle, $\bar{u}$ satisfies the following:

$$\begin{aligned}\bar{u}(Nk, x) = Z * \bar{u}\big((N-1)k, \cdot\big) + \int_{(N-1)k}^{Nk} \mathrm{d}s \int_I Z(Nk - s, x, z) \\ \times \phi\left(\int_I \bar{w}_1(s, z, y)\bar{u}(s, y)\,\mathrm{d}y + \int_I \bar{w}_1(s, x, y)\,\mathrm{d}y\right)\mathrm{d}z.\end{aligned}$$

On the other hand, $\{\widetilde{U}_h^{(n)}\}_n$ defined in (6.2) satisfies the following:

$$\begin{aligned}\widetilde{U}_h^{(N)} &= r(kA_h)\overline{P}_h\bar{u}\big((N-1)k,\cdot\big)\\ &\quad + kr(kA_h)\left[P_h\phi\left(\sum_{l=1}^{L-2} w_{\cdot,l}^{(N-1)}\widetilde{U}_{h(l)}^{(N-1)} + h\sum_{l=1}^{L} w_{\cdot,l}^{(N-1)}\right)\right].\end{aligned}$$

We also consider

$$\begin{aligned}V_h^{(n)} &\equiv r(kA_h)^n P_h u_0\\ &\quad + k\sum_{j=1}^{n} r(kA_h)^{(n-j)}\\ &\quad \times P_h\left[\phi\left(\int_I \bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\right)\right].\end{aligned}$$

Note that $\bar{w}_1$ is a piecewise constant function from its construction, thus the right-hand side above makes sense. Recalling $T = Nk$, we first consider $\|\bar{u}(T) - V_h^{(N)}\|$. We have

$$\begin{aligned}\left\|\bar{u}(T) - V_h^{(N)}\right\| &= \left\|Z(Nk) * u_0 - r(kA_h)^N P_h u_0\right\|\\ &\quad + \left\|\int_0^{Nk}\mathrm{d}s\int_I Z(Nk-s,\cdot,z)\right.\\ &\quad \times \phi\left(\int_I \bar{w}_1(s,z,y)\bar{u}(s,y)\,\mathrm{d}y + \int_I \bar{w}_1(s,z,y)\,\mathrm{d}y\right)\mathrm{d}z\\ &\quad - k\sum_{j=1}^{N} r(kA_h)^{N-j}\\ &\quad \left.\times P_h\left[\phi\left(\int_I \bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\right)\right]\right\|.\end{aligned} \tag{7.2}$$

Based on the estimate presented by Fujita and Mizutani [21], we have

$$\left\|Z(Nk) * u_0 - r(kA_h)^N P_h u_0\right\| \le \frac{c_{81}(h^2+k)}{T}|u_0|, \tag{7.3}$$

where c_{81} is a positive constant. On the other hand, regarding the second term of the right-hand side of (7.2), we have

$$\begin{aligned}&\left\|\int_0^{Nk}\mathrm{d}s\int_I Z(Nk-s,\cdot,z)\right.\\ &\quad \times \phi\left(\int_I \bar{w}_1(s,z,y)\bar{u}(s,y)\,\mathrm{d}y + \int_I \bar{w}_1(s,z,y)\,\mathrm{d}y\right)\mathrm{d}z\\ &\quad - k\sum_{j=1}^{N} r(kA_h)^{N-j}\\ &\quad \left.\times P_h\left[\phi\left(\int_I \bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\right)\right]\right\|\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| \int_0^{(N-1)k} \mathrm{d}s \int_I Z(Nk-s,\cdot,z) \right. \\
&\times \phi\left(\int_I \bar{w}_1(s,z,y)\bar{u}(s,y)\,\mathrm{d}y + \int_I \bar{w}_1(s,z,y)\,\mathrm{d}y \right) \mathrm{d}z \\
&- k \sum_{j=1}^{N-1} \int_I Z\big((N-j)k,\cdot,z\big) \\
&\left. \times \phi\left(\int_I \bar{w}_1(jk,z,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,z,y)\,\mathrm{d}y \right) \mathrm{d}z \right\| \\
&+ \left\| \int_{(N-1)k}^{Nk} \mathrm{d}s \int_I Z(Nk-s,\cdot,z) \right. \\
&\times \phi\left(\int_I \bar{w}_1(s,z,y)\bar{u}(s,y)\,\mathrm{d}y + \int_I \bar{w}_1(s,z,y)\,\mathrm{d}y \right) \mathrm{d}z \\
&\left. - kP_h\left[\phi\left(\int_I \bar{w}_1(Nk,\cdot,y)\bar{u}(Nk,y)\,\mathrm{d}y + \int_I \bar{w}_1(Nk,\cdot,y)\,\mathrm{d}y \right) \right] \right\| \\
&+ k \left\| \sum_{j=1}^{N-1} \int_I Z\big((N-j)k,\cdot,z\big) \right. \\
&\times \phi\left(\int_I \bar{w}_1(jk,z,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,z,y)\,\mathrm{d}y \right) \mathrm{d}z \\
&- \sum_{j=1}^{N-1} r(kA_h)^{N-j} \\
&\left. \times P_h\left[\phi\left(\int_I \bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,\cdot,y)\,\mathrm{d}y \right) \right] \right\| \\
&\equiv \sum_{j=1}^{3} J_j.
\end{aligned}
$$

Regarding the estimate of J_1, let us recall (6.15). Then, following the direction of Hoff and Smoller [30], we have

$$
\begin{aligned}
J_1 &\leq k\phi(0) \int_0^{(N-1)k} \mathrm{d}s \int_I \frac{\partial Z}{\partial t}(Nk-s,x,z)\,\mathrm{d}z \\
&\leq k\phi(0) \int_0^{(N-1)k} (Nk-s)^{-\frac{3}{2}}\,\mathrm{d}s \int_I e^{-\frac{(x-z)^2}{Nk-s}}\,\mathrm{d}z \\
&\leq 2k\phi(0)\left(\frac{1}{\sqrt{T}} - \frac{1}{\sqrt{k}} \right).
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
J_2 \leq \left\| \int_{(N-1)k}^{Nk} \mathrm{d}s \int_I Z(Nk-s,\cdot,z) \right. \\
\left. \times \phi\left(\int_I \bar{w}_1(s,z,y)\bar{u}(s,y)\,\mathrm{d}y + \int_I \bar{w}_1(s,z,y)\,\mathrm{d}y \right) \mathrm{d}z \right.
\end{aligned}
$$

$$
\begin{aligned}
&- k\phi\left(\int_I \bar{w}_1(T,\cdot,y)\bar{u}(T,y)\,\mathrm{d}y + \int_I \bar{w}_1(T,\cdot,y)\,\mathrm{d}y\right)\Bigg\| \\
&+ \Bigg\| k\phi\left(\int_I \bar{w}_1(T,\cdot,y)\bar{u}(T,y)\,\mathrm{d}y + \int_I \bar{w}_1(T,\cdot,y)\,\mathrm{d}y\right) \\
&- kP_h\left[\phi\left(\int_I \bar{w}_1(T,\cdot,y)\bar{u}(T,y)\,\mathrm{d}y + \int_I \bar{w}_1(T,\cdot,y)\,\mathrm{d}y\right)\right]\Bigg\| \\
&= k\Bigg\|\frac{1}{k}\int_{(N-1)k}^{Nk}\mathrm{d}s\int_I Z(Nk-s,x,z) \\
&\times \phi\left(\sum_l w_{k,l}^{(N)}\int_{I_l}\bar{u}(s,y)\,\mathrm{d}y + \sum_{l=1}^{L} w_{k,l}^{(N)}\right)\mathrm{d}z \\
&- \phi\left(\sum_l w_{\cdot,l}^{(N)}\int_{I_l}\bar{u}(T,y)\,\mathrm{d}y + \sum_{l=1}^{L} w_{k,l}^{(N)}\right)\Bigg\| + c\big(|u_0|\big)k.
\end{aligned}
$$

Note that for an integrable function $f(s)$ of s, we have

$$
\frac{1}{k}\int_{(N-1)k}^{Nk} f(s)\,\mathrm{d}s = f(Nk) + o(k).
$$

Thus, we have

$$
\begin{aligned}
&\frac{1}{k}\int_{(N-1)k}^{Nk}\mathrm{d}s\int_I Z(Nk-s,x,z)\phi\left(\sum_l w_{\cdot,l}^{(N-1)}\int_{I_l}\bar{u}(T,y)\,\mathrm{d}y + \sum_{l=1}^{L} w_{k,l}^{(N-1)}\right)\mathrm{d}z \\
&\quad = \phi\left(\sum_l w_{\cdot,l}^{(N-1)}\int_{I_l}\bar{u}(T,y)\,\mathrm{d}y + \sum_{l=1}^{L} w_{k,l}^{(N-1)}\right) + o(k),
\end{aligned}
$$

which yields $|J_2| \le c_{82}(k + ko(k))$ with some $c_{82} > 0$. As for the estimate of J_3, applying (7.3) again together with (7.1) and Corollary 2, we have

$$
\begin{aligned}
\|J_3\| &= k\Bigg\|\sum_{j=1}^{N-1}\Bigg\{\int_I Z\big((N-j)k,\cdot,z\big) \\
&\quad \times \phi\left(\int_I \bar{w}_1(jk,z,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,z,y)\,\mathrm{d}y\right)\mathrm{d}z \\
&\quad - r(kA_h)^{N-j} \\
&\quad \times P_h\left(\phi\left(\int_I \bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\right)\right)\Bigg\}\Bigg\| \\
&\le c_{81}\big(h^2+k\big)\sum_{j=1}^{N-1}\frac{1}{(N-j)}\left|\phi\left(\int_I \bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\right)\right|.
\end{aligned}
$$

Recalling (6.15), the right-most hand side is estimated by

$$\begin{aligned} &c_{81}\bigl|\phi(0)\bigr|\bigl(h^2+k\bigr)\sum_{j=1}^{N-1}\frac{1}{(N-j)} \\ &\quad= c_{81}\bigl|\phi(0)\bigr|\bigl(h^2+k\bigr)\biggl\{\log(N-1)+\frac{1}{2(N-1)}+\frac{1}{2}+\int_1^{N-1}\frac{P_1(t)}{t^2}\,\mathrm{d}t\biggr\}, \end{aligned}$$

where $P_1(t)=\{t\}-\frac{1}{2}$ with $\{x\}$ being the fractional part of its argument, and we have used the Euler–Maclaurin formula [38]. Combining these, under the assumption of the lemma, we arrive at the following:

$$\bigl\|\bar{u}(T)-V_h^{(N)}\bigr\|\le c(h,k), \tag{7.4}$$

where $c(h,k)\to 0$ as $h,k\to 0$ satisfying $h^2=o(\log(T/k)^{-1})$. Next, we estimate $\|V_h^{(N)}-\widetilde{U}_h^{(N)}\|$. Recall the following equalities.

$$\begin{aligned} \widetilde{U}_h^{(N-1)} &= \overline{P}_h e^{-(N-2)kA}u_0 \\ &\quad+k\overline{P}_h\sum_{j=1}^{N-1}e^{-(N-j-1)kA} \\ &\quad\times\biggl[\phi\biggl(\int_I\bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y+\int_I\bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\biggr)\biggr], \\ V_h^{(N-1)} &\equiv r(kA_h)^{N-1}P_hu_0 \\ &\quad+k\sum_{j=1}^{N-1}r(kA_h)^{N-j-1} \\ &\quad\times P_h\biggl[\phi\biggl(\int_I\bar{w}_1(jk,x,y)\bar{u}(jk,y)\,\mathrm{d}y+\int_I\bar{w}_1(jk,x,y)\,\mathrm{d}y\biggr)\biggr]. \end{aligned} \tag{7.5}$$

Thus, we have

$$\begin{aligned} \bigl\|\widetilde{U}_h^{(N-1)}-V_h^{(N-1)}\bigr\| &\le \bigl\|\overline{P}_he^{-(N-2)kA}u_0-r(kA_h)^{N-2}P_hu_0\bigr\| \\ &\quad+k\sum_{j=1}^{N-2}\Bigl\|\overline{P}_he^{-(N-j-1)kA} \\ &\quad\times\biggl[\phi\biggl(\int_I\bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y+\int_I\bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\biggr)\biggr] \\ &\quad-r(kA_h)^{N-j-1} \\ &\quad\times P_h\biggl[\phi\biggl(\int_I\bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y+\int_I\bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\biggr)\biggr]\Bigr\|. \end{aligned} \tag{7.6}$$

Regarding the first term of the right-hand side of (7.6), we have the following inequality.

$$\begin{aligned}&\left\|\overline{P}_h e^{-(N-2)kA}u_0 - r(kA_h)^{N-2}P_h u_0\right\| \\ &\quad\le \left\|\overline{P}_h e^{-(N-2)kA}u_0 - e^{-(N-2)kA}u_0\right\| \\ &\qquad + \left\|e^{-(N-2)kA}u_0 - r(kA_h)^{N-2}P_h u_0\right\|. \end{aligned} \tag{7.7}$$

Because N is sufficiently large, we have $e^{-(N-2)kA}u_0 \in H^2(I)$. For an arbitrary $\varepsilon_2 > 0$, if we take h sufficiently small, we can obtain

$$\|\overline{P}_h f - f\| < \varepsilon_2, \tag{7.8}$$

for a uniformly continuous function f in general. Moreover, we have

$$\left\|e^{-(N-2)kA}u_0 - r(kA_h)^{N-2}P_h u_0\right\| \le c_{81}\left(h^2 + k\right)|u_0|.$$

By combining this and (7.8), and applying to (7.7), we obtain the estimate

$$\left\|\overline{P}_h e^{-(N-2)kA}u_0 - r(kA_h)^{N-2}P_h u_0\right\| \le c\left(|u_0|\right)\left(h^2 + k\right) + \varepsilon_2.$$

Regarding the second term of the right-hand side of (7.6), we have

$$\begin{aligned}&k\sum_{j=1}^{N-2}\left\|\overline{P}_h e^{-(N-j-1)kA}\right. \\ &\qquad\times\left[\phi\left(\int_I \bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\right)\right] \\ &\qquad - r(kA_h)^{N-j-1} \\ &\qquad\left.\times P_h\left[\phi\left(\int_I \bar{w}_1(jk,\cdot,y)\bar{u}(jk,y)\,\mathrm{d}y + \int_I \bar{w}_1(jk,\cdot,y)\,\mathrm{d}y\right)\right]\right\| \\ &\quad\le k\phi(0)\left\{\sum_{j=1}^{N-2}\left\|\overline{P}_h e^{-(N-j-1)kA}1 - e^{-(N-j-1)kA}1\right\|\right. \\ &\qquad\left. + k\sum_{j=1}^{N-2}\left\|e^{-(N-j-1)kA}1 - r(kA_h)^{N-j-1}P_h 1\right\|\right\} \\ &\quad\le \varepsilon_2\phi(0)(T-2k) + c_{81}\phi(0)\sum_{j=1}^{N-2}\frac{h^2+k}{N-j-1} \\ &\quad\le \varepsilon_2 T\phi(0) + c_{81}\phi(0)\left(h^2+k\right)\left\{1 + \log\left(\frac{T}{k} - 2\right)\right\}. \end{aligned} \tag{7.9}$$

Thus, (7.6), (7.7), and (7.9) yield

$$\begin{aligned}\left\|\widetilde{U}_h^{(N-1)} - V_h^{(N-1)}\right\| &\le c\left(|u_0|\right)\left(h^2+k\right)\left\{1 + \log\left(\frac{T}{k} - 2\right)\right\} \\ &\quad + c_{84}\varepsilon_2 + \varepsilon_2 T\phi(0). \end{aligned} \tag{7.10}$$

Next, we proceed to the estimation of $\|\widetilde{U}_h^{(N)} - V_h^{(N)}\|$. Note that $V_h^{(N)}$ satisfies the following recurrence relation:

$$\begin{aligned} V_h^{(N)} &= r(kA_h)V_h^{(N-1)} \\ &\quad + kr(kA_h)P_h\Bigg[\phi\Bigg(\int_I \bar{w}_1(Nk,\cdot,y)\bar{u}(Nk,y)\,\mathrm{d}y \\ &\quad + \int_I \bar{w}_1(Nk,\cdot,y)\,\mathrm{d}y\Bigg)\Bigg]. \end{aligned}$$

Then, by using (6.2), we observe that

$$\begin{aligned} \big\|\widetilde{U}_h^{(N)} - V_h^{(N)}\big\| &= r(kA_h)\big(\widetilde{U}_h^{(N-1)} - V_h^{(N-1)}\big) \\ &\quad + kr(kA_h)P_h\Bigg[\phi\Bigg(h\sum_{l=2}^{L-1} w_{\cdot,l}^{(N-1)}\widetilde{U}_{h(l)}^{(N-1)} + h\sum_{l=1}^{L} w_{\cdot,l}^{(N-1)}\Bigg) \\ &\quad - \phi\Bigg(\int_I \bar{w}_1(Nk,\cdot,y)\bar{u}(Nk,y)\,\mathrm{d}y \\ &\quad + \int_I \bar{w}_1(Nk,\cdot,y)\,\mathrm{d}y\Bigg)\Bigg]. \end{aligned} \tag{7.11}$$

Recalling (6.15), we have that $h\sum_{l=1}^{L} w_{\cdot,l}^{(N-1)} = \int_I \bar{w}_1(Nk,\cdot,y)\,\mathrm{d}y$, and

$$\begin{aligned} \int_I \bar{w}_1\big((N-1)k,x,y\big)\,\mathrm{d}y &= \sum_{l=1}^{L} h\int_{I_l} \bar{w}_1\big((N-1)k,x,y\big)\,\mathrm{d}y \\ &= h\sum_{l=1}^{L} w_{p,l}^{(N-1)} \quad (x\in I_p). \end{aligned}$$

Similarly, if we recall the definitions of $\overline{P}_h$, we have

$$\widetilde{U}_{h(l)}^{(N-1)} = \big[\overline{P}_h\bar{u}\big((N-1)k,\cdot\big)\big]_l = \frac{1}{h}\int_{I_l} \bar{u}\big((N-1)k,y\big)\,\mathrm{d}y.$$

Thus, we have

$$\begin{aligned} &\phi\Bigg(h\sum_{l=2}^{L-1} w_{\cdot,l}^{(N-1)}\widetilde{U}_{h(l)}^{(N-1)} + h\sum_{l=1}^{L} w_{\cdot,l}^{(N-1)}\Bigg) \\ &\quad = \phi\Bigg(\int_I \bar{w}_1((Nk,\cdot,y)\bar{u}\Bigg((Nk,y)\,\mathrm{d}y + \int_I \bar{w}_1(Nk,\cdot,y)\,\mathrm{d}y\Bigg), \end{aligned}$$

which implies that the second term of (7.11) vanishes. Thus, owing to this and (7.10), we can estimate (7.11) as shown below.

$$\begin{aligned} \big\|r(kA_h)\big(\widetilde{U}_h^{(N-1)} - V_h^{(N-1)}\big)\big\| &\le c\big(|u_0|\big)\big(h^2+k\big)\Bigg\{1+\log\Bigg(\frac{T}{k}-2\Bigg)\Bigg\} \\ &\quad + c_{84}\varepsilon_2 + \varepsilon_2 T\phi(0). \end{aligned}$$

Regarding the second term on the right-hand side of (7.11), let us recall (6.15) and the definition of $\overline{P}_h$ to obtain:

$$\begin{aligned}\int_I \bar{w}_1\big((N-1)k,x,y\big)\,\mathrm{d}y &= \sum_{l=1}^{L} h \int_{I_l} \bar{w}_1\big((N-1)k,x,y\big)\,\mathrm{d}y \\ &= h\sum_{l=1}^{L} w_{p,l}^{(N-1)} \quad (x\in I_p).\end{aligned}$$

Similarly, if we recall the definitions of $\overline{P}_h$, $\widetilde{U}_{h(l)}^{(N-1)}$ and

$$\widetilde{U}_{h(l)}^{(N-1)} = \big[\overline{P}_h \bar{u}\big((N-1)k,\cdot\big)\big]_l = \frac{1}{h}\int_{I_l} \bar{u}\big((N-1)k,y\big)\,\mathrm{d}y,$$

we observe

$$h\sum_{l=2}^{L-1} w_{p,l}^{(N-1)}\widetilde{U}_{h(l)}^{(N-1)} = \int_{I_l} \bar{u}\big((N-1)k,y\big)\,\mathrm{d}y.$$

Thus, we arrive at the estimate:

$$\big\|\widetilde{U}_h^{(N)} - V_h^{(N)}\big\| \le c\big(|u_0|\big)\big(h^2+k\big)\left\{1+\log\left(\frac{T}{k}-2\right)\right\} + \varepsilon_2 + \varepsilon_2 T\phi(0).$$

Finally, by combining (7.2) and above, we arrive at the desired inequality of Lemma 5 because $\varepsilon_2 > 0$ is arbitrary. □

Owing to Lemmas 3, 4, and 5, if the spatio-temporal mesh is sufficiently fine and the relationship $h^2 = o(\log(\frac{T}{k})^{-1})$ is satisfied, we can approximate the solution u of (3.3) with the fully discretized one, for example, $((\triangle x)_i,(\triangle t)_i)$, regardless of w_1. This leads us to the proof of Theorem 3 in Sect. 4. Actually, owing to Lemma 4, we can assume that $\|\vec{w}'_{0(h,k)}\| \le R$ with some $R > 0$. Then, we have

$$\begin{aligned}\left|F(\xi) - \int_I \bar{w}_0(x)\bar{u}(T,x;\xi)\,\mathrm{d}x\right| &\le \big|F(\xi) - \vec{w}_{0(h,k)}^{\prime\top}\widetilde{U}_h^{(N)}\big| \\ &\quad + \left|\vec{w}_{0(h,k)}^{\prime\top}\widetilde{U}_h^{(N)} - \int_I \bar{w}_0(x)\bar{u}(T,x;\xi)\,\mathrm{d}x\right| \\ &\le \frac{\varepsilon}{2} + Rc(h,k). \end{aligned} \tag{7.12}$$

Thus, if h and k are set to have sufficiently small values maintaining the relationship $h^2 = o(\log(\frac{T}{k})^{-1})$, the right-hand side of (7.12) can be less than ε. This proves Theorem 3.

Remark 10 The estimate (7.12) above is observed for a fixed value of ν. Actually, the estimate (7.3) in Fujita and Mizutani [21] is obtained by assuming $\nu = 1$. In the general case, let us introduce a transform $\bar{x} = x/\sqrt{\nu}$. Then, a problem

$$u_t - \nu u_{xx} = 0 \quad \text{in } I_T,$$

with an initial value $u_0(x)$ for a function $u(t,x)$ is transformed into the form:

$$\bar{u}_t - \bar{u}_{\tilde{x}\tilde{x}} = 0 \quad \text{in } (0, 1/\sqrt{\nu}) \times (0, T),$$

for a function $\bar{u}(t,\tilde{x})$ with an initial value $\bar{u}_0(\tilde{x}) = u_0(x)$. We can easily find that

$$\|\bar{u}_0\|^2_{L_2(0,1/\sqrt{\nu})} = \frac{1}{\sqrt{\nu}} \|u_0\|^2_{L_2(I)},$$

which means that the right-hand side of (7.12) diverges in case $\nu \to 0$. Therefore, we leave the discussion concerning the convergence of the universal approximation property we have proved here when $\nu \to 0$ as an open problem.

8 Capacity and learnability of the model

In the proof of universal approximation property, we fixed the values of w_1 up to the time right before the terminal moment. However, this does not mean that the temporal direction is not necessary in our model. Universal approximation property is not the only property that a learner should possess; indeed, the learnability and generalization performance are also important. In this section, we will observe that our model possesses the learnability in some sense. In doing so, we will also observe that the estimations to deduce the learnability depend on time.

In this regard, we will discuss other aspects of the proposed model in the sequel. First, we will address its learnability, specifically focusing on classification performance metrics or classes of functions, such as the VC-dimension and Glivenko-Cantelli class. Our discussion is limited to binary classifications. Although we discuss VC-dimension, we delegate its definition to some monographs [8, 65]. Finally, we present the results of our numerical experiments.

8.1 Learnability

In this section, we discuss the learnability of the proposed model. Hereafter, we will often denote the solution to (3.3) (or equivalently, (3.1)–(3.2)) as $u(t,x;w_1,\vec{\xi},\nu)$ for clearly indicating its dependence on w_1, $\vec{\xi}$, and ν. Then, given the terminal moment T and diffusion coefficient ν, we define a hypothesis set. This set comprises functions on $\mathbb{R}^J$ realized by our model:

$$\mathscr{F}_T^{(\nu)} \equiv \left\{ \vec{\xi} \longmapsto \int_I w_0(x) u(T,x;w_1\vec{\xi},\nu)\,\mathrm{d}x \,\middle|\, w_0 \in L_2(I), w_1 \in L_2(\mathcal{H}_T) \right\}. \tag{8.1}$$

Let us discuss the learnability of $\mathscr{F}_T^{(\nu)}$. In the following, the VC-dimension of a hypothesis set $\mathscr{F}$ is denoted as $VC(\mathscr{F})$. Our first result is

Theorem 4 *Suppose that the assumptions of Theorem 3 are satisfied. Let T and ν be arbitrary positive numbers. Subsequently, for our proposed PDE-based neural network,*

$$VC\big(\mathscr{F}_T^{(\nu)}\big) = +\infty.$$

Proof Suppose that we are given an arbitrary $N \in \mathbb{N}$ and a dataset $\{\vec{\xi}_i, y_i\}_{i=1}^N \in \mathbb{R}^J \times \{\pm 1\}$. Then, let us take $\varepsilon > 0$ so that $B(\vec{\xi}_i;\varepsilon) \cap B(\vec{\xi}_i;\varepsilon) = \phi (i \neq j)$. In virtue of Theorem 3, by suitably

taking w_0 and w_1, we can make a continuous function $f \in \mathscr{F}_T^{(\nu)}$ which associates each element in $B(\vec{\xi}_i;\varepsilon)$ with y_i for all $i = 1, 2, \ldots, N$. This means that the set $\mathscr{F}_T^{(\nu)}$ shatters the given dataset with an arbitrary $N \in \mathbb{N}$. □

Theorem 4 also implies that we require an infinite amount of training data, which is practically impossible, and that our model is not PAC-learnable in the classical sense [65]. However, using the concept of a structural risk minimization (SRM) scheme, we can still make it nonuniformly learnable [65]. A relaxation of the concept of learnability of this kind has also been applied to support vector machines [8].

To discuss this in more detail, we introduce certain notations. In general, the "risk" over a loss function $l(\cdot)$ and a general hypothesis set $\mathscr{F}$ is defined by the following:

$$L_D(h) = E_{z\sim\mathcal{D}}\big[l(h;z)\big], \tag{8.2}$$

where $\mathcal{D}$ is an unknown data-generating distribution defined as follows: $\mathcal{Z} \equiv \mathcal{X} \times \{\pm 1\}$, with $\mathcal{X}$ being a set of inputs. The notation $z \sim \mathcal{D}$ means that a random variable z is drawn from $\mathcal{D}$. Similarly, we use the notation $S \sim \mathcal{D}^m$ to denote that a dataset S of sample size m is i.i.d. drawn from $\mathcal{D}$. If some $h \in \mathscr{F}$ attains (8.2), we call it a *Bayesian hypothesis*. However, we usually do not know the actual distribution $\mathcal{D}$. For this reason, we usually try to minimize the surrogate quantity, which is called the *empirical risk*:

$$L_S(h) \equiv \frac{1}{m}\sum_{i=1}^{m} l(h;\vec{\xi}_i, y_i),$$

where $S = \{(\vec{\xi}_i, y_i)\}_{i=1}^m \subset \mathcal{Z}$ represents the training data drawn from the original unknown distribution $\mathcal{D}$. This framework is called *empirical risk minimization* (ERM). Utilizing the law of large numbers, $L_S(h)$ converges to the true risk as $m \to +\infty$ for each h. We also define

$$\hat{h}_S = \mathrm{ERM}_{\mathscr{F}}(S) \in \operatorname*{argmin}_{h\in\mathscr{F}} L_S(h),$$

where $\mathrm{ERM}_{\mathscr{F}}(S)$ denotes a hypothesis returned as (one of) the minimizer(s) of the empirical risk under training dataset S.

To evaluate the "goodness" of the training data, we define the following concept.

Definition 1 A training set S is called *ε-representative* with respect to the domain $\mathcal{Z} \equiv \mathcal{X} \times \{\pm 1\}$, hypothesis set $\mathscr{F}$, loss function $l(\cdot)$, and distribution $\mathcal{D}$ if the following holds.

$$\big|L_S(h) - L_{\mathcal{D}}(h)\big| \le \varepsilon \quad \forall h \in \mathscr{F}.$$

To determine the conditions under which the ERM scheme works well, we need the following definition (please refer to [65], Definition 4.3).

Definition 2 We say that a hypothesis set $\mathscr{F}$ possesses the *uniform convergence property* with respect to the domain $\mathcal{Z}$ and loss function $l(\cdot)$ if there exists a function $m_{\mathscr{F}}^{UC} : (0,1)^2 \to \mathbb{N}$, which is called the *sample complexity*, such that for each $\varepsilon, \delta \in (0,1)$ and for

every probability distribution $\mathcal{D}$ over $\mathcal{Z}$, if S is a sample of $m \geq m_{\mathscr{F}}^{UC}(\varepsilon, \delta)$ elements that are drawn i.i.d. according to $\mathcal{D}$, then, with a probability of at least $1 - \delta$, S is ε-representative.

A well-known theorem states that (see [65], Theorem 6.7) the uniform convergence property is equivalent to the fact that the VC-dimension of the hypothesis set is finite. Thus, together with Theorem 4 above, our hypothesis set $\mathscr{F}_T^{(v)}$ does not satisfy the uniform convergence property itself (consequently, neither PAC nor agnostic PAC is learnable, although we omit the definitions of these terms here). However, we can also consider a relaxed concept of learnability [65].

Definition 3 A hypothesis set $\mathscr{F}$ is said to be *non-uniformly learnable* if there exists a learning algorithm A that associates a dataset S with a hypothesis $A(S) \in \mathscr{F}$ and a function $m_{\mathscr{F}} : (0,1)^2 \times \mathscr{F} \to \mathbb{N}$, such that for every $\varepsilon, \delta \in (0,1)$, and for every $h \in \mathscr{F}$, if $m \geq m_{\mathscr{F}}(\varepsilon, \delta, h)$ then for every distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, with a probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$, it is ensured that

$$L_{\mathcal{D}}\big(A(S)\big) \leq L_{\mathcal{D}}(h) + \varepsilon.$$

The following theorem [65] describes an important characterization of nonuniform learnability.

Theorem 5 *Let $\mathscr{F}$ be a hypothesis set that can be written as a countable union of the individual hypothesis sets.*

$$\mathscr{F} = \bigcup_{n \in \mathbb{N}} \mathscr{F}_n,$$

where each $\mathscr{F}_n$ exhibits a uniform convergence property. Then, $\mathscr{F}$ is nonuniformly learnable.

Returning to our specific case, we can show that our hypothesis set $\mathscr{F}_T^{(v)}$ defined in (8.1) *is* nonuniformly learnable. To demonstrate this, we will introduce a sequence of hypothesis sets.

$$\mathscr{F}_T^{(v)}(n) \equiv \left\{ \vec{\xi} \longmapsto \int_I w_0(x) u(T, x; w_1, \cdot, v)\, \mathrm{d}x \,\middle|\, \|w_0\|_{L_2(I)}, \|w_1\|_{L_2(\mathcal{H}_T)} \leq n \right\}$$
$$(n = 1, 2, \ldots). \tag{8.3}$$

Evidently, these sets form the following relationships.

$$\mathscr{F}_T^{(v)}(1) \subset \mathscr{F}_T^{(v)}(2) \subset \ldots,$$
$$\mathscr{F}_T^{(v)} = \bigcup_{n=1}^{\infty} \mathscr{F}_T^{(v)}(n) \quad \forall T, v > 0. \tag{8.4}$$

Next, we demonstrate that each set $\mathscr{F}_T^{(v)}(n)$ in (8.4) satisfies uniform convergence property. We also use the notations

$$\begin{aligned}\mathscr{L}(n) &\equiv \tilde{l} \circ \mathscr{F}_T^{(v)}(n) \\ &= \left\{ (\vec{\xi}, y) \longmapsto \tilde{l}\left(\int w_0(x) u(T, x; w_1, \vec{\xi}, v)\, \mathrm{d}x, y \right) \middle| \, \|w_0\|_{L_2(I)}, \|w_1\|_{L_2(\mathcal{H}_T)} \leq n \right\} \\ &\quad (n = 1, 2, \ldots).\end{aligned}$$

To assess the uniform convergence property of $\mathscr{F}_T^{(v)}(n)$ with respect to the loss function $\tilde{l}(\cdot)$, it is necessary and sufficient to check that the set $\mathscr{L}(n)$ is a Glivenko–Cantelli class [65].

Hereafter, we denote a probability space as $(\Omega, \mathscr{A}, P)$, where Ω is the sample space, $\mathscr{A}$, the σ-algebra with respect to probability measure P. We also denote a corresponding empirical measure as $P_m(A) = \frac{1}{m} \sum_{j=1}^{m} \delta_{\vec{\xi}_j}(A)$ for a Borel set A with $\delta(\cdot)$ being Dirac measure, and define

$$Pf = \int_\Omega f \, \mathrm{d}P, \qquad \|P_m - P\|_{\mathscr{F}} \equiv \sup_{f \in \mathscr{F}} \sqrt{m} |P_m f - Pf|.$$

Definition 4 Given a probability space $(\Omega, \mathscr{A}, P)$ and a set of integrable real-valued functions $\mathscr{F}$, we say that $\mathscr{F}$ is a Glivenko–Cantelli class for P if and only if

$$\|P_m - P\|_{\mathscr{F}} \to 0 \quad (m \to +\infty)$$

holds almost uniformly.

In case of binary classification, being a Glivenko–Cantelli class is equivalent to satisfying the uniform convergence property [65]. Moreover, the following theorem is known [16]. Here, $I^d = [0,1]^d$ with $d \in \mathbb{N}$.

Theorem 6 *Let $K > 0$ and $\mathscr{F}_{1,K}(I^d)$ be a set of the Lipschitz continuous functions on I^d:*

$$\mathscr{F}_{1,K}(I^d) = \left\{ f \in C(I^d) \middle| \sup_x |f(x)| + \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|_{\mathbb{R}^d}} \leq K \right\}.$$

Then, $\mathscr{F}_{1,K}(I^d)$ is a Glivenko–Cantelli class for any probability measure P on I^d.

Thus, if we impose Lipschitz continuity on the loss function, we can guarantee that the set $\mathscr{L}(n)$ becomes a Glivenko–Cantelli class for each n.

Theorem 7 *Suppose that the assumptions of Theorem 3 hold. Let $T > 0$ be arbitrary, $\mathcal{Z} = \mathcal{X} \times \pm 1$ with $\mathcal{X} \subset \mathbb{R}^J$ being compact, and a loss function $l(\cdot) : \mathcal{Z} \times L_2(I) \times L_2(\mathcal{H}_T) \to \mathbb{R}$ of the form*

$$l\big((\vec{\xi}, y), w_0, w_1\big) = \tilde{l}\left(\int_I w_0(x) u(T, x; w_1, \vec{\xi}, v)\, \mathrm{d}x, y \right)$$

with a function $\tilde{l}(a, y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ being Lipschitz continuous with respect to (a, y) with Lipschitz coefficient L. Then, the set $\mathscr{L}(n)$ is a Glivenko–Cantelli class.

Proof The following is a simple denotation: $\|w_0\|_{L_2(I)}$, $\|w_1\|_{L_2(\mathcal{H}_T)}$ by $|w_0|$, $|w_1|$, respectively. Without losing generality, we can assume that $\mathcal{X} = I^J$. Under the assumptions of the theorem, we have

$$\begin{aligned}&\left|l\left((\vec{\xi}_1, y_1), w_0, w_1\right) - l\left((\vec{\xi}_2, y_2), w_0, w_1\right)\right| \\ &\quad \leq L\left\{|y_1 - y_2| + \left|\int_I w_0(x)u(t,x;w_1,\vec{\xi}_1,\nu) - \int_I w_0(x)u(t,x;w_1,\vec{\xi}_2,\nu)\right|\right\}. \end{aligned} \tag{8.5}$$

In order to verify the continuity of $u(T,x;w_1,\vec{\xi},\nu)$ with respect to $\vec{\xi}$, we appeal to a standard energy estimate. Let us denote $u(t,x;w_1,\vec{\xi}_i,\nu)$ $(i = 1,2)$ by $u_i(t,x)$ and $\tilde{u}(t,x) \equiv u_1(t,x) - u_2(t,x)$. Then, we have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left|\tilde{u}(t,\cdot)\right|^2 + \frac{\nu}{2}\left|\nabla\tilde{u}(t,\cdot)\right|^2 \leq L\left\|w_1(t,\cdot,\cdot)\right\|_{L_2(I\times I)}\left|\tilde{u}(t,\cdot)\right|^2, \quad t \in (0,T].$$

By the Gronwall's inequality, we obtain [74]

$$\left|\tilde{u}(T,\cdot)\right|^2 \leq \left|\tilde{u}(0,\cdot)\right|^2 e^{n\sqrt{T}}. \tag{8.6}$$

By noting $|u(0;\vec{\xi}_i)|^2 = \frac{1}{J}\|\vec{\xi}_i\|^2_{\mathbb{R}^J}$, and consequently, $|\tilde{u}(0)|^2 = \frac{1}{J}\|\vec{\xi}_1 - \vec{\xi}_2\|^2_{\mathbb{R}^J}$, and combining (8.5) and (8.6), we obtain

$$\left|l(\vec{\xi}_1, y_1) - l(\vec{\xi}_2, y_2)\right| \leq L\left(1 + \frac{nLe^{\frac{n\sqrt{T}}{2}}}{\sqrt{J}}\right)\left(|y_1 - y_2| + \|\vec{\xi}_1 - \vec{\xi}_2\|_{\mathbb{R}^J}\right).$$

By Theorem 6, this implies that $\mathcal{L}(n)$ forms a Glivenko–Cantelli class. □

Theorem 7 implies that our model achieves uniform convergence property of the hypothesis set under the boundedness of $\|w_0\|_{L_2(I)}$ and $\|w_1\|_{L_2(\mathcal{H}_T)}$ and the compactness of the input space $\mathcal{X}$ with which $\mathcal{D}$ is defined. Thus, for each $n \in \mathbb{N}$, we establish that $\mathscr{F}_T^{(\nu)}(n)$ has a uniform convergence property with respect to this $l(\cdot)$ and $\mathcal{D}$.

Before introducing another theorem, let us present a known lemma [80] concerning the covering number $N(\cdot)$ and bracketing number $N_{[]}(\cdot)$. We delegate the definitions of these quantities to other references (see, for instance, [16, 25, 80]).

Lemma 6 *Let $\mathcal{F} = \{f_t | t \in \mathcal{T}\}$ be a class of functions defined on a set $\mathcal{X}$ satisfying Lipschitz continuity in the index parameter*:

$$\left|f_s(x) - f_t(x)\right| \leq d(s,t)F(x) \quad \forall x \in \mathcal{X}, \forall s,t \in \mathcal{T}, \tag{8.7}$$

for some fixed function $F(\cdot)$, where $d(\cdot,\cdot)$ is a metric in the index space $\mathcal{T}$. Then, for any norm $\|\cdot\|$, $N_{[]}(2\varepsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\varepsilon, \mathcal{T}, d)$.

We also introduce the following lemma concerning the metric entropy of a set of functions.

Lemma 7 *Let $B_M \equiv \{u \in H^1(I) | \|u\|_{H^1(I)} \leq M\}$. Then, B_M is relatively compact in $L_2(I)$ and satisfies*

$$\log N\big(\varepsilon, B_M, L_2(I)\big) \leq \frac{KM}{\varepsilon} \quad \forall \varepsilon > 0,$$

where K is a constant.

Proof This lemma can be proved if we take $p = q = 2$ in Theorem 4.3.36 of [25] and note that the inclusion of function spaces $H^1(I) \subset B^{1,W}_{2\infty}(I)$, where $B^{1,W}_{2\infty}(I)$ is the Besov space defined in [25]. □

In the optimization procedure, it is often the case that w_0 is determined depending on w_1, and consequently $u(T,x;w_1)$. Based on Lemmas 6 and 7, we can assert the following theorem.

Theorem 8 *Under the assumptions of Theorem 7, suppose that $\tilde{l}(a,y)$ is Lipschitz continuous with respect to its first argument a and w_0 can be determined as a functional of $u(T,x;w_1)$: $w_0 = w_0(u(T,x;w_1))$ and satisfies*

$$\big\| w_0\big(u(T,\cdot;w_1)\big) - w_0\big(u\big(T,\cdot;w_1'\big)\big)\big\|_{L_2(I)} \leq L_w \big\| u(T,\cdot;w_1) - u\big(T,\cdot;w_1'\big)\big\|_{L_2(I)},$$

with some $L_w > 0$. Then, the the set $\mathscr{L}(n)$ is a Glivenko–Cantelli class.

Proof Let us simply denote $w_0 = w_0(u(T,x;w_1))$ and $w_0(x)'(x) = w_0(u(T,x;w_1'))$. We first show that

$$\begin{aligned} &\left|\tilde{l}\left(\int_I w_0(x)u(T,x;w_1,\vec{\xi},\nu)\,\mathrm{d}x,y\right) - \tilde{l}\left(\int_I w_0'(x)u\big(T,x;w_1',\vec{\xi},\nu\big)\,\mathrm{d}x,y\right)\right| \\ &\quad \leq c_T^{(\nu)} \big\| u(T,\cdot;w_1,\vec{\xi},\nu) - u\big(T,\cdot;w_1',\vec{\xi},\nu\big)\big\|_{L_2(I)}, \end{aligned} \tag{8.8}$$

where $c_T^{(\nu)} > 0$ is some constant depending on T and ν. Here, we have used the assumption on w_0 as well as the assumption $\|w_0\|_{L_2(I)} \leq n$, and the boundedness of $u(T,\cdot)$, which can be derived as follows.

Applying the standard energy estimate to (3.3) yields the following equation:

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\big|u(t,\cdot;w_1,\vec{\xi},\nu)\big|^2 + \frac{\nu}{2}\big|\nabla u(t,\cdot;w_1,\vec{\xi},\nu)\big|^2 \\ &\quad \leq 2L\big\|w_1(t,\cdot,\cdot)\big\|_{L_2(I\times I)}\big|u(t)\big|^2 + \frac{1}{\nu}\big(2L\big\|w_1(t,\cdot,\cdot)\big\|_{L_2(I\times I)} + \sqrt{2}c_1\big)^2. \end{aligned}$$

By introducing the notation $c_1 = |\phi(0)|^2$, together with Gronwall's inequality, we obtain [74]

$$\begin{aligned} \big|u(T,\cdot;w_1,\vec{\xi},\nu)\big|^2 \leq{}& \left\{\big|u(0)\big|^2 + \frac{2}{\nu}\int_0^T \big(2L\big\|w_1(\tau,\cdot,\cdot)\big\|_{L_2(I\times I)} + \sqrt{2}c_1\big)^2\,\mathrm{d}\tau\right\} \\ &\times \exp\left(4L\int_0^T \big\|w_1(\tau,\cdot,\cdot)\big\|_{L_2(I\times I)}\,\mathrm{d}\tau\right). \end{aligned} \tag{8.9}$$

By noting $|u(0;\vec{\xi})|^2 = \frac{1}{J}\|\vec{\xi}\|^2_{\mathbb{R}^J}$, and together with (8.9), we obtain the following:

$$\left|u(T,\cdot;w_1,\vec{\xi},\nu)\right|^2 \leq \left\{\frac{1}{J}\|\vec{\xi}\|^2_{\mathbb{R}^J} + \frac{2}{\nu}(2nL + c_1\sqrt{2T})^2\right\}\exp\left(4nLT^{\frac{1}{2}}\right).$$

Moreover, we can estimate the right-hand side of (8.8) (we omit the procedure of this estimate, for it is quite similar to the deduction of (8.6)). This, combined with (8.8), implies that the assumption of Lemma 6 is satisfied if we regard $\mathcal{L}(n)$ as a set of functions indexed by a set of functions of the form $u(T,\cdot;w_1) \in H^1(I)$. Indeed, in this case, (8.7) holds, where $d(\cdot,\cdot)$ is $L_2(I)$-norm and $F(x)$ is a constant. Thus, Lemma 6 implies $N_{[]}(2\varepsilon c_T^{(\nu)}, \mathcal{L}(n), |\cdot|) \leq N(\varepsilon, \mathcal{B}^M_{H^1(I)}, \|\cdot\|_{L_2(I)}) \leq \frac{KM}{\varepsilon}$, where $\mathcal{B}^M_{H^1(I)}$ denotes a ball in $H^1(I)$ with radius M. Because a finite bracketing number implies that the function space is a Glivenko-Cntelli class, this completes the proof. □

Remark 11 Note that in the proof of Theorem 8, the estimate above depends on T and ν, which implies that the generalization performance may depend on them. As a special case, when $\phi(\cdot)$ is bounded, we obtain the following:

$$\left|u(T,\cdot;w_1,\vec{\xi},\nu)\right|^2 \leq e^{-\nu T}\|\vec{\xi}\|^2_{\mathbb{R}^J} + \nu^{-1}\left(1 - e^{-\nu T}\right),$$

which implies that the increase of T may lead to a smaller covering number.

We have seen that under some conditions, $\mathscr{L}(n)$ is a Glivenko–Cntelli class, and consequently, $\mathcal{F}_T^{(\nu)}(n)$ has a finite VC-dimension and sample complexity, say d_n and $m^{UC}_{\mathscr{F}_T^{(\nu)}(n)}(\epsilon,\delta)$, respectively. To examine nonuniform learnability of $\mathcal{F}_T^{(\nu)}$, let us consider

$$\varepsilon_n(m,\delta) = \min_{\varepsilon\in(0,1)}\left\{m^{UC}_{\mathscr{F}_T^{(\nu)}(n)}(\epsilon,\delta) \leq m\right\}.$$

Then, it clearly holds that for each $n \in \mathbb{N}$.

$$\left|L_D(h) - L_S(h)\right| \leq \varepsilon_n(m,\delta) \quad \forall h \in \mathscr{F}_T^{(\nu)}(n).$$

In addition, if we consider a family of functions $w(n) : \mathbb{N} \to [0,1]$ that satisfies $\sum_{n=1}^{\infty} w(n) \leq 1$, we have the following approach called *structural risk minimization* (SRM) (Algorithm 1) [65]:

Algorithm 1 SRM scheme

Require: Training dataset $\{(\vec{\xi}_j, y_j)\}_{j=1}^m \sim \mathcal{D}^m$, confidence δ, and a sequence $\{w(n)\}_n$ that satisfies $\sum_n w(n) \leq 1$

Ensure: $h \in \operatorname{argmin}_{h\in\mathscr{F}}[L_S(h) + \varepsilon_{n(h)}(m,\delta w(n(h)))]$

$n = 1$

while $n < N'$ **do**

$n \Leftarrow n + 1$

The value of the loss function can be determined using the current parameter values.

$\varepsilon_n(m,\delta) \leftarrow \min_{\varepsilon\in(0,1)}\{m^{UC}_{\mathscr{F}_T^{(\nu)}(n)}(\epsilon,\delta) \leq m\}$

end while

Theorem 9 *Let $\mathscr{F}$ be a hypothesis class, such that $\mathscr{F} = \bigcup_n \mathscr{F}_n$, where each $\mathscr{F}_n$ has uniform convergence property with sample complexity $m^{UC}_{\mathscr{F}_n}$. Let $w : \mathbb{N} \to [0,1]$ be defined as $w(n) = 6/n^2\pi^2$. Then, $\mathscr{F}$ becomes nonuniformly learnable using the SRM scheme at a rate*

$$m^{NUC}_{\mathscr{F}}(\varepsilon, \delta, h) \leq m^{UC}_{\mathscr{F}_n}\left(\frac{\varepsilon}{2}, \frac{6\delta}{(\pi n(h))^2}\right).$$

Theorem 9 with $\mathscr{F}_n$ replaced by $\tilde{l} \circ \mathscr{F}^{(v)}_T(n)$ guarantees that our PDE-based neural network has nonuniform learnability.

8.2 Numerical computation

Finally, we conducted some numerical experiments to evaluate the performance of our model using practical datasets. Because the main focus of the present paper is the theoretical argument, this is the first example to check the effectiveness of our model. In the following section, we first clarify the setting of our numerical experiment and then state the results.

8.2.1 Settings

In this experiment, we focused exclusively on binary classification. The proposed model was implemented using Python 3.7 on a Windows Server 2019 (64 bits), 12th Gen Intel (R) Core (TM) i7-12700, 2.11 GHz, RAM 96.0 GB. In this experiment, we used the time difference $\triangle t = 5 \times 10^{-4}$ and a range of values for the number of temporal and spatial grids, denoted as N and L, respectively. At the output layer, we employed a logistic regression scheme with L_1 regularization using statsmodels [71]. The optimization of w_1 in our model involved optimizing the values $w_1(i_1, i_2, i_3)$ ($i_1 = 1, 2, \ldots, N$, $i_2, i_3 = 1, 2, \ldots, L$), each of which is a temporally discretized version of $w_1(t, x, y)$. Optimization was conducted using a genetic algorithm with the deap library [13] in Python.

8.2.2 Datasets

Numerical simulations are conducted with "adult income" [5] and "diabetes" [15] datasets, which are well-known benchmarks of binary classification.

The former dataset contains 121 adult attributes and their annual income. The task is to predict whether their income is larger than 50 thousands dollars (which corresponds to the label "1") or not ("0"). The latter dataset contained eight attributes with the human subjects and a binary label indicating whether each subject had the symptoms of diabetes.

Table 1 presents an overview of the datasets. For both datasets, we employed 70% of the training dataset, and the remaining part of the training dataset was used to check the test accuracy.

We applied min–max scaler, which transforms the values of each attribute onto the interval $[0, 1]$.

Table 1 Overview of datasets

	Adult income	Diabetes
Sample size (train)	34,189	375
Sample size (test)	14,653	162
Input dimension (J)	121	8

Table 2 Results of "adult income" dataset

	(ν, N, L)	Training accuracy / Test accuracy / AUC
Proposed method	(0.01, 2, 121)	0.858 / **0.860** / **0.915**
	(0.01, 5, 121)	0.858 / **0.860** / **0.915**
	(0.01, 3, 363)	0.758 / 0.788 / 0.854
	(0.01, 30, 121)	0.858 / 0.859 / **0.915**
Existing methods	SVC	0.764 / 0.764 / 0.770
	RFC	0.833 / 0.837 / 0.702
	LightGBM	**0.865** / 0.856 / 0.780
	XGBoost	0.864 / 0.857 / 0.774

Table 3 Results of "diabetes" dataset

	(ν, N, L)	Training accuracy / Test accuracy / AUC
Proposed method	(0.01, 2, 8)	0.758 / 0.788 / 0.854
	(0.01, 5, 8)	0.758 / 0.788 / 0.854
	(0.01, 3, 24)	0.823 / 0.801 / 0.899
	(0.01, 3, 80)	0.838 / **0.913** / **0.968**
	(0.01, 30, 8)	0.758 / 0.788 / 0.854
Existing methods	SVC	0.778 / 0.789 / 0.746
	RFC	0.762 / 0.792 / 0.757
	LightGBM	**0.957** / 0.775 / 0.746
	XGBoost	0.857 / 0.784 / 0.753

8.2.3 Results of experiments

Tables 2 and 3 show the results of the training and test accuracies, the area under the curve (AUC) (boldface indicates the largest value for each indicator) under a range of values T, and the number of points in the discretization of both spatial and temporal directions. The performance of the proposed method was comparable to that of the existing methods (Random Forest Classifier (RFC), Support Vector Classifier (SVC) with RFB kernel, XGBoost, and LightGBM) in terms of test accuracy and AUC. Note that in the existing methods, we tuned the hyperparameters by using cross-validation and grid-search.

The values of generations and population size in the genetic algorithm are 5 and 10, respectively, for "adult income" dataset, and 10 and 200 for 'diabetes' dataset. This is due to fact that the "adult income" dataset is larger and requires much longer computation time. From Tables 2 and 3, we observe that the performance of our model varies depending on the values of T.

In summary, the considerations in this section imply the following issues:

(i) Although our model achieves an infinite VC-dimension, it is still non-uniformly learnable under some assumptions about the underlying distribution behind the dataset. This property is also observed in some well-known machine learning algorithms, such as Support Vector Machine (SVM) with kernels.

(ii) By adjusting the parameters, we can adjust the generalization performance of our model. On the one hand, optimal values of the parameters yield a model with lower generalization error. On the other hand, this enlarges the search space during optimization, leading to the concern that we might not attain a (sub-)optimal solution within a realistic computation time. Therefore, in our future work, we will continue to search for an effective approach to optimize our model.

9 Conclusion

This study demonstrates the universal approximation property of our PDE-based neural network. It has been demonstrated that any continuous function on a compact set in $\mathbb{R}^J$ can be approximated by the output of a neural network with arbitrary precision.

We have also discussed the learnability of our model. Moreover, we implemented our model on a computer and performed certain numerical experiments. It showed a comparable performance to that of the existing models, such as RFC, SVC, LightGBM, and XGBoost. It was shown that the generalization performance could be adjusted by some parameters of the model. The exploration of more effective optimization procedures can be performed in the future.

Future work will consider the limit when ν tends to zero, in which case the proposed model could be considered the continuous limit of the usual neural network or one with an artificial diffusion term. Although we observed weak convergence of our solution, we should appeal to the theory of singular perturbation to factor in the boundary condition of a thin layer.

There is room for improvement in optimization procedure. We are planning to explore Bayesian optimization approaches that we have already attempted using ODE-based neural networks [33]. Therefore, it is important to discuss the PAC-Bayes perspective of the proposed model as well.

Additionally, we intend to extend our PDE-based neural network to multidimensional Euclidean spaces. As stated in Remark 5 at the end of Sect. 4, this is necessary when considering a GNN in which the elements are treated in the matrix form.

Appendix A: Summary of notations

We summarize the notations used in this paper, which are not presented in Sect. 2 in Table 4 below.

Appendix B: Proofs of existence

B.1 Proof of Theorem 1

Before introducing our first result, we shall define the Galerkin approximation [12].

Definition 5 Let V be a separable Hilbert space and $\{V_m\}_{m=1}^{\infty}$ be a family of finite dimensional vector spaces satisfying the assumptions (i) and (ii) below.

(i) $V_m \subset V$, $\dim V_m < +\infty$.

(ii) $V_m \to V$ $(m \to \infty)$ in the sense below: there exists a dense subspace of V, every element v of which has a corresponding sequence $\{v_m\}_{m=1}^{\infty} \subset V_m$ satisfying $\|v_m - v\|_V \to 0$ $(m \to +\infty)$.

Then, each space V_m $(m = 1, 2, \ldots)$ is called the Galerkin approximation of order m of V.

Table 4 Notations of function spaces and operators

$\mathfrak{W}(T)$	$\{u \mid u \in L_2(0,T;H^1(I)), \frac{du}{dt} \in L_2(0,T;H^{-1}(I))\}$
$\mathfrak{S}_h \equiv \{S_h\}_h$	A family of finite-dimensional subspaces of $H_0^1(I)$ with parameter $h < 1$ that tends to 0
$r(kA_h)$	Padé approximation $(I_d + kA_h)^{-1}$
$\mathcal{G}_{\infty}^{(k)}$	$\{f \in L_2(I) \mid f \perp \mathrm{Span}\{r(kA)u((N-1)k, \cdot)\}\}$
$\mathcal{G}_h^{(k)}$	$\{f \in L_2(I) \mid f \perp \mathrm{Span}\{r(kA_h)\overline{P}_h u((N-1)k, \cdot)\}\}$
$S_R^{(k)}$	$\{f \in L_2(I) \mid \|f\| = R, f \in \mathcal{G}_{\infty}^{(k)}\}$ with R satisfying $R > \|\overline{P}_h \widetilde{\sigma}_0\|$

Now, we prove Theorem 1. First, let us introduce a space

$$\mathfrak{W}(T) \equiv \left\{ u \middle| u \in L_2\big(0,T;H^1(I)\big), \frac{\mathrm{d}u}{\mathrm{d}t} \in L_2\big(0,T;H^{-1}(I)\big) \right\}.$$

We note the fact that (see, [12], Chapter XVIII, Theorem 1):

$$\mathfrak{W}(T) \subset C\big(0,T;L_2(I)\big), \tag{B.1}$$

holds. We shall seek a T_{u_0} and $v \in \mathfrak{W}(T_{u_0})$ that solves (3.3) in the following sense:

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}(v(\cdot),w) + \sigma(v(\cdot,w)) \\ \quad = (\phi(\int_I w_1(t,x,y)v(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y), w) \\ \qquad \text{on } (0,T_{u_0}), \\ v(0) = \tilde{u}_0 \quad \text{on } I, \end{cases} \tag{B.2}$$

in the sense of $(C_0^\infty(0,T))'$ for all $w \in H^1(I)$. Note that due to (B.1), the initial condition in the second equation of (B.2) has a meaning. We shall prove this in the following steps [12, 64]. First, assuming the temporally local solvability of the problem, we prove the uniqueness of the local solution. Second, we prove the existence of a local solution up to a certain time T_{u_0}. Let us assume that we have temporally local two solutions to (B.2) on a time interval $[0,T^*]$, say, $v^{(1)}$ and $v^{(2)}$, which belong to the space mentioned in Theorem 1 and subsequent Remark 1.

We introduce a notation $\tilde{v} \equiv v_{(1)} - v_{(2)}$. This should satisfy:

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}(\tilde{v}(\cdot),w) + \sigma(\tilde{v}(\cdot),w) = (\Phi(\cdot),w), \\ \tilde{v}(0) = 0, \end{cases} \tag{B.3}$$

where

$$\begin{aligned} \Phi(t,x) \equiv{}& \phi\left(\int_I w_1(t,x,y)v_{(1)}(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y \right) \\ & - \phi\left(\int_I w_1(t,x,y)v_{(2)}(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y \right). \end{aligned}$$

Replacing w with $\tilde{v}(t,x)$ on both sides of (B.3), and applying the Schwartz's inequality, we observe:

$$\frac{\mathrm{d}}{\mathrm{d}t}\big|\tilde{v}(t)\big|^2 + v\big|\tilde{v}_x(t)\big|^2 \le L\int_I \big\|w_1(t,\cdot,\cdot)\big\|_{L_2(I\times I)}\big|\tilde{v}(t)\big|^2\,\mathrm{d}t,$$

where $L>0$ is the Lipschitz constant of $\phi(\cdot)$. This, together with the Gronwall's inequality [57] and the fact that $\tilde{v}|_{t=0} = 0$, yields

$$\tilde{v}(t) \equiv 0 \quad \forall t \in \big(0,T^*\big),$$

which implies the uniqueness of the solution.

Next, we prove the existence of a local solution. Let $\{V_m\}_{m=1}^{\infty}$ be an increasing family of d_m dimensional subspaces of $H^1(I)$, in which each $v \in H^1(I)$ has its approximating sequence $\{v^{(m)}\}_{m=1}^{\infty}$ such that $v^{(m)} \in V_m$ for each m, and $\|v^{(m)} - v\|_{H^1(I)} \to 0$ as $m \to \infty$. Because V_m is a Galerkin approximation of $L_2(I)$ as well, we have a sequence $\{\tilde{u}_{0m}\}_{m=1}^{\infty}$ such that

$$\tilde{u}_{0m} \in V_m,$$
$$\tilde{u}_{0m} \to \tilde{u}_0 \quad \text{in } L_2(I).$$

Let $\{W_{jm}\}_{j=1}^{d_m}$ be a basis in V_m. We seek $v^{(m)}$ and $\tilde{u}_{0m}$ of the form of linear combinations of $\{W_{jm}\}_{j=1}^{d_m}$ that solve

$$\begin{cases} (\frac{\mathrm{d}v^{(m)}}{\mathrm{d}t}, W_{jm}) + \sigma(v^{(m)}, W_{jm}) \\ \quad = (\phi(\int_I w_1(t,x,y)v^{(m)}(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y), W_{jm}), \\ v^{(m)}|_{t=0} = \tilde{u}_{0m} \quad (j = 1, 2, \dots, d_m). \end{cases} \tag{B.4}$$

Because W_{jm} are linearly independent with each other, (B.4) is assured to have a local solution $v^{(m)} \in C(0, T_{u_0}; V_m)$ with some T_{u_0}. It also satisfies $\frac{\mathrm{d}v^{(m)}}{\mathrm{d}t} \in L_2(0, T_{u_0}; V_m)$ under the assumptions of the theorem.

Next, we observe the *a-priori* estimate. Let us multiply the coefficient of $v^{(m)}$ on both sides of (B.4) for each j, and sum up with respect to $j = 1, 2, \dots, d_m$. Then, we have

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\left|v^{(m)}(t)\right|^2 + \nu\left|v_x^{(m)}(t)\right|^2 \\ &\quad \le \left|v^{(m)}(t)\right|\left|\phi\left(\int_I w_1(t,\cdot,y)v^{(m)}(t,y)\,\mathrm{d}y + \int_I w_1(t,\cdot,y)\,\mathrm{d}y\right)\right|. \end{aligned}$$

Regarding the right-hand side, with a notation $c_1 = |\phi(0)|^2$, we estimate from above as follows.

$$\begin{aligned} &\int_I \left|\phi\left(\int_I w_1(t,x,y)v^{(m)}(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y\right)\right|^2 \mathrm{d}x \\ &\quad \le 2\int_I \left|\phi\left(\int_I w_1(t,x,y)v^{(m)}(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y\right) - \phi(0)\right|^2 \mathrm{d}x \\ &\qquad + 2\left|\phi(0)\right|^2 \\ &\quad \le 4L^2\int_I \left|\int_I w_1(t,x,y)v^{(m)}(t,y)\,\mathrm{d}y\right|^2 \mathrm{d}x \\ &\qquad + 4L^2\int_I \left|\int_I w_1(t,x,y)\,\mathrm{d}y\right|^2 \mathrm{d}x + 2c_1 \\ &\quad \le 4L^2\left\|w_1(t,\cdot,\cdot)\right\|_{L_2(I\times I)}^2\left|v^{(m)}(t)\right|^2 + 4L^2\left\|w_1(t,\cdot,\cdot)\right\|_{L_2(I\times I)}^2 + 2c_1. \end{aligned}$$

This yields

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\left|v^{(m)}(t)\right|^2+\nu\left|v_x^{(m)}(t)\right|^2\\&\quad\le 2L\left|v^{(m)}(t)\right|^2\left\|w_1(t,\cdot,\cdot)\right\|_{L_2(I\times I)}+\left\{2L\left\|w_1(t,\cdot,\cdot)\right\|_{L_2(I\times I)}+\sqrt{2c_1}\right\},\end{aligned}$$

from which, together with the Gronwall's inequality again, we obtain

$$\begin{aligned}\left|v^{(m)}(t)\right|^2\le{}&\left\{|\tilde{u}_{0m}|^2+\frac{1}{2}\int_0^t\left(2L\left\|w_1(\tau,\cdot,\cdot)\right\|_{L_2(I\times I)}+\sqrt{2c_1}\right)\mathrm{d}\tau\right\}\\&\times\exp\left(2L\int_0^t\left\|w_1(\tau,\cdot,\cdot)\right\|_{L_2(I\times I)}\mathrm{d}\tau\right).\end{aligned}$$

This enables us to extract a subsequence $\{v^{(m')}\}\subset\{v^{(m)}\}$ satisfying the following issues with some $v_\infty\in L_2(I)$:

$$\begin{aligned}&v^{(m')}\to v_\infty\quad\text{weakly in } L_2\big(0,T_{u_0};H^1(I)\big),\\&v^{(m')}\to v_\infty\quad\text{weakly* in } L_\infty\big(0,T_{u_0};L_2(I)\big),\\&Av^{(m')}\to Av_\infty\quad\text{weakly in } L_2\big(0,T_{u_0};H^{-1}(I)\big).\end{aligned}\tag{B.5}$$

In virtue of the Relich's theorem, we have

$$v^{(m)}\to v_\infty\quad\text{strongly in } L_2\big(0,T_{u_0};L_2(I)\big).$$

Now, we are in a position to check that this v_∞ certainly solves (B.2). In order for this, we take an arbitrary smooth function $\zeta(t)\in C_0^\infty(0,T)$ and $\breve{w}\in H^1(I)$, a sequence $\{w_m\}_m\subset H^1(I)$ satisfying

$$\lim_{m\to\infty}w_m=\breve{w}\quad\text{in } H^1(I),$$

and define $\psi_m\equiv\zeta(t)w_m$ and $\psi\equiv\zeta(t)\breve{w}$ (note that because we consider in one-dimensional space where $H^1(I)$ can be embedded into $C(I)$, we can regard $\mathfrak{V}=H^1(I)$ in Definition 1 [12]). It is clear that as $m\to+\infty$,

$$\begin{aligned}&\psi_m\to\psi\quad\text{strongly in } L_2\big(0,T_{u_0};H^1(I)\big),\\&\frac{\mathrm{d}\psi_m}{\mathrm{d}t}\to\frac{\mathrm{d}\psi}{\mathrm{d}t}\quad\text{strongly in } L_2\big(0,T_{u_0};L_2(I)\big).\end{aligned}\tag{B.6}$$

For now, we can replace m above with m' prescribed. Thus, from (B.4), after integration by parts (note that $\zeta(t)\in C_0^\infty(0,T)$), we have

$$\begin{aligned}&-\int_0^{T_{u_0}}\left(v^{(m')}(t),\frac{\mathrm{d}\psi_{m'}(t)}{\mathrm{d}t}\right)\mathrm{d}t+\int_0^{T_{u_0}}\sigma\left(v^{(m')}(t),\psi_{m'}(t)\right)\mathrm{d}t\\&\quad=\int_0^{T_{u_0}}\left(\phi\left(\int_I w_1(t,x,y)v^{(m')}(t,y)\,\mathrm{d}y+\int_I w_1(t,x,y)\,\mathrm{d}y\right),\psi_{m'}(t)\right)\mathrm{d}t.\end{aligned}\tag{B.7}$$

In virtue of (B.5) and (B.6), as $m \to +\infty$, we have

$$
\begin{aligned}
&-\int_0^{T_{u_0}} \big(\nu_\infty(t), \breve{w}\big)\zeta'(t)\,\mathrm{d}t + \int_0^{T_{u_0}} \sigma\big(\nu_\infty(t), \breve{w}\big)\zeta(t)\,\mathrm{d}t \\
&\quad = \int_0^{T_{u_0}} \left(\phi\left(\int_I w_1(t,x,y)\nu_\infty(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y\right), \breve{w}\right)\zeta(t)\,\mathrm{d}t. \qquad \text{(B.8)}
\end{aligned}
$$

The equality above holds for any $\breve{w} \in H^1(I)$, and thus, this ν_∞ solves (B.2).

Now, (B.8) can be rewritten as follows.

$$
\begin{aligned}
&-\int_0^{T_{u_0}} \big(\nu_\infty(t), \breve{w}\big)\zeta'(t)\,\mathrm{d}t \\
&\quad = \int_0^{T_{u_0}} \left(\phi\left(\int_I w_1(t,x,y)\nu_\infty(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y\right) - A\nu_\infty, \breve{w}\right)\zeta(t)\,\mathrm{d}t.
\end{aligned}
$$

We can easily see

$$
\frac{\mathrm{d}\nu_\infty}{\mathrm{d}t} \in L_2\big(0, T; H^{-1}(I)\big),
$$

which, together with (B.1), yields the fact that ν_∞ belongs to the same space mentioned in Theorem 1 and subsequent Remark 1.

Finally, we verify that ν_∞ above satisfies the initial condition. Let $\eta(t) \in C^\infty(0, T_{u_0})$ be a function that satisfies $\eta(t) = 0$ near T_{u_0} and $\eta(0) \neq 0$. We again take $\breve{w} \in H^1(I)$ and a sequence $\{w_m\}_m \subset H^1(I)$ satisfying

$$
\lim_{m\to\infty} w_m = \breve{w} \quad \text{in } H^1(I).
$$

Then, $\psi = \eta(t)\breve{w} \in \mathfrak{W}(T_{u_0})$ and by integration by parts, we have

$$
\int_0^{T_{u_0}} \left(\frac{\mathrm{d}\nu_\infty}{\mathrm{d}t}(t), \eta(t)\breve{w}\right)\mathrm{d}t = -\int_0^{T_{u_0}} \big(\nu_\infty(t), \breve{w}\big)\eta'(t)\,\mathrm{d}t - \big(\nu_\infty(0), \breve{w}\big)\eta(0). \qquad \text{(B.9)}
$$

From Equation (B.2), we can derive

$$
\begin{aligned}
&\int_0^{T_{u_0}} \left(\frac{\mathrm{d}\nu_\infty(t)}{\mathrm{d}t}, \eta(t)\breve{w}\right)\mathrm{d}t \\
&\quad = \int_0^{T_{u_0}} \left(\phi\left(\int_I w_1(t,x,y)\nu_\infty(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y\right), \breve{w}\right)\eta(t)\,\mathrm{d}t \\
&\qquad - \int_0^{T_{u_0}} \sigma\big(\nu_\infty(t), \breve{w}\big)\eta(t)\,\mathrm{d}t. \qquad \text{(B.10)}
\end{aligned}
$$

Moreover, from (B.4) we have

$$\begin{aligned}
&\int_0^{T_{u_0}} \left(\frac{\mathrm{d}v^{(m')}(t)}{\mathrm{d}t}, w_{m'} \right) \eta(t)\,\mathrm{d}t \\
&\quad = \int_0^{T_{u_0}} \left(\phi\left(\int_I w_1(t,x,y) v^{(m')}(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y \right), w_{m'} \right) \eta(t)\,\mathrm{d}t \\
&\qquad - \int_0^{T_{u_0}} \sigma\left(v^{(m')}(t), w_{m'} \right) \eta(t)\,\mathrm{d}t.
\end{aligned} \tag{B.11}$$

The left-hand side of (B.11) has another representation:

$$\begin{aligned}
&\int_0^{T_{u_0}} \left(\frac{\mathrm{d}v^{(m')}(t)}{\mathrm{d}t}, w_{m'} \right) \eta(t)\,\mathrm{d}t \\
&\quad = -\int_0^{T_{u_0}} \left(v^{(m')}(t), w_{m'} \right) \eta'(t)\,\mathrm{d}t - (\tilde{u}_{0m'}, w_{m'}) \eta(0).
\end{aligned} \tag{B.12}$$

Making m' tend to $+\infty$, on the one hand, (B.11) yields

$$\begin{aligned}
&\lim_{m' \to +\infty} \int_0^{T_{u_0}} \left(\frac{\mathrm{d}v^{(m')}}{\mathrm{d}t}(t), w_{m'} \right) \eta(t)\,\mathrm{d}t \\
&\quad = \int_0^{T_{u_0}} \left(\phi\left(\int_I w_1(t,x,y) v_\infty(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y \right), \breve{w} \right) \eta(t)\,\mathrm{d}t \\
&\qquad - \int_0^{T_{u_0}} \sigma\left(v_\infty(t), \breve{w} \right) \eta(t)\,\mathrm{d}t \\
&\quad = \int_0^{T_{u_0}} \left(\frac{\mathrm{d}v_\infty}{\mathrm{d}t}(t), \eta(t)\breve{w} \right) \mathrm{d}t,
\end{aligned} \tag{B.13}$$

where we used (B.12) to deduce the right-most hand side. On the other hand, (B.12) yields

$$\begin{aligned}
&\lim_{m' \to +\infty} \int_0^{T_{u_0}} \left(\frac{\mathrm{d}v^{(m')}}{\mathrm{d}t}(t), w_{m'} \right) \eta(t)\,\mathrm{d}t \\
&\quad = -\int_0^{T_{u_0}} \left(v_\infty(t), \breve{w} \right) \eta'(t)\,\mathrm{d}t - (\tilde{u}_0, \breve{w}) \eta(0).
\end{aligned} \tag{B.14}$$

By comparing (B.12), (B.13), and (B.14), we arrive at

$$\left(v_\infty(0), \breve{w} \right) = (\tilde{u}_0, \breve{w}) \quad \forall \breve{w} \in H^1(I). \tag{B.15}$$

Because $H^1(I)$ is dense in $L_2(I)$, (B.15) holds for all $\breve{w} \in L_2(I)$, which implies

$$v_\infty(0) = \tilde{u}_0.$$

This is the desired result.

B.2 Proof of Theorem 2

Here, we prove Theorem 2. Because the local solvability is assured in Theorem 1, we assume that for some $T^* > 0$, we have a solution v of (3.3) on the interval $[0, T^*]$. Now, let us

first construct a variable:

$$\breve{\nu}(t,x) \equiv e^{-\lambda t}\nu(t,x), \quad t \in [0,T^*],$$

with $\lambda \in \mathbb{R}$ specified later, which solves

$$\begin{cases} \breve{\nu}_t - \nu\breve{\nu}_{xx} = -\lambda\breve{\nu} + e^{-\lambda t}\phi\left(\int_I w_1(t,x,y)e^{\lambda t}\breve{\nu}(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y\right) & \text{in } I_{T^*}, \\ \breve{\nu}(0,x) = \tilde{u}_0 \quad \text{on } I, \\ \breve{\nu} = 0 \quad \text{on } \partial I\ \forall t \in (0,T^*). \end{cases} \tag{B.16}$$

By multiplying $\breve{\nu}$ on both sides in (B.16), we can deduce an estimation as below:

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left|\breve{\nu}(t)\right|^2 + \nu\left|\breve{\nu}_x(t)\right|^2 + \lambda\left|\breve{\nu}(t)\right|^2 \\ &\quad \le e^{-\lambda t}\left|\breve{\nu}(t)\right|\left\|\phi\left(\int_I w_1(t,\cdot,y)e^{\lambda t}\breve{\nu}(t,y)\,\mathrm{d}y + \int_I w_1(t,\cdot,y)\,\mathrm{d}y\right)\right\|_{L_2(I)}. \end{aligned} \tag{B.17}$$

By introducing a notation $c_1 = |\phi(0)|^2$ again, by applying the Schwartz inequality and the Lipschiz continuity of ϕ, we have

$$\begin{aligned} &\int_I \left|\phi\left(\int_I w_1(t,x,y)e^{\lambda t}\breve{\nu}(t,y)\,\mathrm{d}y + \int_I w_1(t,x,y)\,\mathrm{d}y\right)\right|^2 \mathrm{d}x \\ &\quad \le 4L^2e^{2\lambda t}\left\|w_1(t,\cdot,\cdot)\right\|^2_{L_2(I\times I)}\left|\breve{\nu}(t)\right|^2 + 4L^2\left\|w_1(t,\cdot,\cdot)\right\|^2_{L_2(I\times I)} + 2c_1, \end{aligned}$$

if we substitute this to (B.17), we obtain

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left|\breve{\nu}(t)\right|^2 + \nu\left|\breve{\nu}_x(t)\right|^2 + \lambda\left|\breve{\nu}(t)\right|^2 \\ &\quad \le 2L\left\|w_1(t,\cdot,\cdot)\right\|_{L_2(I\times I)}\left|\breve{\nu}(t)\right|^2 \\ &\qquad + \frac{e^{-2\lambda t}}{2\lambda}\left(2L\left\|w_1(t,\cdot,\cdot)\right\|_{L_2(I\times I)} + c_1\right)^2 + \frac{\lambda}{2}\left|\breve{\nu}(t)\right|^2. \end{aligned} \tag{B.18}$$

If we denote

$$G(t) \equiv \frac{e^{-2\lambda t}}{2\lambda}\left(2L\left\|w_1(t,\cdot,\cdot)\right\|_{L_2(I\times I)} + c_1\right)^2,$$

by the Gronwall's inequality, we have

$$\left|\breve{\nu}(t)\right|^2 \le \left(|\tilde{u}_0|^2 + 2\int_0^t G(\tau)\,\mathrm{d}\tau\right)\exp\left(4L\int_0^t \left\|w_1(\tau,\cdot,\cdot)\right\|_{L_2(I\times I)}\,\mathrm{d}\tau - \lambda t\right). \tag{B.19}$$

Applying the Schwartz's inequality to the right-hand side of (B.19) and taking λ so that

$$\lambda \ge \frac{4L\|w_1\|_{L_2(\mathcal{H}_\infty)}}{\sqrt{T^*}}$$

holds, then (B.19) yields

$$\left|\check{\nu}\left(T^*\right)\right|^2 \leq |\tilde{u}_0|^2 + 2\int_0^\infty G(\tau)\,\mathrm{d}\tau.$$

This implies that the norm $|\check{\nu}(T^*)|$ does not depend on T^*. Tracing the same argument as in [64], we have the statement of the theorem.

Acknowledgements
We thank the anonymous reviewers whose comments and suggestions greatly help improve and clarify this manuscript. We also appreciate Mamoru Miyazawa, who contributed to the numerical experiments in this study.

Funding
This work was supported by Toyo University Top Priority Research Program.

Availability of data and materials
The datasets generated and/or analyzed during the current study are available in: (i) UCI Machine Learning Repository, [https://doi.org/10.24432/C5XW20], (ii) Kaggle repository, [https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset].

Declarations

Ethics approval and consent to participate
There are no ethics approval.

Competing interests
The author declares no competing interests.

Author contributions
Only one author for this paper. The author read and approved the final manuscript.

Received: 20 December 2022 Accepted: 6 October 2023 Published online: 25 October 2023

References

1. Aizawa, Y., Kimura, M.: Universal approximation properties for ODENet and ResNet. CoRR (2021). arXiv:2101.10229
2. Annunziato, M., Borzì, A.: A Fokker–Planck control framework for multidimensional, stochastic processes. J. Comput. Appl. Math. **237**, 487–507 (2013). https://doi.org/10.1016/j.cam.2012.06.019
3. Baker, G.A., Bramble, J.H., Thomee, V.: Single step Galerkin approximations for parabolic problems. Math. Comput. **31**, 818–847 (1977). https://doi.org/10.2307/2006116
4. Barbu, V.: Analysis and Control of Nonlinear Infinite Dimensional Systems. Academic Press, London (2012)
5. Barry, B., Ronny, K.: Adult income dataset, UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20
6. Baum, E.B., Haussler, D.: What size net gives valid generalization? Neural Comput. **1**, 151–160 (1989). https://doi.org/10.1162/neco.1989.1.1.151
7. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Singapore (2006)
8. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. (1998). https://doi.org/10.1023/A:1009715923555
9. Chamberlain, B.P., et al.: GRAND: graph neural diffusion. In: Proc. ICML 2021 (2021)
10. Chen, R.T.Q., et al.: Neural ordinary differential equations. Adv. Neural Inf. Process. Syst. **31**, 6572–6583 (2018)
11. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. **2**, 303–314 (1989). https://doi.org/10.1007/BF02551274
12. Dautray, R., Lions, L.J.: Mathematical Analysis and Numerical Methods for Science and Technology, vol. 5. Springer, Berlin (1991)
13. Deap (2023). https://deap.readthedocs.io/en/master/
14. DeVore, R., Hanin, B., Petrova, G.: Neural network approximation. Acta Numer. **30**, 327–444 (2021). https://doi.org/10.1017/S0962492921000052
15. Diabetes dataset: Kaggle (2020). https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset
16. Dudley, R.M.: Uniform Central Limit Theorems. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge (1999). https://doi.org/10.1017/CBO9780511665622
17. Dupont, E., Doucet, A., Teh, Y.W.: Augmented neural ODEs. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Red Hook (2019)
18. Esteve-Yagüe, C., et al.: Large-time asymptotics in deep learning (2021). https://hal.archives-ouvertes.fr/hal-02912516
19. Esteve-Yagüe, C., Geshkovski, B.: Sparse approximation in learning via neural ODEs. (2021). arXiv:2102.13566
20. Fernández-Cara, E., et al.: Null controllability of linear heat and wave equations with nonlocal spatial terms. SIAM J. Control Optim. **54**, 2009–2019 (2016). https://doi.org/10.1137/15M1044291
21. Fujita, H., Mizutani, A.: On the finite element method for parabolic equations, I; approximation of holomorphic semi-groups. J. Math. Soc. Jpn. **28**, 749–771 (1976). https://doi.org/10.2969/jmsj/02840749
22. Funahashi, K.: On the approximate realization of continuous mappings by neural networks. Neural Netw. **2**, 183–192 (1989). https://doi.org/10.1016/0893-6080(89)90003-8

23. Funahashi, K., Nakamura, Y., Networks, N.: Neural Networks, Approximation Theory, and Dynamical Systems (Structure and Bifurcation of Dynamical Systems), Suuri-kaiseki kenykuujo Kokyuroku, 18–37 (1992). http://hdl.handle.net/2433/82914
24. Geshkovski, B., Zuazua, E.: Turnpike in optimal control of PDEs, ResNets, and beyond. Acta Numer. **31**, 135–263 (2022). https://doi.org/10.1017/S0962492922000046
25. Giné, E., Nickl, R.: Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2015). https://doi.org/10.1017/CBO9781107337862
26. González-Burgos, M., de Teresa, L.: Some results on controllability for linear and nonlinear heat equations in unbounded domains. Adv. Differ. Equ. **12**, 1201–1240 (2007). https://doi.org/10.57262/ade/1355867413
27. Haber, E., Ruthotto, L.: Stable architectures for deep neural networks. Inverse Probl. **34**, 014004 (2017). https://doi.org/10.1088/1361-6420/aa9a90
28. Han, E.W., Han, J., Li, Q.: A mean-field optimal control formulation of deep learning. Res. Math. Sci. **6**, 10 (2019). https://doi.org/10.1007/s40687-018-0172-y
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE Comput. Soc., Los Alamitos (2016). https://doi.org/10.1109/CVPR.2016.90
30. Hoff, D., Smoller, J.: Error bounds for finite-difference approximations for a class of nonlinear parabolic systems. Math. Comput. **45**, 35–49 (1985). https://doi.org/10.2307/2008048
31. Honda, H.: On continuous limit of neural network. In: Proc. of NOLTA 2020 (2020)
32. Honda, H.: On a partial differential equation based neural network. IEICE Commun. Express **10**, 137–143 (2021). https://doi.org/10.1587/comex.2020XBL0174
33. Honda, H., et al.: An ODE-based neural network with bayesian optimization. JSIAM Lett. **15**, 101–104 (2023). https://doi.org/10.1587/comex.2020XBL0174
34. Honda, H.: Approximating a multilayer neural network by an optimal control of a partial differential equation. Preprint
35. Hornik, K., et al.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**, 359–366 (1989). https://doi.org/10.1016/0893-6080(89)90020-8
36. Irie, B., Miyake, S.: Capabilities of three-layered perceptrons. In: Proc. IEEE Int. Conf. on Neural Networks, pp. 641–648 (1988). https://doi.org/10.1109/ICNN.1988.23901
37. Ito, S.: Fundamental solutions of parabolic differential equations and boundary value problems. Jpn. J. Math., Trans. Abstr. **27**, 55–102 (1957). https://doi.org/10.4099/jjm1924.27.055
38. Kac, V.G., Cheung, P.: Quantum Calculus. Springer, New York (2001)
39. Kato, T.: Perturbation Theory for Linear Operators, 2nd edn. Springer, New York (1976)
40. Koenderink, J.J.: The structure of images. Biol. Cybern. **50**, 363–370 (1984). https://doi.org/10.1007/BF00336961
41. Kolmogorov, A.N.: On the representation of continuous function of many variables by superposition of continuous function of one variable and addition. Dokl. Akad. Nauk SSSR **144**, 679–681 (1957)
42. Laakmann, F., Petersen, P.C.: Efficient approximation of solutions of parametric linear transport equations by ReLU DNNs. Adv. Comput. Math. **47**, 11 (2021)
43. Leshno, M., et al.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Netw. **6**, 303–314 (1993). https://doi.org/10.1016/S0893-6080(05)80131-5
44. Li, Q., et al.: Maximum principle based algorithms for deep learning. J. Mach. Learn. Res. **18**, 5998–6026 (2017)
45. Li, Q., Lin, T., Shen, Z.: Deep learning via dynamical systems: an approximation perspective. J. Eur. Math. Soc. (2019). https://doi.org/10.4171/jems/1221
46. Li, Z., Shi, Z.: Deep residual learning and PDEs on manifold (2017). arXiv:1708.05115
47. Lions, J.L.: Perturbations Singulières dans les Problèmes aux Limites et en Contrôle Optimal. Springer, Berlin (1973)
48. Lions, J.L.: Exact controllability, stabilization and perturbations for distributed systems. SIAM Rev. **30**, 1–68 (1988). https://doi.org/10.1137/1030001
49. Lions, J.L., Magenes, E.: Non-homogeneous Boundary Values Problems and Applications I. Springer, Berlin (1972)
50. Lions, P.L.: Une vision mathématique du Deep Learning (2018). https://www.college-de-france.fr/fr/agenda/seminaire/mathematiques-appliquees/une-vision-mathematique-du-deep-learning
51. Lippmann, R.: An introduction to computing with neural nets. IEEE ASSP Mag. **4**, 4–22 (1987). https://doi.org/10.1109/MASSP.1987.1165576
52. Liu, H., Markowich, P.: Selection dynamics for deep neural networks. J. Differ. Equ. **269**, 11540–11574 (2020). https://doi.org/10.1016/j.jde.2020.08.041
53. Lohéac, J., Zuazua, E.: From averaged to simultaneous controllability. Ann. Fac. Sci. Toulouse, Math. **25**, 785–828 (2016)
54. Neal, R.M.: Bayesian Learning for Neural Networks. Springer, Berlin (1996)
55. Nirenberg, L.: Topics in Nonlinear Functional Analysis. Am. Math. Soc., Providence (2001)
56. Oono, K., Suzuki, T.: Graph neural networks exponentially lose expressive power for node classification (2020). https://api.semanticscholar.org/CorpusID:209994765
57. Pachpatte, B.G., Ames, W.F.: Inequalities for Differential and Integral Equations. Academic Press, London (1997)
58. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. **12**, 629–639 (1990). https://doi.org/10.1109/34.56205
59. Rodriguez, I.D.J., Ames, A.D., Yue, Y.: Lyanet: a Lyapunov framework for training neural ODEs. CoRR (2022). arXiv:2202.02526
60. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. **65**, 386–408 (1958)
61. Ruiz-Balet, D., Zuazua, E.: Neural ODE control for classification, approximation and transport. SIAM Rev. **65**, 735–773 (2023). https://doi.org/10.1137/21M1411433
62. Rusch, T.K., et al.: Graph-coupled oscillator networks. CoRR (2022). arXiv:2202.02296
63. Ruthotto, L., Haber, E.: Deep neural networks motivated by partial differential equations. J. Math. Imaging Vis. **62**, 352–364 (2020). https://doi.org/10.1007/s10851-019-00903-1
64. Ryu, S.U., Yagi, A.: Optimal control of Keller–Segel equations. J. Math. Anal. Appl. **256**, 45–66 (2001)

65. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning. Cambridge University Press, Padstow Cornwall (2014)
66. Shen, Z., Yang, H., Zhang, S.: Nonlinear approximation via compositions. CoRR (2019). arXiv:1902.10170
67. Sonoda, S., Murata, N.: Double continuum limit of deep neural networks. In: Proc. of ICML 2017, Workshop on Principled Approaches to Deep Learning (2017)
68. Sonoda, S., Murata, N.: Transport analysis of infinitely deep neural network. J. Mach. Learn. Res. **20**, 1–52 (2019)
69. Sontag, E., Sussmann, H.: Complete controllability of continuous-time recurrent neural networks. Syst. Control Lett. **30**, 177–183 (1997). https://doi.org/10.1016/S0167-6911(97)00002-9
70. Sprecher, D.A.: On the structure of continuous functions of several variables. Trans. Am. Math. Soc. **115**, 340–355 (1965). https://doi.org/10.2307/1994273
71. Statsmodels (2023). https://www.statsmodels.org/
72. Stelzer, F., et al.: Deep neural networks using a single neuron: folded-in-time architecture using feedback-modulated delay loops. Nat. Commun. **12**, 1–10 (2021). https://doi.org/10.1038/s41467-021-25427-4
73. Tabuada, P., et al.: Universal approximation power of deep residual neural networks through the lens of control. IEEE Trans. Autom. Control **68**, 2715–2728 (2023). https://doi.org/10.1109/TAC.2022.3190051
74. Temam, R.: Infinite-Dimensional Dynamical Systems in Mechanics and Physics. Springer, New York (1997)
75. Teshima, T., et al.: Coupling-based invertible neural networks are universal diffeomorphism approximators. CoRR (2020). arXiv:2006.11469
76. Teshima, T., et al.: Universal approximation property of neural ordinary differential equations (2020). arXiv:2012.02414
77. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Springer, Berlin (2006)
78. Thorpe, M., van Gennip, Y.: Deep limits of residual neural networks. Res. Math. Sci. **10**, 6 (2023). https://doi.org/10.1007/s40687-022-00370-y
79. Trotter, H.F.: Approximation of semi-groups of operators. Pac. J. Math. **8**, 887–919 (1958)
80. Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer Series in Statistics. Springer, New York (1996). https://doi.org/10.1007/978-1-4757-2545-2
81. Vainikko, G.: Funktionalanalysis der Diskretisierungsmethoden. Teubner, Leipzig (1976)
82. Weickert, J.: Anisotropic Diffusion in Image Processing (1998). https://www.mia.uni-saarland.de/weickert/Papers/book.pdf
83. Weinan, E.: A proposal on machine learning via dynamical systems. Commun. Math. Stat. **5**, 1–11 (2017). https://doi.org/10.1007/s40304-017-0103-z
84. Williams, C.: Computing with infinite networks. In: Mozer, M., Jordan, M., Petsche, T. (eds.) Advances in Neural Information Processing Systems, vol. 9. MIT Press, Cambridge (1996)
85. Yun, B.I.: A neural network approximation based on a parametric sigmoidal function. Mathematics **7**, 262 (2019). https://www.mdpi.com/2227-7390/7/3/262
86. Yunjin, C., Thomas, P.: Trainable nonlinear reaction diffusion: a flexible framework for fast and effective image restoration. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1256–1272 (2017). https://doi.org/10.1109/TPAMI.2016.2596743
87. Zeidler, E.: Nonlinear Functional Analysis and Its Applications. Springer, New York (1986)
88. Zhang, H., et al.: Approximation capabilities of neural ODEs and invertible residual networks. In: Daumé, H., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 11086–11095 (2020)

Publisher's Note