sciendo

CAIM

ISSN 2038-0909

Commun. Appl. Ind. Math. 13 (1), 2022, 96–120

DOI: 10.2478/caim-2022-0008

Research Article

# Continuous limits of residual neural networks in case of large input data

**Michael Herty[1*], Anna Thünen[2],Torsten Trimborn[3],Giuseppe Visconti[4]**

[1]Institut für Geometrie und Praktische Mathematik (IGPM) – RWTH Aachen University – Templergraben 55, 52062 Aachen (Germany)

[2]Institut für Mathematik – TU Clausthal – Erzstraße 1, 38678 Clausthal-Zellerfeld (Germany)

[3]NRW.BANK – Kavalleriestraße 22, 40213 Düsseldorf (Germany)

[4]Department of Mathematics "G. Castelnuovo" – Sapienza University of Rome – P.le Aldo Moro 5, 00185 Roma (Italy)

[*]Email address for correspondence: herty@igpm.rwth-aachen.de

Communicated by Lorenzo Pareschi

Received on 07 11, 2022. Accepted on 11 12, 2022.

## Abstract

Residual deep neural networks (ResNets) are mathematically described as interacting particle systems. In the case of infinitely many layers the ResNet leads to a system of coupled system of ordinary differential equations known as neural differential equations. For large scale input data we derive a mean–field limit and show well–posedness of the resulting description. Further, we analyze the existence of solutions to the training process by using both a controllability and an optimal control point of view. Numerical investigations based on the solution of a formal optimality system illustrate the theoretical findings.

*Keywords:* `Neural networks, mean--field limit, well--posedness, optimal control, controllability`

*AMS subject classification:* `35Q83, 49J15, 49J20, 92B20`

## 1. Introduction

In the last years, there has been a growing interest in machine learning and data science applications [1–3], e.g. in the fields of recognition of human speech [4], competition at the highest level in strategic game systems [5], intelligent routing in content delivery networks [6], and autonomously operating cars [7]. The intersection between mathematics and artificial intelligence has been mainly based on using machine learning tools to resolve bottlenecks in existing numerical methods, e.g. to replace parameter optimization, parameter–identification and data assimilation methods, or shock–detection techniques for non–oscillatory reconstructions, or to model physics–based operators through experimental data and uncertainty quantification. We refer to [8–15] for additional references on these topics. Here, we contribute to a framework for a particular class of learning–based methods, the deep residual neural networks (ResNets), using a description based on partial differential equations, more precisely, linear kinetic equations. This

formulation allows to apply different techniques to analyze theoretical properties of the underlying neural network.

First, we briefly recall the residual neural networks (ResNet), see also equation (1). Given a set of input data or measurements $x_i^0, i = 1, \dots, M$, the ResNet propagates those through discrete entities, the layers $\kappa = 0, \dots, L+1$, to provide a state prediction $x_i(L+1)$. The dynamics depends on (a large set of) parameters, called weights $w(\kappa)$ and bias $b(\kappa)$. Their values are chosen in an optimization procedure called training to predict given reference data $y_i$. This training procedure is performed by minimizing a given distance $\ell$ between predictions of the network and the given reference values.

Neural networks, and in general machine learning models, are typically processed on very large data sets, formally $M \to \infty$, making both the forward and the training processes computationally costly. An attempt to provide a more synthetic, and statistical, description of the neural network dynamics has been investigated in [16], where a kinetic formulation of neural differential equations is proposed. However, neither the training nor the well–posedness have been analyzed so far. Using a mean–field or kinetic description of large–scale neural networks has so far only be discussed for particular examples in a few recent manuscripts [17–22]. A general investigation in particular in view of large input data is to the best of our knowledge still open.

In this work we contribute towards the mean–field description of neural networks by discussing well–posedness of the mean–field residual neural network using results obtained in the context of pedestrian dynamics [23]. The arising equation is similar to the mean–field model formally proposed in [16]. However, therein the training process has not been discussed. Here, we follow two directions. Training is formulated as a controllability problem for the mean–field equation. This problem allows for solutions for very particular initial and reference data. In the general case, the training process is considered as an optimal control problem with constraints given by the mean–field equation. Here, the derived continuous dependence on the parameters is used to show existence of optimal weights and bias. A numerical method for computing those based on the mean–field equation is implemented and computational results are presented.

The structure of the paper is shortly summarized here. In Section 2 we introduce deep residual neural networks and discuss formally the equations resulting in time and in the mean–field continuous limits. Section 3 is the main part of this work since we provide a rigorous analysis of the mean–field equation. In Section 4 we discuss a computational technique for the training of the mean–field neural network and numerical experiments are performed. Finally, we conclude the paper in Section 5 proposing also research perspectives.

## 2. Continuous limits of residual neural networks

Let us consider a set of $M \in \mathbb{N}$, $M \gg 1$, input data characterized by $d$ measurements. In terms of neural networks, each measurement represents a feature of the given input. Without loss of generality it is possible to assume that the value of each feature is one–dimensional so that each input data can be described by $x_i^0 \in \mathbb{R}^d$, $i = 1, \dots, M$.

As starting point we consider deep Residual Neural Networks (ResNets). Their struc-

ture is given by $L$ hidden layers, with labels $\{1,\dots,L\}$, and in each layer the number of neurons is given by $N_\kappa$, $\forall \kappa = 1,\dots,L$. We use the indices $\kappa = 0$ and $\kappa = L+1$ to denote the input layer and the output layer, respectively. The state of the $i$–th input signal at the $\kappa$–th layer is $x_i(\kappa) \in \mathbb{R}^{N_\kappa}$ and $N_0 = d$. The final state $x_i(L+1) \in \mathbb{R}^{N_{L+1}}$ is called *output* or prediction of the network.

Each input signal $x_i^0$ propagates according to the deterministic dynamics [24]:

$$\begin{cases} x_i(\kappa+1) = A(\kappa)x_i(\kappa) + \Delta t\, \sigma\big(w(\kappa)x_i(\kappa) + b(\kappa)\big), \quad \kappa = 0,\dots,L \\ x_i(0) = x_i^0. \end{cases} \tag{1}$$

Here, $w(\kappa) \in \mathbb{R}^{N_{\kappa+1}\times N_\kappa}$ are the weights and $b(\kappa) \in \mathbb{R}^{N_{\kappa+1}}$ the bias and $\Delta t > 0$ indicates a (pseudo) time step. The vectors $(w, b)$ define the parameters of the network. The matrix $A(\kappa) \in R^{N_{\kappa+1}\times N_\kappa}$ is a deterministic matrix which can be reduced to an identity matrix under Assumption 2.1, cf. the next section, and therefore we do not provide a rigorous definition. The function $\sigma : \mathbb{R} \to \mathbb{R}$ is called activation function of the neurons and it is applied component wise in (1). Examples of activation functions include the identity function $\sigma(x) = x$, the rectified linear unit (ReLU) function $\sigma(x) = \max\{0, x\}$, the sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$, the hyperbolic tangent function $\sigma(x) = \tanh(x)$ and the growing cosine unit (GCU) function $\sigma(x) = x\cos(x)$.

The parameters $w$ and $b$ are chosen in a *training process* in order to have ResNet solve a given learning problem. In *supervised training* we have that *desired* outputs, the targets or reference values, are provided along with the input data. The network processes the input data and then compares the predictions against the targets $\{y_i\}_{i=1}^M$. The error is then propagated back through the network with the aim of optimizing the parameters. This process occurs many times on a set of data which is typically named as training data set and several approaches are known, as e.g. stochastic gradient descent [25] or ensemble Kalman filter [26–28], and they are related to the choice of the *loss functions*. Typical loss functions [29], for instance such as Mean Squared Error, the Mean Absolute Error and the Categorical Cross–Entropy, can be all written as

$$\frac{1}{M}\sum_{i=1}^{M} \ell(x_i(L+1) - y_i), \tag{2}$$

for suitable choices of a (differentiable) function $\ell : \mathbb{R}^{N_{L+1}} \to \mathbb{R}_0^+$.

### 2.1. *Neural differential equations and mean–field limit*

In (1) the layers define a discrete structure within the ResNet. In order to compute the time continuous limit of (1) we interpret the layers as discrete times where the propagation of the input signal is evaluated. To this end we need to introduce the following assumption.

**Assumption 2.1.** *The number of neurons in each layer is fixed and determined by the dimension of the input data. Namely, $N_\kappa = N = d$, $\forall\, \kappa = 1,\dots,L+1$. In addition, $A(\kappa) = I_N$, where $I_N \in R^{N\times N}$ is the identity matrix.*

This assumption typically underlies the derivation of neural differential equations, see also [30]. Note that even the choice $d = 1$ is possible and, in addition, that networks of this type have been already proved to satisfy different formulations of the universal approximation theorem [31–33]. Further, they have been also applied to several (real–world) training problems [34,35].

Under Assumption 2.1 and interpreting $\Delta t$ as the size of a time step, (1) is an explicit Euler discretization of an underlying differential equation. Namely, in the limit $\Delta t \to 0^+$ and $L \to \infty$ such that $\Delta t(L+2) \to T$, (1) formally leads to

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} x_i(t) = \sigma\big(w(t)x_i(t) + b(t)\big), \quad t \in [0,T] \\ x_i(0) = x_i^0, \end{cases} \tag{3}$$

for each $i = 1, \dots, M$. The system of differential equations (3) describes the time propagation of each measurement $x_i(t) \in \mathbb{R}^d$, starting from the initial condition $x_i^0 \in \mathbb{R}^d$ fixed by the input data. It is known as *neural differential equation* [30].

The parameters of the network are given by the time dependent weights $w(t) \in \mathbb{R}^{d\times d}$ and by the time dependent bias $b(t) \in \mathbb{R}^d$, $\forall\, t \geq 0$. By the Picard–Lindelöf Theorem, existence and uniqueness of a solution to (3) is guaranteed as long as the activation function $\sigma$ satisfies the Lipschitz condition and $t \mapsto w(t)$, $t \mapsto b(t)$ are continuous. Notice that the loss functional (2) reads

$$\frac{1}{M} \sum_{i=1}^{M} \ell(x_i(T) - y_i), \tag{4}$$

where $x_i(T)$ represents the state at time $T > 0$ obtained with (3).

It is clear from (3) that the computational and memory cost of the neural network still increases with the dimension of the data set, i.e. $M$. A way to overcome this problem is introducing a statistical interpretation of the neural network by computing the mean–field limit of the neural differential equations (3) for $M \to \infty$. In the limit of infinitely many data we formally obtain the linear equation

$$\begin{cases} \partial_t f(t,x) + \nabla_x \cdot \Big(\sigma\big(w(t)x + b(t)\big) f(t,x)\Big) = 0, \quad t > 0 \\ f(0,x) = f_0(x), \end{cases} \tag{5}$$

which describes the evolution of the distribution $f : \mathbb{R}_{\geq 0} \times \mathbb{R}^d \to \mathbb{R}^d_{\geq 0}$ of the data. The initial condition $f_0(x)$ is obtained as limit of the input data. Since (5) preserves the mass, $f(t,x)$ is a probability distribution $\forall\, t > 0$ provided that $f_0$ is. We point out that in the mean–field limit any information on the network output of a precise measurement is lost. In fact, (5) provides only a statistical information on the neural network propagation and, thus, of the learning problem. The well–posedness of equation (5) and the convergence of (3) to (5) as $M \to \infty$ is proven in 1–Wasserstein distance, see Section 3.1.

| Microscopic | | Mean–field | |
|---|---|---|---|
| Trajectories of (3) | $x_i(t) \in \mathbb{R}^d$ | Weak solution of (5) | $f_t \in \mathcal{P}_1(\mathbb{R}^d)$ |
| Augmented trajectories of (6) | $(x_i(t), \tau_i(t)) \in \mathbb{R}^{d+1}$ | Weak solution of (9) | $F_t \in \mathcal{P}_1(\mathbb{R}^{d+1})$ |
| Target data | $y_i \in \mathbb{R}^d$ | Target measure | $g \in \mathcal{P}_1(\mathbb{R}^d)$ |
| Loss function (4) | $\ell\colon \mathbb{R}^d \to \mathbb{R}$ | Loss function (15) | $\tilde{\ell}\colon \mathbb{R}^d \to \mathbb{R}$ |

## 3. Analysis of the mean–field limit

In this section, we discuss the mean–field limit of the system (3) and related minimization problems following results of [23, Section 6 and 7]. We start with recalling some preliminary notation and refer to [36,37] for more details.

Let $\mathcal{P}(\mathbb{R}^d)$ the set of real–valued probability measures defined on $\mathbb{R}^d$ and, for $p \geq 1$, we denote by $\mathcal{P}_p(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$ the set of probability measures with finite $p$–th moment, i.e.

$$\mathcal{P}_p(\mathbb{R}^d) = \left\{ \mu \in \mathcal{P}(\mathbb{R}^d)\colon \int_{\mathbb{R}^d} |x|^p \mathrm{d}\mu(x) < +\infty \right\}.$$

Throughout the paper we denote by $\mu_t$ a time dependent probability measure for $t \in \mathbb{R}_0^+$.

Given a map $\gamma\colon \mathbb{R}^d \to \mathbb{R}^d$, the push–forward of $\mu \in \mathcal{P}(\mathbb{R}^d)$ through $\gamma$ is defined for every Borel set $A \subset \mathbb{R}^d$ as the unique probability measure $\gamma\#\mu$ such that $\gamma\#\mu(A) \colon = \mu(\gamma^{-1}(A))$. Given two probability measures $\mu$, $\nu \in \mathcal{P}(\mathbb{R}^d)$, a probability measure $\pi$ on the product space $\mathbb{R}^d \times \mathbb{R}^d$ is said to be an admissible transport plan from $\mu$ to $\nu$ if the following properties hold:

$$\int_{y \in \mathbb{R}^d} \mathrm{d}\pi(x,y) = \mathrm{d}\mu(x), \quad \int_{x \in \mathbb{R}^d} \mathrm{d}\pi(x,y) = \mathrm{d}\nu(y).$$

We denote the set of admissible transport plans from $\mu$ to $\nu$ by $\Pi(\mu,\nu)$. Note that the set $\Pi(\mu,\nu)$ is always nonempty, since the product $\mu\nu \in \Pi(\mu,\nu)$. The cost of each admissible transport plan $\pi$ from $\mu$ to $\nu$ can be defined as follows:

$$J[\pi] \colon = \int_{\mathbb{R}^{2d}} |x-y|^p \mathrm{d}\pi(x,y),$$

where $|\cdot|$ represents the Euclidean norm on $\mathbb{R}^d$. A minimizer of $J$ in $\Pi(\mu,\nu)$ always exists. Thus for any two measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, one can define the following metric

$$W_p(\mu,\nu) \colon = \left( \min_{\pi \in \Pi(\mu,\nu)} J[\pi] \right)^{\frac{1}{p}},$$

which is called the $p$–Wasserstein distance. The set of transport plans $\pi \in \Pi(\mu,\nu)$ achieving this optimal value is denoted by $\Pi_0(\pi,\nu)$ and is referred to as the set of optimal transport plans between $\mu$ and $\nu$. The space of probability measures $\mathcal{P}_p(\mathbb{R}^d)$ endowed with the $p$–Wasserstein distance is called the Wasserstein space of order $p$.

Finally, in order to help the reader, we report in Table 1 a list of the microscopic and mean–field objects used in the analysis performed in the subsequent sections.

3.1. *Well–posedness of weak solutions*

We notice that the microscopic system (3) describing a neural differential equation can be recast as an autonomous system using the auxiliary variables $\tau_i = \tau_i(t) \in \mathbb{R}$ for $i = 1, \dots, M$:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} x_i(t) = \sigma\big(w(\tau_i(t))x_i(t) + b(\tau_i(t))\big), & x_i(0) = x_i^0 \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\tau_i(t) = 1, & \tau_i(0) = 0. \end{cases} \tag{6}$$

In the following, the right hand side of (6) will be compactly denoted using the function

$$\begin{aligned} G : \mathbb{R}^{d+1} &\to \mathbb{R}^{d+1} \\ (x, \tau) &\mapsto \Big(\sigma\big(w(\tau)x + b(\tau)\big), 1\Big)^\top. \end{aligned} \tag{7}$$

**Definition 3.1.** Let $T > 0$ be fixed. Assume that $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$. We say that the time dependent measure $F_t \in C([0,T]; \mathcal{P}_1(\mathbb{R}^{d+1}))$ is a weak solution to the mean–field equation

$$\partial_t F_t + \nabla_x \cdot \Big(\sigma\big(w(\tau)x + b(\tau)\big)F_t\Big) + \partial_\tau F_t = 0 \tag{8}$$

with initial condition $F_0$ if for all $\phi = \phi(x,\tau) \in C_0^\infty(\mathbb{R}^{d+1})$ and for all $t \in [0,T]$ the following equality holds:

$$\begin{aligned} \int_{\mathbb{R}^{d+1}} \phi(x,\tau)\mathrm{d}F_t(x,\tau) = &\int_{\mathbb{R}^{d+1}} \phi(x,\tau)\mathrm{d}F_0(x,\tau) \\ &+ \int_0^t \int_{\mathbb{R}^{d+1}} \nabla_{(x,\tau)}\phi(x,\tau) \cdot G(x,\tau)\mathrm{d}F_s(x,\tau)\mathrm{d}s. \end{aligned} \tag{9}$$

Existence and uniqueness of a weak solution $F_t$ of the mean–field equation (5) is obtained under the following assumptions, see [23, Section 6.1 and Section 6.2] and Proposition 3.1 below:

$$\begin{aligned} &\text{(A1)} \qquad \sigma \in C^{0,1}(\mathbb{R}^d), \hat{A} \quad w, b \in C^{0,1}(\mathbb{R}); \\ &\text{(A2)} \qquad |\sigma(x)| \le C_0, \ \forall x \in \mathbb{R}^d. \end{aligned}$$

**Remark 3.1.** We observe that Assumption (A2) requires that the activation function $\sigma$ is bounded. This property is verified for some choices of the activation function, e.g. if $\sigma$ is the hyperbolic tangent function or the sigmoid function, but not in general. However, the results of this section are true if the kinetic measure $F_0$ has compact support, which implies that any $\sigma$ is bounded on the support of $F_t$ for all $t \ge 0$.

**Definition 3.2.** We define the flow associated to the mean–field equation (8) as the map $\Phi_t : (x,\tau) \in \mathbb{R}^{d+1} \mapsto \Phi_t(x,\tau) \in \mathbb{R}^{d+1}$ such that

$$\begin{cases} \partial_t \Phi_t(x,\tau) = G(\Phi_t(x,\tau)) \\ \Phi_0(x,\tau) = (x,\tau). \end{cases} \tag{10}$$

**Proposition 3.1.** *Let $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ be given and let $T > 0$. Then, under the assumptions (A1) and (A2), there exists a unique solution $F_t \in C([0,T];\mathcal{P}_1(\mathbb{R}^{d+1}))$ of the mean–field equation (8), in particular $F_t = \Phi_t \# F_0$ and $F_t$ is continuously dependent on the initial data $F_0$ with respect to the $1-$Wasserstein distance.Furthermore, the solution of the dynamical system (6)–(7) converges to $F_t$ in $1-$Wasserstein for $M \to \infty$.*

**Proof.** Under the assumptions (A1) and (A2) we have that $G$ defined by equation (7) is also Lipschitz and uniformly bounded for all $(x,\tau) \in \mathbb{R}^{d+1}$. Hence, due to [23, Lemma 6.1], the flow $\Phi_t$ introduced in Definition 3.2 is well–defined and Lipschitz in $(x,\tau)$ and $F_t = \Phi_t \# F_0$ for $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ is the unique weak solution of equation (8) in the sense of (9). Furthermore, under the assumptions (A1) and (A2), any two weak solutions $F_t^{(1)}, F_t^{(2)}$ in the sense of equation (9), obtained from initial conditions $F_0^{(1)}, F_0^{(2)}$, respectively, fulfill the Dobrushin's stability estimate in $1-$Wasserstein distance. The Dobrushin's inequality allows us to prove the convergence of the solutions of the dynamical system (6)–(7) to $F_t$. In fact, we first observe that if we consider the initial condition

$$\mathrm{d}F_0^M(x,\tau) = \frac{1}{M}\sum_{i=1}^{M} \delta(x - x_i^0)\delta(\tau) \tag{11}$$

with $x_i^0$ prescribed by (6), then the following empirical measure

$$\mathrm{d}F_t^M(x,\tau) = \frac{1}{M}\sum_{i=1}^{M} \delta(x - x_i(t))\delta(\tau - \tau_i(t))$$

is a weak solution of (8) in the sense of (9), where $(x_i(t), \tau_i(t))$ are the trajectories given by the dynamical system (6) for any $i = 1, \dots, M$. The previous consideration follows from a classical derivation, see e.g. [38]. Hence, if the initial empirical measure (11) converges in $1-$Wasserstein distance $W_1$ to some $\bar{F}_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ for $M \to \infty$, using the Dobrushin's estimate

$$W_1\big(\bar{F}_t, F_t^M\big) \le C\, W_1\big(\bar{F}_0, F_0^M\big),$$

with $C$ being a constant and $\bar{F}_t = \Phi_t \# \bar{F}_0$, we obtain that (8) is the mean–field limit of the particle dynamics (6) for $M \to \infty$. □

The previous proposition shows that the mean–field limit can be obtained provided that the controls $w, b \in C^{0,1}(\mathbb{R})$. As further result we establish the dependence on the functions $(w,b)$.

**Proposition 3.2.** *Let $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ be given and let $T > 0$. Then, under the assumptions (A1) and (A2), the unique solution $F_t \in C([0,T];\mathcal{P}_1(\mathbb{R}^{d+1}))$ of the mean–field equation (8) is continuously dependent on $(w,b)$.*

**Proof.** Denote by $\Phi^{(w,b)}$ the flow defined by equation (10) with given $(w,b)$. Then, for any $(w,b)$ fulfilling (A1) and $\sigma$ fulfilling (A2) the assumptions of [23, Proposition 7.2] are satisfied and we obtain for $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$

$$W_1(\Phi^{(w,b)}\#F_0, \Phi^{(\bar{w},\bar{b})}\#F_0) \leq \frac{\exp(Lt-1)}{L}\|(w,b)-(\bar{w},\bar{b})\|_{C^0(0,T)}, \tag{12}$$

where $L = \max\{L_{G(w,b)}, L_{G(\bar{w},\bar{b})}\}$ is the maximum of the Lipschitz constants of $G$ defined by equation (7). □

**Proposition 3.3.** *If a weak solution $F_t \in C([0,T];\mathcal{P}_1(\mathbb{R}^{d+1}))$ of (8) fulfills*

$$\mathrm{d}F_t(x,\tau) = \mathrm{d}f_t(x)\delta(\tau - t) \tag{13}$$

*with $f_t \in C([0,T];\mathcal{P}_1(\mathbb{R}^d))$, and if $\mathrm{d}F_0(x,\tau) = \mathrm{d}f_0(x)\delta(\tau)$ with $f_0 \in \mathcal{P}_1(\mathbb{R}^d)$, then, under the assumptions (A1) and (A2), $f_t$ is a weak solution of the mean–field equation (5) with initial condition $f_0$.*

**Proof.** Using the assumptions on $F_t$ and $F_0$ in (9) we find for all $\phi = \phi(x) \in C_0^\infty(\mathbb{R}^d)$:

$$\begin{aligned}\int_{\mathbb{R}^d} \phi(x)\mathrm{d}f_t(x) = &\int_{\mathbb{R}^d} \phi(x)\mathrm{d}f_0(x)\\ &+ \int_0^t \int_{\mathbb{R}^d} \nabla_x\phi(x)\cdot\sigma\big(w(t)x + b(t)\big)\mathrm{d}f_s(x)\mathrm{d}s,\end{aligned}$$

which is exactly the weak form of the mean–field equation (5) with initial condition $f_0$. □

Provided that the initial data has the decomposition given in the previous Proposition, the above computation shows that then there exists a solution $\mathrm{d}F_t$ fulfilling (13) and $f_t$ being a solution to equation (5).

### 3.2. *Mean–field controllability problems*

In the continuous formulation of the neural network, the training step can be seen as controllability problem in the sense of the following definition, see also [39,40].

**Definition 3.3.** Let $f_0, g \in \mathcal{P}_1(\mathbb{R}^d)$ be given. Let $T > 0$ be fixed. We say that the mean–field equation (5) is controllable if there exist $w \in C^{0,1}([0,T];\mathbb{R}^{d\times d})$ and $b \in C^{0,1}([0,T];\mathbb{R}^d)$ such that $(\Phi_T\#f_0) = g$ where $\Phi_t : \mathbb{R}^d \to \mathbb{R}^d$ is the Lipschitz continuous characteristic flow of (5).

In this section we focus on the controllability problem at the mean–field level. We show that for simple problems it is possible to recover explicit results on the solution of the controllability problem. However, so far, a general theory is not available.

**Proposition 3.4.** *Let $T > 0$, $\beta > 0$ be fixed constants such that $\beta/T$ belongs to the image of $\sigma$. Further, denote by $B = (\beta,\dots,\beta)^t \in \mathbb{R}^d$ and $f_0, g \in \mathcal{P}_1(\mathbb{R}^d)$ be given such that $\mathrm{d}g(x) = \mathrm{d}f_0(x-B)$. Then, the mean–field equation (5) is controllable in the sense of Definition 3.3.*

**Proof.** According to Definition 3.3 we only need to show that there exist $w \in C^{0,1}([0,T];\mathbb{R}^{d\times d})$ and $b \in C^{0,1}([0,T];\mathbb{R}^d)$ such that $g$ is the push–forward of $f_0$ under the flow of (5). Taking $w(t) \equiv 0$, the flow $\Phi_t : \mathbb{R}^d \to \mathbb{R}^d$ defined by

$$\partial_t \Phi_t(x) = \sigma(b(t)),\ \Phi_0(x) = x,$$

yields the desired result provided that $b$ fulfills $\int_0^T \sigma(b(t))\mathrm{d}t = B$. There is at least one $b(t) = b_0$ such that the later equality holds. □

Clearly, Proposition 3.4 does not guarantee uniqueness of the choice of the pair $(w, b)$. Consider, e.g., $d = 1$ and the identity activation function $\sigma(x) = x$. Then for any $\beta > 0$ the training process can be solved with $b(t) = t^2 + 1$. Namely, there exists a time $T$ at which $f_T = g$. However, for the same time $T$, the training process is also solved with $b(t) = \beta/T$.

The next proposition is recalled from [16, Proposition 1] and it characterizes steady states of the mean–field equation that are not necessarily unique. The proposition illustrates that if initial and terminal states are a sum of weighted Dirac measures the system is trivially controllable with parameters $(\overline{w}, \overline{b})$ given below. However, as the proposition shows, this only possible if the activation function has sufficiently many zeros.

**Proposition 3.5.** *Let $f_t \in \mathcal{P}_1(\mathbb{R}^d)$ be a compactly supported weak solution of the mean–field equation (5). Assume that the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ have $n$ zeros $z_i$, i.e. $\sigma(z_i) = 0$ for $i = 1, \dots, n$. Let $b(t) = \bar{b} \in \mathbb{R}^d$ and $w(t) = \bar{w} \in \mathbb{R}^{d\times d}$ for all $t \geq 0$. Moreover, assume that $\bar{w}$ has maximum rank. Then,*

$$f_\infty = \sum_{i=1}^{n^d} \rho_i \delta_{y_i}$$

*is a steady state solution of (5) in the sense of measures provided that $y$ is the solution to the system $\bar{w}y + \bar{b} = z$, where $z$ is any disposition with repetition of the $n$ zeros, and where $\rho_i \in [0,1]$, $\forall\, i = 1, \dots, n$, with $\sum_{i=1}^{n^d} \rho_i = 1$.*

The next lemma shows that for particular choices of $\sigma, w, b$ the mean of $f_t$ is preserved.

**Lemma 3.1.** *Let $f_0 \in \mathcal{P}_1(\mathbb{R}^d)$ and let $f_t \in \mathcal{P}_1(\mathbb{R}^d)$ be a solution of the mean–field equation (5) where $m(t) = \int_{\mathbb{R}^d} x\mathrm{d}f_t(x)$, $\forall t \in [0,T]$. Assume that $\sigma(x) = x$ and $b(t) = -w(t)m(t)$, $\forall t \in [0,T]$. Then, $m(t) = m(0)$, $\forall t \in [0,T]$.*

**Proof.** Using equation (5) the first moment $m(t)$ of the probability density $f_t$ satisfies the evolution equation

$$\frac{\mathrm{d}}{\mathrm{d}t} m(t) = w(t)m(t) + b(t).$$

Taking $b(t) = -w(t)m(t)$ the right–hand side vanishes and the assertion follows. □

This lemma can be used to obtain a further controllability result in Proposition 3.6 below. However, we restrict the case to $d = 1$ since for $d > 1$ one has to assume additional assumptions on the matrix $w(t)$ in order to write the solution formula of the flow explicitly.

**Proposition 3.6.** *Let $T > 0$, $\alpha \in \mathbb{R}$ be fixed constants. Let $f_0, g \in \mathcal{P}_1(\mathbb{R})$ be given such that $g = (F^{-1}\#f_0)$ where $F : x \in \mathbb{R} \mapsto F(x) = xe^{\alpha} + (1 - e^{\alpha})m_0 \in \mathbb{R}$ and $m_0 = \int_{\mathbb{R}} x\mathrm{d}f_0(x)$. Assume that $\sigma(x) = x$. Then the mean–field equation (5) is controllable in the sense of Definition 3.3.*

**Proof.** Observe that with the definition of $g$ we have

$$\int_{\mathbb{R}} x\mathrm{d}g(x) = \int_{\mathbb{R}} e^{\alpha} x\mathrm{d}(F^{-1}\#f_0)(x) = \int_{\mathbb{R}} F^{-1}(x)|\mathrm{det}J_{F^{-1}}|\mathrm{d}f_0(x) = m_0.$$

In order to have $m(t) = m_0$ for all times we set $b(t) = -w(t)m_0$ and $\sigma(x) = x$. Further, we choose $w \in C^{0,1}([0,T];\mathbb{R}^{d\times d})$ such that $\int_0^T w(t)\mathrm{d}t = -\alpha$. The flow $\Phi_t : \mathbb{R} \to \mathbb{R}$ is defined by

$$\partial_t \Phi_t(x) = w(t)\left(\Phi_t(x) - m_0\right), \ \Phi_0(x) = x,$$

which yields $\Phi_t(x) = e^{\int_0^t w(s)\mathrm{d}s}x + m_0 \int_0^t -e^{-\int_0^\xi w(s)\mathrm{d}s}w(\xi)\mathrm{d}\xi$ and $\Phi_T^{-1}(y) = xe^{\alpha} - m_0 \int_0^T -e^{-\int_0^\xi w(s)\mathrm{d}s}w(\xi)\mathrm{d}\xi$. Hence,

$$\mathrm{d}f_T(x) = \mathrm{d}(\Phi_T\#f_0)(x) = \mathrm{d}f_0\left(xe^{\alpha} + (1 - e^{\alpha})m_0\right).$$

and the system is controllable, i.e. $f_T = g$. □

Proposition 3.6 shows that the sign of $\alpha$ influences the behavior of the second moment $\int x^2\mathrm{d}f_t(x)$ of $f_t$. In particular, for $\alpha > 0$ it is possible to show that the second moment of $f_t$ decreases and $f_t$ concentrates at the first moment, whereas for $\alpha < 0$ the second moment increases. In fact, without loss of generality, assume that $m_0 = 0$ and let $g$ defined as in Proposition 3.6, then $\int x^2\mathrm{d}g(x) = e^{-3\alpha}\int x^2\mathrm{d}f_0(x)$.

Proposition 3.4 and Proposition 3.6 of this section discuss some prototype situations in which the problem of recovering the target distribution $g$ through the mean–field equation (5) with initial condition $f_0$ is explicitly possible. In general, we follow an alternative method introduced in the subsequent sections to compute the parameters.

### 3.3. *Existence of solutions to the mean–field minimization problem*

As presented in Section 2 the training procedure of a neural network aims to find optimal weights and bias in order to minimize a given distance, e.g. (4). The mean–field interpretation of this training procedure requires first to derive the mean–field limit of the loss function which is given by the following proposition. Note that we have enforced a possibly strong condition on the independence of the data. If this assumption does not hold true, we still obtain a mean field limit of the cost functional. Then, we need to expand the phase space of the mean field equation by an additional dimension $y$ that

links to the distribution $g^M$ in the space of the target values. However, in the following we consider the simpler case, where the data is actually independent.

**Proposition 3.7.** *Let $\{x_i(t), \tau_i(t)\}_{i=1}^M$ be the trajectories given by the dynamical system (6)-(7) with initial conditions $\{x_i^0, 0\}_{i=1}^M$ and let $\{y_i\}_{i=1}^M$ be the given target values which are assumed to be obtained as statistically independent. Let $F_0^M \in \mathcal{P}_1(\mathbb{R}^{d+1})$ and $g^M \in \mathcal{P}_1(\mathbb{R}^d)$ be the empirical measures associated to the initial conditions and the target values, respectively. Furthermore, let $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ and $g \in \mathcal{P}_1(\mathbb{R}^d)$ be such that $W_1(F_0, F_0^M) \to 0$, $W_1(g, g^M) \to 0$, as $M \to \infty$. Then, under the assumptions (A1) and (A2), the mean–field limit of the loss function (4) is*

$$\int_{(x,\tau)\in\mathbb{R}^{d+1}} \tilde{\ell}(x) \mathrm{d}F_T(x,\tau)$$

*where $\tilde{\ell}(x) = \int_{\mathbb{R}^d} \ell(x-y)\mathrm{d}g(y)$ and $F_t$ is the weak solution of (8) obtained with initial condition $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$.*

**Proof.** We notice that the loss function (4) can be written as

$$\frac{1}{M}\sum_{i=1}^M \ell(x_i(T) - y_i) = \int_{\mathbb{R}^{2d+1}} \ell(x-y)\mathrm{d}\mu_T^M(x,y,\tau)$$

where $\mu_t^M \in \mathcal{P}_1(\mathbb{R}^{2d+1})$ is the time dependent empirical measure

$$\mathrm{d}\mu_t^M(x,y,\tau) = \frac{1}{M}\sum_{i=1}^M \delta(x - x_i(t))\delta(y-y_i)\delta(\tau - t) \tag{14}$$

We observe that $\mu_T^M$ has marginals

$$\int_{(x,\tau)\in\mathbb{R}^{d+1}} \mathrm{d}\mu_T^M(x,y,\tau) = \frac{1}{M}\sum_{i=1}^M \delta(y-y_i) = \mathrm{d}g^M(y),$$
$$\int_{y\in\mathbb{R}^d} \mathrm{d}\mu_T^M(x,y,\tau) = \frac{1}{M}\sum_{i=1}^M \delta(x - x_i(T))\delta(\tau - t) = \mathrm{d}F_T^M(x,\tau).$$

The prediction and target values are assumed to be statistically independent and therefore $x_i(T)$ solely depends on $x_i(0)$ which yields

$$\mathrm{d}\mu_T^M(x,y,\tau) = \mathrm{d}F_T^M(x,\tau)\mathrm{d}g^M(y).$$

By the Glivenko–Cantelli's theorem, see e.g. [41,42], we have that there exists $g \in \mathcal{P}(\mathbb{R}^d)$ and a time dependent measure $\mu_t \in \mathcal{P}(\mathbb{R}^{2d+1})$ such that $W_1(g^M, g) \to 0$ and $W_1(\mu_t^M, \mu_t) \to 0$, as $M \to \infty$. In addition, under the assumptions (A1) and (A2), the mean–field convergence result of Section 3.1 implies $W_1(F_t^M, F_t) \to 0$, $\forall t \in [0,T]$ as

$M \to \infty$, with $F_t \in \mathcal{P}_1(\mathbb{R}^{d+1})$ weak solution of (8) obtained with initial condition $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ such that $W_1(F_0, F_0^M) \to 0$, as $M \to \infty$. Then, noticing that

$$W_1(\mu_T^M, F_T g) \leq W_1(F_T^M g^M, F_T^M g) + W_1(F_T^M g, F_T g) \to 0 \quad \text{as } M \to \infty,$$

we conclude that the mean–field limit of the loss function (4) is

(15)
$$\int_{(x,y,\tau)\in\mathbb{R}^{2d+1}} \ell(x-y)\mathrm{d}\mu_T(x,y,\tau) = \int_{(x,\tau)\in\mathbb{R}^{d+1}} \tilde{\ell}(x)\mathrm{d}F_T(x,\tau), \quad \tilde{\ell}(x) = \int_{\mathbb{R}^d} \ell(x-y)\mathrm{d}g(y). \qquad \square$$

Consider now the cost functional $J : \mathcal{P}_1(\mathbb{R}^{d+1}) \to \mathbb{R}$ given by

$$J(\mu) = \int_{\mathbb{R}^{d+1}} \tilde{\ell}(x)\mathrm{d}\mu(x,\tau), \tag{16}$$

cf. the definition of the cost of the transport plan at the beginning of Section 3. In the following we discuss the existence of solutions to the mean–field minimization problem

$$(w,b) \mapsto \min J(F_T)$$

on a suitable subset $X$ of controls $(w,b)$, where $F_T$ is the unique weak solution to equation (8) for fixed initial datum $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$. We observe that the mean–field cost functional $\tilde{\ell}$ derived in (15) is bounded and Lipschitz continuous provided that $\ell \in C^{0,1}(\mathbb{R}^d)$ is bounded from below. In fact:

$$\|\tilde{\ell}(x)-\tilde{\ell}(z)\| \leq \int_{\mathbb{R}^d} \|\ell(x-y)-\ell(z-y)\|\mathrm{d}g(y) \leq L\int_{\mathbb{R}^d} \|x-z\|\mathrm{d}g(y) = L\|x-z\|, \quad \forall x,z \in \mathbb{R}^d,$$

with $L$ Lipschitz constant of $\ell$. In order to simplify the notation we denote by

$$\begin{aligned} u &:= (w,b) \in C^{0,1}([0,T];\mathbb{R}^{d\times d}\times\mathbb{R}^d), \\ \mu_u &:= F_T \in \mathcal{P}_1(\mathbb{R}^{d+1}). \end{aligned}$$

For fixed $T>0$ and $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ the reduced cost functional is then defined by

$$j(u) = J(\mu_u) = \int_{\mathbb{R}^{d+1}} \tilde{\ell}(x)\mathrm{d}\mu_u(x,\tau).$$

Since $\tilde{\ell}$ is bounded from below, we obtain that $j$ is bounded from below. Since $\tilde{\ell}$ is Lipschitz with constant $L$ we obtain the following estimate for $u, v \in C^{0,1}([0,T];\mathbb{R}^{d\times d}\times\mathbb{R}^d)$ constrained to $u(0)=v(0)=0$:

$$\begin{aligned} |j(u)-j(v)| &= L\left\|\int_{\mathbb{R}^{d+1}} \frac{\tilde{\ell}(x)}{L}\mathrm{d}(\mu_u-\mu_v)\right\| \\ &\leq L\sup\left\{\int_{\mathbb{R}^{d+1}} \phi(x)\mathrm{d}(\mu_u-\mu_v) : \phi \text{ 1–Lipschitz}\right\} \\ &= LW_1(\mu_u,\mu_v) \leq C(L_u,L_v)\|u-v\|_{C^0}, \end{aligned}$$

where the last inequality follows by (12) and the constant $C$ depends on the Lipschitz constants of $u$ and $v$. Thus, the loss function $j$ is continuous with respect to the $C^0-$norm. The previous results can be used to establish existence of minimizers using the direct method of variation.

**Proposition 3.8.** *Assume that the assumptions (A1) and (A2) are fulfilled. Let $\tilde{\ell} \in C^{0,1}(\mathbb{R}^d)$ and bounded from below. Assume $T > 0, L > 0$ and $F_0 \in \mathcal{P}_1(\mathbb{R}^{d+1})$ are given. Then, there exists a solution to the minimization problem*

$$\min_{(w,b)\in X} \int_{\mathbb{R}^{d+1}} \tilde{\ell}(x) \mathrm{d}F_T(x,\tau),$$

*where $F_t \in C([0,T];\mathcal{P}_1(\mathbb{R}^{d+1}))$ is the weak solution to equation (8) and where*

$$X = \{(w,b) \in C^{0,1}([0,T];\mathbb{R}^{d\times d} \times \mathbb{R}^d) : L_w + L_b \leq L,\ w(0) = b(0) = 0\} \tag{17}$$

*with $L_w, L_b$ Lipschitz constants of $w, b$, respectively.*

**Proof.** For the proof of the previous proposition we proceed as follows. Since the cost functional is bounded from below, there exists a minimizing sequence $(u_n)_{n\geq 0} \subset X$. According to the definition of $X$ we have that $\|u_n\|_{C^0} \leq L$ for all $n$. Further, we have that $u_n$ is uniformly Lipschitz continuous due to definition of $X$. Hence, the assertion of the Arzela–Ascoli are fulfilled and $u_n$ converges in $C^0$ to $u \in C^0([0,T];\mathbb{R}^{d\times d} \times \mathbb{R}^d)$. Furthermore, it holds that $u \in C^{0,1}([0,T];\mathbb{R}^{d\times d} \times \mathbb{R}^d)$ with Lipschitz constant bounded by $L$. Finally, the continuity of $j$ shown above yields that $u$ is the minimizer, i.e., $j(u) = \lim_{n\to\infty} \hat{A}\ j(u_n)$. This finishes the proof. □

## 4. Computational approach to the mean–field training procedure

In this section we explicitly formulate the training processes for the mean–field limit (5) of the neural differential equation (3) in terms of an optimal control problem. While Section 3.3 discusses existence of minimizers to the mean–field control problem, here, we formally derive a first–order optimality system in order to design a numerical method for the optimization of the parameters $w(t)$ and $b(t)$ of the neural network. Note that the previous theorem does not allow for a characterization due to a lack of regularity of the solution in terms of the parameters.

### 4.1. *Formulation of the computational approach*

Our computational approach is based on the minimization problem of a general functional of equation (16) for $(w,b) \in X$, cf. (17):

$$\min_{(w,b)\in X} \int_{\mathbb{R}^{d+1}} \tilde{\ell}(x) \mathrm{d}F_T(x,\tau) + \frac{\gamma}{2} \int_0^T \|w\|^2 + \|b\|^2 \mathrm{d}t, \hat{\mathrm{A}} \text{ subject to}$$

$F_T \hat{\mathrm{A}}$ being the solution at time $T$ of equation (8) and initial condition $F_0$.

We assume the initial $F_0$ and solution $F_T$ are of the type (13) $\forall(t,x,\tau)$. Furthermore, we use a Tikhonov regularization for the controls with a parameter $\gamma > 0$ but the proposed computational approach below works also in the case $\gamma = 0$. Under the structural assumption (13) the previous problem reduces to a constrained optimal control problem

$$\min_{(w,b)\in X} \int_{\mathbb{R}^d} \tilde{\ell}(x)\mathrm{d}f_T(x) + \frac{\gamma}{2}\int_0^T \|w\|^2 + \|b\|^2 \mathrm{d}t\hat{A}$$
$$\text{subject to} \quad \begin{cases} \partial_t f_t(x) + \nabla_x \cdot \Big(\sigma\big(w(t)x + b(t)\big) f_t(x)\Big) = 0, \\ f_{t=0}(x) = f_0(x). \end{cases}$$

Only formally, a first–order optimality system in strong form is derived

$$\partial_t f_t(x) + \nabla_x \cdot \Big(\sigma\big(w(t)x + b(t)\big) f_t(x)\Big) = 0, \quad f_{t=0}(x) = f_0(x), \tag{18a}$$

$$\partial_t \Lambda_t(x) + \nabla_x \Lambda_t(x) \cdot \sigma\big(w(t)x + b(t)\big) = 0, \quad \Lambda_{t=T}(x) = \tilde{\ell}(x), \tag{18b}$$

$$\gamma b_j(t) + \int_{\mathbb{R}^d} \partial_{x_j}\Lambda_t(x)\sigma_j'\big(w(t)x + b(t)\big) f_t(x)\mathrm{d}x = 0, \quad j = 1,\dots,d, \tag{18c}$$

$$\gamma w_{j,k}(t) + \int_{\mathbb{R}^d} \partial_{x_j}\Lambda_t(x)\sigma_j'\big(w(t)x + b(t)\big) x_k f_t(x)\mathrm{d}x = 0, \quad j,k = 1,\dots,d, \tag{18d}$$

where $b_j$ represents the $j$–th component of the bias vector $b$, $w_{j,k}$ is the entry $(j,k)$ of the weight matrix $w$ and, finally, $\sigma_j'$ represents the derivative of the activation function $\sigma$ computed on the $j$–th component of its argument. Note that the constraint $w(0) = 0 \in \mathbb{R}^{d\times d}$ and $b(0) = 0 \in \mathbb{R}^d$ are enforced in the numerical method. We further formally differentiate the equation for $\Lambda_t$ with respect to $x_j$ for $j = 1,\dots,d$. This also yields a conservative formulation for each $\partial_{x_j}\Lambda_t$. Furthermore, we transform time $t \mapsto T - t$ in equation (18b) in order to obtain an initial value problem. Since we implement numerical results in the case $d = 1$ we state the resulting system where $\lambda_t(x) = \partial_x \Lambda_{T-t}(x)$

$$\partial_t f_t(x) + \partial_x\Big(\sigma\big(w(t)x + b(t)\big) f_t(x)\Big) = 0, \quad f_{t=0}(x) = f_0(x), \tag{19a}$$

$$\partial_t \lambda_t(x) - \partial_x\Big(\sigma\big(w(T-t)x + b(T-t)\big)\lambda_t(x)\Big) = 0, \quad \lambda_{t=0}(x) = \partial_x\tilde{\ell}(x), \tag{19b}$$

$$\gamma b(t) + \int_{\mathbb{R}} \lambda_{T-t}(x)\sigma'\big(w(t)x + b(t)\big) f_t(x)\mathrm{d}x = 0, \tag{19c}$$

$$\gamma w(t) + \int_{\mathbb{R}} \lambda_{T-t}(x)\sigma'\big(w(t)x + b(t)\big) x f_t(x)\mathrm{d}x = 0. \tag{19d}$$

Observe that (19a) and (19b) are decoupled due to the definition of the loss function, namely the initial state of (19b) does not depend on the final state of (19a). This is due to the fact that the mean–field loss function is linear in the state.

#### 4.1.1. *Numerical discretization scheme*

The optimality system is solved in a block Gauss–Seidel fashion, i.e., at the $k$–th iteration $t \mapsto (w,b)^k(t)$ we compute $f_t^k$ and $\lambda_t^k$ as numerical solution to equation (19a)

and (19b), respectively. Details on the numerical scheme will be presented below. The new iterates $(w,b)^{k+1}$ are found through iteration on equations (19c)–(19d), namely for $t>0$ we define

$$b^*(t;\rho) = b^k(t) - \rho^* \left( \gamma b^k(t) + \int_{\mathbb{R}} \lambda_{T-t}(x)\sigma'\big(w^k(t)x + b^k(t)\big) f_t(x)\mathrm{d}x \right),$$

$b^*(0)=0$, and similarly $w^*$. Here, $\rho^*>0$ is a stepsize parameter chosen using backtracking line search, e.g. the Armijo rule, in order to minimize the reduced cost functional

$$\rho^* = \underset{\rho>0}{\operatorname{argmin}} \left( \int_{\mathbb{R}} \tilde{\ell}(x)\mathrm{d}f_T(x;\rho) + \frac{\gamma}{2}\int_0^T \|w^*(t;\rho)\|^2 + \|b^*(t;\rho)\|^2 \mathrm{d}t \right)$$

where $f_T(x;\rho)$ is the solution to (19a) for $(w,b)=(w^*(t;\rho),b^*(t;\rho))$. The new iterates are then obtained by

$$(w,b)^{k+1}(t) := (w^*,b^*)(t;\rho^*),\ \forall t\geq 0.$$

The procedure is repeated $k=0,1,\dots$ until the error $e^{(k)}$ is below a given tolerance $TOL$:

$$e^{(k+1)} := \frac{\|(w,b)^{k+1}-(w,b)^k\|_{C^0(0,T)}}{\|(w,b)^{k+1}\|_{C^0(0,T)}} \leq TOL.$$

For more details on the iterative scheme described above we refer, e.g., to [43].

**Remark 4.1.** In the case $\sigma(x)=x$ further simplifications are possible. In fact, we may derive explicit equations for the evolution of the $k$–th moment of $\lambda\ f$ given by

$$\partial_t \int_{\mathbb{R}} x^k \lambda_{T-t}(x) f_t(x)\mathrm{d}x = -(k+1)w(t)\int_{\mathbb{R}} x^k \lambda_{T-t}(x) f_t(x)\mathrm{d}x.$$

This allows to obtain $(w,b)$ in closed form

$$b(t) = \frac{1}{\gamma}\exp(w(t))\int_{\mathbb{R}} \lambda_T(x) f_0(x)\mathrm{d}x, \tag{20}$$

$$w(t)\exp(-2w(t)) = \frac{1}{\gamma}\int_{\mathbb{R}} x\lambda_T(x) f_0(x)\mathrm{d}x. \tag{21}$$

In this case it is sufficient to iterate equations (19b) and equations (20)–(21) removing the need to solve equation (19a).

The numerical solution of the PDEs (19a) and (19b) is computed with a third–order finite volume scheme [44], which is briefly described below. Both equations are recast in the following compact formulation:

$$\partial_t u(t,x) + \partial_x \mathcal{L}(u(t,x),t,x) = 0, \tag{22}$$

with $\mathcal{L}$ linear operator with respect to $u$. Since the two PDEs are decoupled, they can be solved simultaneously. Application of the method of lines to (22) on discrete cells $\Omega_j$, defining a discretization of the physical domain $\Omega$, leads to the coupled system of ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{U}_j(t) = -\frac{1}{\Delta x}\left[\mathcal{F}_{j+\frac{1}{2}}(t) - \mathcal{F}_{j-\frac{1}{2}}(t)\right], \tag{23}$$

where $\overline{U}_j(t)$ is the approximation of the cell average of the exact solution $u$ in the cell $\Omega_j$ at time $t$. Here, $\mathcal{F}_{j+\frac{1}{2}}(t)$ approximates $\mathcal{L}(u(t, x_{j+\frac{1}{2}}), t, x_{j+\frac{1}{2}})$ with suitable accuracy and is computed as a function of the boundary extrapolated data $U^{\pm}_{j+\frac{1}{2}}(t)$, i.e.

$$\mathcal{F}_{j+\frac{1}{2}}(t) = \mathcal{F}(U^{+}_{j+\frac{1}{2}}(t), U^{-}_{j+\frac{1}{2}}(t))$$

and $\mathcal{F}$ is a consistent and monotone numerical flux, evaluated on two estimates of the solution at the cell interface. We focus on the class of central schemes, in particular we consider $\mathcal{F}$ as a local Lax–Friedrichs flux. In order to construct a third–order scheme the values $U^{\pm}_{j+1/2}(t)$ at the cell boundaries are computed with the third–order CWENO reconstruction [44].

System (22) is finally solved by the classical third–order (strong stability preserving) SSP Runge–Kutta with three stages [45]. At each Runge–Kutta stage, the cell averages are used to compute the reconstructions via the CWENO procedure and the boundary extrapolated data are fed into the Lax–Friedrichs numerical flux. The initial data are computed with the three point Gaussian quadrature. The time step $\Delta t$ is chosen fixed in order to have a fixed grid in time and to avoid a reconstruction in time of the control functions $w(t)$ and $b(t)$ between different iterates of the Gauss–Seidel approach. All the simulations are run with a CFL of 0.45. The other parameters of the simulations are specified in each numerical example separately.

### 4.2. *Computational results*

We present three numerical experiments in order to illustrate the numerical solution of the training of the mean–field neural network and to numerically observe the theoretical findings on the controllability approach presented in Section 3.2.

As loss function we use

$$\ell(x - y) = |x - y|^2 \tag{24}$$

and the initial condition of equation (19b) is then given by

$$\lambda_0(x) = 2x \int_{\mathbb{R}} \mathrm{d}g(y) - 2 \int_{\mathbb{R}} y \mathrm{d}g(y) = 2x - 2m_g$$

where $m_g$ denotes the expected value of the target $g$.

**Remark 4.2.** The initial condition of the equation for $\lambda$ depends only on the expected value of the target. Hence, if we consider two different targets $g_1$ and $g_2$ such that

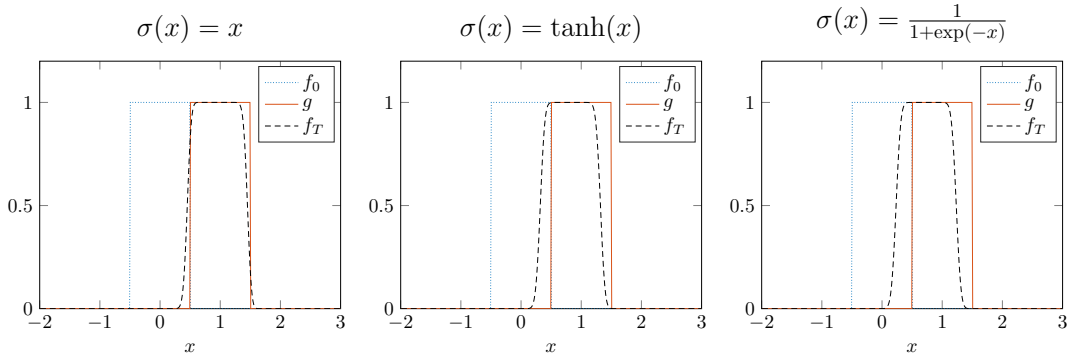


Figure 1. Final states $f_T$ (black dashed lines) at time $T = 1$ obtained with the optimal controls $w(t)$ and $b(t)$ computed by the block Gauss–Seidel approach with tolerance $TOL = 10^{-4}$. Three different activation functions are considered and specified in the panel titles. The blue dotted lines represent the initial state $f_0$, whereas the red solid lines represent the target $g$.

$m_{g_1} \neq m_{g_2}$, we expect to be able to recover correctly the expected value only. Similarly, if we consider two different targets $g_1$ and $g_2$ such that $m_{g_1} = m_{g_2}$, we expect not to learn, e.g., $g_2$ from $g_1$. For example, an $L^2_x$ distance between the final state $f_T$ and the target $g$ would also lead to a dependence on the full state. However, the formal mean–field of the discrete loss function does not include this choice as shown in the previous section.

**Test 1.** In the first example we provide a numerical evidence of the controllability problem proposed in Proposition 3.4. We choose the initial condition

$$f_0(x) = \chi_{[-\frac{1}{2},\frac{1}{2}]}(x)$$

on the physical domain $x \in \Omega = [-2, 3]$, and the target is

$$g(x) = f_0(x - \beta)$$

with $\beta = 1$. The time at which we aim to recover the target $g$ is $T = 1$. We consider a fixed time step $\Delta t = 10^{-2}$ and the space domain is discretized with 200 cells. The regularization parameter is $\gamma = 10^{-3}$ and the tolerance for the stopping criterion is $TOL = 10^{-4}$. The maximum number of iteration of the Armijo–stepsize rule is 10. Three different activation functions are considered, $\sigma(x) = x$, $\sigma(x) = \tanh(x)$ and $\sigma(x) = \frac{1}{1+\exp(-x)}$. The initial guess of the controls is $w^0(t) = 0$ and $b^0(t) = 0$, $\forall\, t \in [0, 1]$.

In Figure 1 we compare the final states $f_T$ (black dashed lines) obtained with the three different activation functions and the optimal controls $w(t)$, $b(t)$ which are shown in the bottom panels of Figure 2. We observe that using the identity activation function provides a better approximation of the target $g$, whereas for the other two activation functions additional iterations of the optimization procedure are required. In fact, the

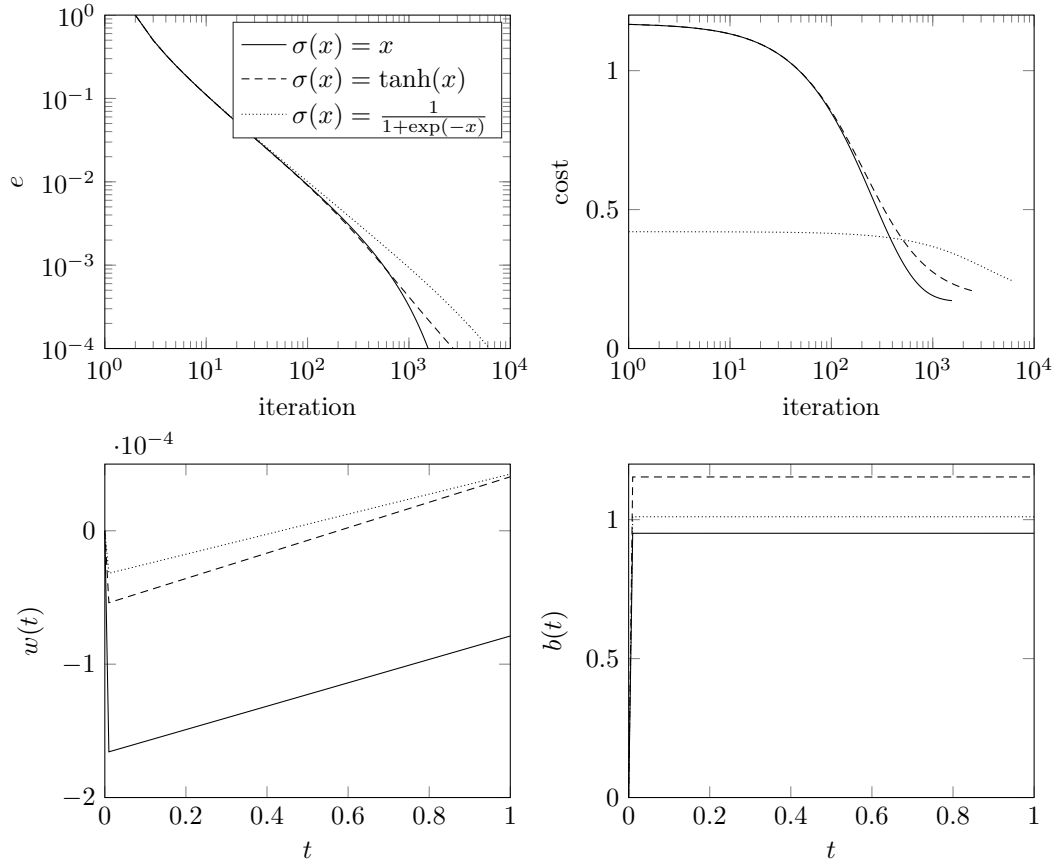


Figure 2. Top left panel: behavior of the relative errors between two consecutive iterations of the controls. Top right panel: value of the cost functionals at each iteration. Bottom panels: optimal controls $w(t)$ (left) and $b(t)$ (right). In all plots, the solid line represent the case of the identity activation function, the dashed line the hyperbolic tangent, and the dotted line the sigmoid.

top panels of Figure 2 shows that, while the relative error between two iterates of the controls reaches the given tolerance $TOL$ for all the activation functions, the values of the cost functional for the case of the hyperbolic tangent and of the sigmoid are larger than the value of the cost functional obtained with the identity and are still decreasing towards a minimum value. Furthermore, we observe that the initial guess of the controls is a better choice for the sigmoid activation. The optimal controls are depicted in the bottom panels of Figure 2 and they are $w(t) \approx 0$ and $b(t) = C$ with $C$ positive constant which depends on the choice of the activation function. Observe, in particular, that $C \approx 1$ for the identity activation, which means that equation (19a) reduces

$$\partial_t f_t(x) + C\partial_x f_t(x) = 0$$

whose solution at $T = 1$ is $f_T(x) = f_0(x-C) \approx g(x) = f_0(x-1)$. This result is consistent

with Proposition 3.4.

**Test 2.** In the second example we provide a numerical evidence of the controllability problem proposed in Proposition 3.6. We choose the initial condition

$$f_0(x) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-\mu)^2}{2s^2}}$$

on the physical domain $x \in \Omega = [-2, 3]$ with $s = 0.1$ and $\mu = 1$. The target is

$$g(x) = f_0(xe^{\alpha} + (1 - e^{\alpha})\mu)e^{\alpha}$$

with $\alpha = 0.25$. The time at which we aim to recover the target $g$ is $T = 1$. We consider a fixed time step $\Delta t = 10^{-2}$ and the space domain is discretized with 400 cells. The regularization parameter is $\gamma = 10^{-3}$ and the tolerance for the stopping criterion is $TOL = 10^{-4}$. The maximum number of iteration of the Armijo–stepsize rule is 10. Due to Lemma 3.1 and Proposition 3.6 we take the identity activation functions $\sigma(x) = x$. The initial guess of the controls is $w(t) = 0$ and $b(t) = 0$, $\forall\, t \in [0, 1]$.

In the top left panel of Figure 3 we observe that the final state $f_T$ (black dashed line) recovers the target $g$. The optimal controls $w(t)$ and $b(t)$ are shown in the top right panel and chosen when the stopping criterion is met, see the bottom left panel. According to Lemma 3.1 and Proposition 3.6 we expect to have $w(t) = -b(t)\mu$ and, since $\mu = 1$, we notice that indeed $w(t) + b(t) \approx 0$. The relative error between cost functional, see the bottom right panel of Figure 3, is monotone decreasing.

**Test 3.** In the last numerical example we build an artificial test and consider exact controls

$$w_e(t) = e^t - 1, \quad b_e(t) = -5t^2 + t$$

to evolve the PDE (19a) up to time $T = 1$ starting from a Beta distribution as initial condition:

$$f_0(x) = \frac{x^{a_1-1}(1-x)^{a_2-1}}{B(a_1, a_2)}$$

where $B$ is the Beta function, and $a_1 = 2$, $a_2 = 5$. We obtain a numerical final state that we use as target to initialize the adjoint equation (19b). Finally, the optimality system is solved with the block Gauss–Seidel approach in order to recover the exact controls $w(t)$ and $b(t)$. The physical domain is again $x \in \Omega = [-2, 3]$. We consider a fixed time step $\Delta t = 10^{-2}$ and the spatial domain is discretized with 400 cells. The regularization parameter is considered different for the two controls, precisely we set $\gamma_w = 1$ and $\gamma_b = 10^{-4}$. The tolerance for the stopping criterion is $TOL = 10^{-4}$. The maximum number of iteration of the Armijo–stepsize rule is 10. For this numerical test we choose the sigmoid activation function, i.e. $\sigma(x) = \frac{1}{1+\exp(-x)}$.

In Figure 4 we show the results of the numerical experiment obtained with two different initial controls. In particular, the top row panels refer to $w^0(t) = b^0(t) = t$, $\forall\, t \in [0, 1]$, whereas the bottom row panels refer to the case $w^0(t) = b^0(t) = 0$, $\forall\, t \in [0, 1]$.

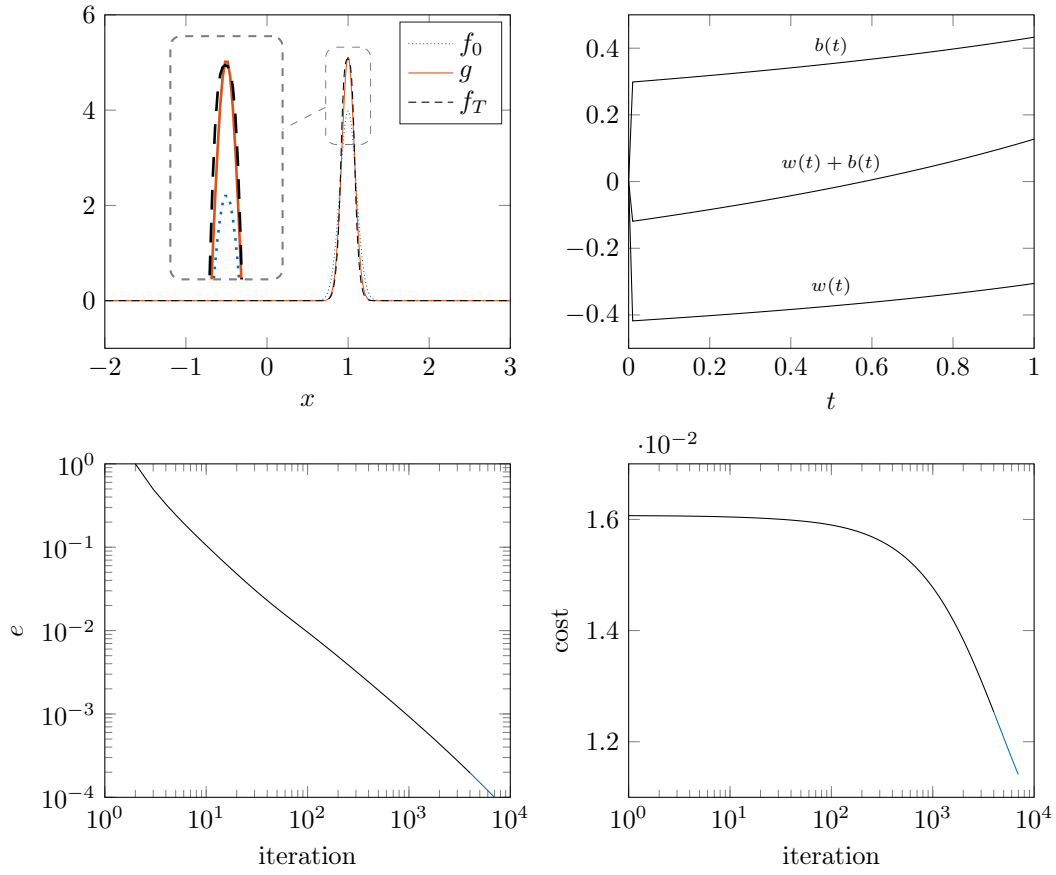


Figure 3. Top left: initial state $f_0$ (blue dotted line), target $g$ (solid red line) and final state $f_T$ (black dashed line) at final time $T = 1$. Top right: optimal controls $w(t)$ and $b(t)$, and their sum on the time interval $[0, 1]$. Bottom left: relative error of two consecutive iterations of the controls. Bottom right: behavior of the cost functional over the iterations.

We notice that in both cases the final state $f_T$, black dashed line in the top left panel, reproduces the expected value of the target, but the method is failing in estimating the variance and the height of the extremal point. This is a consequence of the particular choice of the loss function, as already pointed out in Remark 4.2. The optimal controls computed at the end of the optimization procedure are shown in the center column panels and compared with the exact controls $w_e(t)$ and $b_e(t)$. In both cases, the method provides a constant weight, precisely $w(t) \approx 10^{-3}$, whereas $b(t)$ differs, depending on the choice of the initial guess. This result show the possible existence of multiple optimal controls solving the same task of recovering the target $g$.

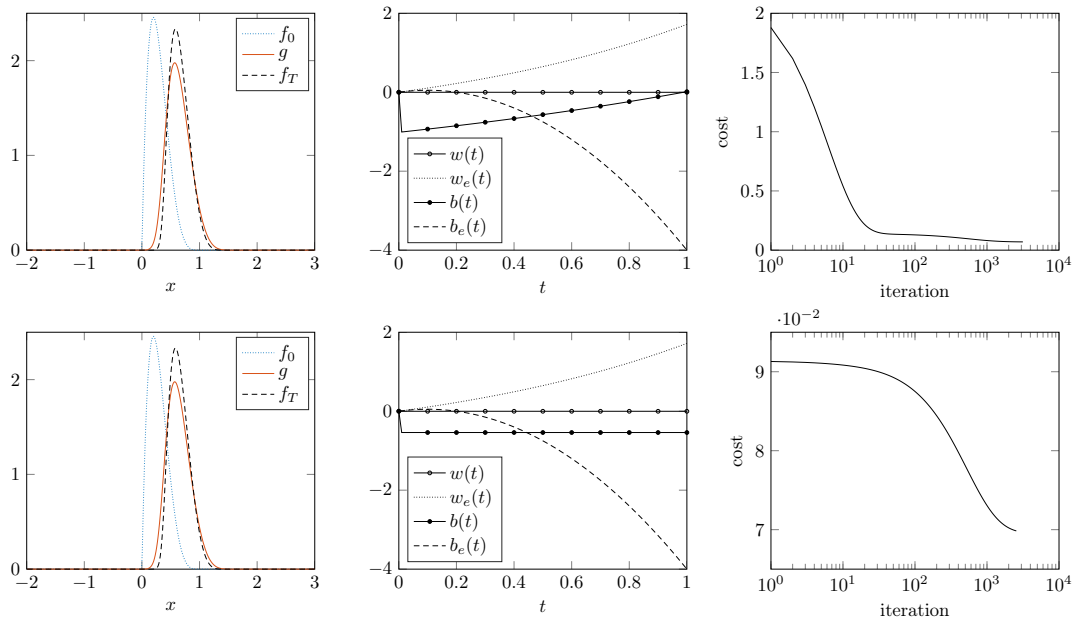


Figure 4. Top row: initial guess for the controls is $w(t) = b(t) = t$, $\forall\, t \in [0,1]$. Bottom row: initial guess for the controls is $w(t) = b(t) = 0$, $\forall\, t \in [0,1]$. Left column: initial state $f_0$ (blue dotted line), target $g$ (solid red line) and final state $f_T$ (black dashed line) at time $T = 1$. Center column: optimal controls $w(t)$ and $b(t)$, and the exact controls $w_e(t)$ and $b_e(t)$ on the time interval $[0,1]$. Right column: behavior of the cost functional over the iterations.

## 5. Conclusion and Future Work

In this work we have proposed and analyzed a mean–field description of residual neural networks. The limit is performed on the number of data, and the well–posedness of the resulting Vlasov–type equation is discussed. We have proved existence and uniqueness of weak solutions, continuous dependence on the initial condition and on the parameters, and the convergence of the solution of the discrete system to the solution of the PDE.

Furthermore, we have tackled the problem of the training of the mean–field neural network using a controllability and an optimal control point of view. We have shown existence of the minimizers and proposed a computational approach based on first–order optimality conditions to numerically optimize the unknown parameters. Finally, we have performed numerical experiments on the derived equations.

We expect that further analysis of the mathematical formulations of machine learning models at different scales is a useful tool to break the complexity of the methods on discrete level and to provide theoretical foundations, in–depth understanding, analysis and improvements of existing approaches. In particular, the present work opens several research perspectives, as for instance the study of the convergence of the optimal solutions of the discrete training process to the solutions of the mean–field optimal control problem via Gamma–convergence, or the definition of different loss functions at

the mean–field level and the computational comparison between the discrete and the mean–field training.

**Acknowledgments**

M.H. thanks the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for the financial support through 320021702/GRK2326, 333849990/IRTG-2379, B04, B05 and B06 of CRC1481, HE5386/18-1,19-2,22-1,23-1, ERS SFDdM035 and under GermanyÂ's Excellence Strategy EXC-2023 Internet of Production 390621612 and under the Excellence Strategy of the Federal Government and the LÃ□nder. Support through the EU DATAHYKING is also acknowledged. The authors acknowledge the support of the Banff International Research Station (BIRS) for the Focused Research Group [22frg198] Â"Novel perspectives in kinetic equations for emerging phenomenaÂ", July 17-24, 2022, where part of this work was done. G.V. acknowledges the support of the INdAM–GNCS group and of the PRIN2017 project 2017KKJP4X funded by the Italian MUR (Ministry of University and Research). The work of A.T. was supported by a postdoc fellowship of the German Academic Exchange Service (DAAD) (PKZ 91817986).

## 6. Bibliography

### References

1. M. Wooldridge, Artificial Intelligence requires more than deep learning - but what, exactly?, *Artificial Intelligence*, vol. 289, p. 103386, 2020.
2. S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review, *Chaos Solitons Fractals*, vol. 139, pp. 110059, 6, 2020.
3. V. C. Müller and N. Bostrom, Future progress in artificial intelligence: a survey of expert opinion, in *Fundamental issues of artificial intelligence*, vol. 376 of *Synth. Libr.*, pp. 553–570, Springer, [Cham], 2016.
4. K. T. Mengistu and F. Rudzicz, Comparing humans and automatic speech recognition systems in recognizing dysarthric speech, in *Advances in artificial intelligence*, vol. 6657 of *Lecture Notes in Comput. Sci.*, pp. 291–300, Springer, Heidelberg, 2011.
5. C. Li, Y. Xing, F. He, and D. Cheng, A strategic learning algorithm for state-based games, *Automatica J. IFAC*, vol. 113, pp. 108615, 9, 2020.
6. Z. M. Fadlullah, B. Mao, F. Tang, and N. Kato, Value iteration architecture based deep learning for intelligent routing exploiting heterogeneous computing platforms, *IEEE Trans. Comput.*, vol. 68, no. 6, pp. 939–950, 2019.
7. R. E. Stern, S. Cui, M. L. Delle Monache, R. Bhadani, M. Bunting, M. Churchill, N. Hamilton, R. Haulcy, H. Pohlmann, F. Wu, B. Piccoli, B. Seibold, J. Sprinkle, and D. B. Work, Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments, *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 205 – 221, 2018.
8. S. Mishra, A machine learning framework for data driven acceleration of computations of differential equations, *Math. Eng.*, vol. 1, no. 1, pp. 118–146, 2019.

9. K. O. Lye, S. Mishra, and D. Ray, Deep learning observables in computational fluid dynamics, *J. Comput. Phys.*, vol. 410, pp. 109339, 26, 2020.
10. D. Zhang, L. Guo, and G. E. Karniadakis, Learning in modal space: solving time-dependent stochastic PDEs using physics-informed neural networks, *SIAM J. Sci. Comput.*, vol. 42, no. 2, pp. A639–A665, 2020.
11. M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *J. Comput. Phys.*, vol. 378, pp. 686–707, 2019.
12. N. Discacciati, J. S. Hesthaven, and D. Ray, Controlling oscillations in high-order discontinuous Galerkin schemes using artificial viscosity tuned by neural networks, *J. Comput. Phys.*, vol. 409, pp. 109304, 30, 2020.
13. D. Ray and J. S. Hesthaven, Detecting troubled-cells on two-dimensional unstructured grids using a neural network, *J. Comput. Phys.*, vol. 397, pp. 108845, 31, 2019.
14. J. Magiera, D. Ray, J. S. Hesthaven, and C. Rohde, Constraint-aware neural networks for Riemann problems, *J. Comput. Phys.*, vol. 409, pp. 109345, 27, 2020.
15. D. Ray and J. S. Hesthaven, An artificial neural network as a troubled-cell indicator, *J. Comput. Phys.*, vol. 367, pp. 166–191, 2018.
16. M. Herty, T. Trimborn, and G. Visconti, Mean-field and kinetic descriptions of neural differential equations, *Foundations of Data Science*, vol. 4, no. 2, pp. 271–298, 2022.
17. J. Crevat, Mean-field limit of a spatially-extended Fitzhugh-Nagumo neural network, *Kinet. Relat. Models*, vol. 12, no. 6, pp. 1329–1358, 2019.
18. S. Mei, A. Montanari, and P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 33, pp. E7665–E7671, 2018.
19. J. Sirignano and K. Spiliopoulos, Mean field analysis of neural networks: a law of large numbers, *SIAM J. Appl. Math.*, vol. 80, no. 2, pp. 725–752, 2020.
20. J. Sirignano and K. Spiliopoulos, Mean field analysis of neural networks: a central limit theorem, *Stochastic Process. Appl.*, vol. 130, no. 3, pp. 1820–1852, 2020.
21. F. Baccelli and T. Taillefumier, Replica-mean-field limits for intensity-based neural networks, *SIAM J. Appl. Dyn. Syst.*, vol. 18, no. 4, pp. 1756–1797, 2019.
22. T. Trimborn, S. Gerster, and G. Visconti, Spectral methods to study the robustness of residual neural networks with infinite layers, *Foundations of Data Science*, vol. 2, no. 3, pp. 257–278, 2020.
23. E. Cristiani, B. Piccoli, and A. Tosin, *Multiscale Modeling of Pedestrian Dynamics*. Springer, Cham, 2014.
24. K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.
25. E. Haber, F. Lucka, and L. Ruthotto, Never look back - A modified EnKF method and its application to the training of neural networks without back propagation. Preprint arXiv:1805.08034, 2018.
26. N. B. Kovachki and A. M. Stuart, Ensemble Kalman inversion: a derivative-free

technique for machine learning tasks, *Inverse Probl.*, vol. 35, no. 9, p. 095005, 2019.

27. K. Watanabe and S. G. Tzafestas, Learning algorithms for neural networks with the Kalman filters, *J. Intell. Robot. Syst.*, vol. 3, no. 4, pp. 305–319, 1990.
28. A. Yegenoglu, S. Diaz, K. Krajsek, and M. Herty, Ensemble Kalman filter optimizing deep neural networks, in *Conference on Machine Learning, Optimization and Data Science*, vol. 12514, 2020.
29. K. Janocha and W. M. Czarnecki, On loss functions for deep neural networks in classification, *Schedae Informaticae*, vol. 2016, no. Volume 25, 2017.
30. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, Neural ordinary differential equations, in *Advances in neural information processing systems*, pp. 6571–6583, 2018.
31. H. Lin and S. Jegelka, Resnet with one-neuron hidden layers is a universal approximator, p. 6172â€“6181, Red Hook, NY, USA: Curran Associates Inc., 2018.
32. Y. Lu and J. Lu, A universal approximation theorem of deep neural networks for expressing probability distributions, in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 3094–3105, Curran Associates, Inc., 2020.
33. P. Kidger and T. Lyons, Universal approximation with deep narrow networks, in *Conference on Learning Theory*, 2020.
34. C. Gebhardt, T. Trimborn, F. Weber, A. Bezold, C. Broeckmann, and M. Herty, Simplified ResNet approach for data driven prediction of microstructure-fatigue relationship, *Mechanics of Materials*, vol. 151, p. 103625, 2020.
35. K. Bobzin, W. Wietheger, H. Heinemann, S. Dokhanchi, M. Rom, and G. Visconti, Prediction of particle properties in plasma spraying based on machine learning, *Journal of Thermal Spray Technology*, 2021.
36. L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH ZÃ¼rich, BirkhÃ¤user, 2. ed ed., 2008.
37. C. Villani, *Optimal Transport: Old and New*. Springer-Verlag, 2009.
38. F. Golse, On the dynamics of large particle systems in the mean field limit, in *Macroscopic and large scale phenomena: coarse graining, mean field limits and ergodicity*, pp. 1–144, Springer, 2016.
39. J. M. Coron, *Control and nonlinearity*. American Mathematical Society, 2007.
40. E. Zuazua, Controllability and observability of partial differential equations: Some results and open problems, vol. 3 of *Handbook of Differential Equations: Evolutionary Equations*, pp. 527–621, North-Holland, 2007.
41. N. Fournier and A. Guillin, On the rate of convergence in wasserstein distance of the empirical measure, *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 707–738, 2015.
42. E. Boissard, Simple bounds for convergence of empirical and occupation measures in 1-wasserstain distance, *Electronic Journal of Probability*, vol. 16, pp. 2296–2333, 2011.

43. J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer New York, 2010.
44. I. Cravero, G. Puppo, M. Semplice, and G. Visconti, CWENO: uniformly accurate reconstructions for balance laws, *Math. Comp.*, vol. 87, no. 312, pp. 1689–1719, 2018.
45. G.-S. Jiang and C.-W. Shu, Efficient implementation of weighted ENO schemes, *J. Comput. Phys.*, vol. 126, pp. 202–228, 1996.