symmetry

MDPI

Perspective

From Optimal Control to Mean Field Optimal Transport via Stochastic Neural Networks

Luca Di Persio [1] **and Matteo Garbelli** [1,2,*]

[1] Department of Computer Science, University of Verona, Strada le Grazie 15, 37134 Verona, Italy; luca.dipersio@univr.it
[2] Department of Mathematics, University of Trento, Via Sommarive 14, Povo, 38123 Trento, Italy
* Correspondence: matteo.garbelli@unitn.it

Abstract: In this paper, we derive a unified perspective for Optimal Transport (OT) and Mean Field Control (MFC) theories to analyse the learning process for Neural Network algorithms in a high-dimensional framework. We consider a Mean Field Neural Network in the context of MFC theory referring to the mean field formulation of OT theory that may allow the development of efficient algorithms in a high-dimensional framework while providing a powerful tool in the context of explainable Artificial Intelligence.

Keywords: neural network; machine learning; optimal transport; mean field control; mean field optimal transport

Citation: Di Persio, L.; Garbelli, M. From Optimal Control to Mean Field Optimal Transport via Stochastic Neural Networks. *Symmetry* **2023**, *15*, 1724. https://doi.org/10.3390/sym15091724

Academic Editor: Alexander Zaslavski, Savin Treanta, Octav Olteanu and Sergei D. Odintsov

Received: 2 May 2023
Revised: 28 August 2023
Accepted: 31 August 2023
Published: 8 September 2023

1. Introduction

In recent years, parametric Machine Learning (ML) applications have shown brilliant performance in capturing relevant symmetries and hidden patterns characterizing a specific knowledge base. Specifically, Neural Networks (NNs), i.e., systems of interconnected artificial neurons, constitute a fundamental tool to capture complex patterns and to make accurate predictions for various applications, ranging from computer vision and natural language processing to robotics and reinforcement learning. Their growing popularity has prompted an increasing demand for a deep mathematical description of the underlying training procedures, specifically in high dimensions to tackle the curse of dimensionality.

For this latter research challenge, we consider a novel class of NNs, termed *Mean Field Neural Networks* (MFNNs), which are defined as the limiting object of a *population of NNs* when its number of components tends to infinity. Our aim concerns deriving a unified perspective for this class of models based on existing symmetries between Mean Field Control (MFC) theory and the Optimal Transport (OT) method. Our approach is based on an *infinite dimensional lifting* which allows new insights to be gained into relationships between data in the corresponding finite-dimensional scenario.

We start the analysis by looking at the continuous idealization of a specific class of NNs, namely Residual NNs, also named ResNets, whose training process in a supervised learning scenario is stated as a Mean Field Optimal Control Problem (MFOCP). We consider a deterministic dynamic that evolves in terms of an ordinary differential equation (ODE). Moreover, the training problem of a ResNet is shown to be equivalent to an MFOCP of Bolza type, see [1,2] for further details.

The next passage in our analysis concerns introducing a noisy component into the dynamics of the ODE, moving to a Stochastic Differential Equation (SDE) that allows us to consider the inherent uncertainty connected to the variations in the real-world data, simultaneously allowing for the integration of stochastic aspects into the learning process. Although this second model does not include any mean field terms, it allows the development of a class of algorithms known as Stochastic NNs (SNNs). In [3], the authors develop a sample-wise backpropagation method for SNNs based on backward SDE that models

the gradient (with respect to the parameters) process of the loss function, representing a feasible tool for quantifying the uncertainty of the learning process. Another possible approach for probabilistic learning is studied in [4], where the authors develop the so-called Stochastic Deep Network (SDN), namely an NN architecture that can use as input data not only single vectors but also *random vectors* to model the probability distribution of given inputs. Following their analysis, the SDN is considered as an architecture based on the composition of maps between probability measures performing inference tasks and solving ML problems over the space of probability measures.

In the last passage, we merge the stochastic aspect with the mean field one by considering the so-called Mean Field Optimal Transport (MFOT) formulation, recently introduced in [5]. We describe the MFC tools relevant to formalize the training process; hence, we formulate the training problem as MFOT in an infinite-dimensional setting. Considering the collective interactions and distributions of the network's parameters may facilitate the analysis of the network behavior on a macroscopic level, hence improving the interpretability, scalability, and robustness of NNs models, while adding knowledge by highlighting the hidden symmetries and relations between data.

We highlight that the symmetry between mean field models and ML algorithms is also studied in [6], where the authors establish a mathematical relationship between the MFG framework and normalizing flows, a popular method for generative models composed of a sequence of invertible mappings. Similarly, in [7], the authors analyze Generative Adversarial Networks (GANs) from the perspectives of MFGs, providing a theoretical connection between GANs, OT, and MFG and numerical experiments.

This paper is organized as follows: In Section 2, we introduce the mathematical formalism of the supervised learning paradigm while providing the description of the continuous idealization of a Residual NN stated as an MFOCP; in Section 3, we introduce a noisy component into the network dynamic, thus focusing on Stochastic NNs formalized as stochastic optimal control problems; in Section 4, we review the MFG setting in a cooperative scenario defined in terms of MFC theory. Then, we consider recently developed Mean Field Optimal Transport methods that allow MFC problems to be rephrased into OT ones. We also illustrate related approximation schemes and possible connection to an abstract class of NNs that respect the MFOT structure. We conclude by reviewing some methods to *learn*, i.e., approximate, mean field functions that depend on probability distribution, obtained as the limiting object of empirical measures.

2. Residual Neural Networks as a Mean Field Optimal Control Problem

In this section, we present the workflow to treat a feed-forward NN, specifically a Residual NN, as a dynamical system based on the work in [8]. The main reference for this part is the well-known paper in [2], where the authors introduce a continuous idealization of Deep Learning (DL) to study the Supervised Learning (SL) procedure; this is stated as an optimal control problem by considering the associated population risk minimization problem.

2.1. The Supervised Learning Paradigm

Following [9,10], the SL problem aims at estimating the function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$, commonly known as the Oracle. The space $\mathcal{X}$ can be identified with a subset of $\mathbb{R}^d$ related to input arrays (such as images, string texts, or time series), while $\mathcal{Y}$ is the corresponding target set. Here, for simplicity, we consider $\mathcal{X}$ and $\mathcal{Y}$ Euclidean spaces with different dimensions. Thus, training begins with a set of N input–target pairs $\{x_0^i, y_T^i\}_{i=1}^N$ where:

- $x_0^i \in \mathbb{R}^d$ denotes the inputs of the NN;
- $x_T^i = \mathcal{F}(x_0^i) \in \mathbb{R}^d$ denotes the outputs of the NN;
- $y_T^i \in \mathbb{R}^l$ denotes the corresponding targets.

We assume the same dimension of the Euclidean space for NN inputs and outputs, allowing us to explicitly write a dynamic in terms of a difference equation. Hence, for a ResNet (see [11] for more details) with T layers, the feed-forward propagation is given by

$$x_{t+1} = x_t + f(x_t, \theta_t) \qquad t = 0, \dots, T-1 \tag{1}$$

with $f : \mathbb{R}^d \to \mathbb{R}^d$ being a parameterized function and θ_t being the trainable parameters, e.g., bias, weights of the t-th layer that belong to a measurable set $\mathcal{U}$ with values in a subset of the Euclidean space $\mathbb{R}^m$.

Remark 1. *Following [12], we report an example of a domain for parameters of NN with ReLU activation functions. We define the following parameter domain*

$$\Theta = \left\{ (a, w, b) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} : a^2 < ||w|| + b^2 \right\}$$

with activation functions $\phi : \Theta \to \mathbb{R}$ *defined as*

$$\phi(\theta; x) = a\sigma(w^T x + b), \quad \theta = (a, w, b), \quad \sigma(z) = z^+ = max\{z, 0\},$$

2.2. Empirical Risk Minimization

We aim at minimizing, over the set of measurable parameters Θ, a *terminal loss function* $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ plus a regularization term, L, to derive a Supervised Learning problem as an Empirical Risk Minimization (ERP) problem, namely

$$\min_{\theta \in \mathcal{U}} \left[\frac{1}{K} \sum_{i=1}^{N} \Phi(x_T^i, y^i) + \sum_{t=0}^{T-1} L(\theta_t) \right] \tag{2}$$

over N training data samples indexed by i. We write $\theta = [\theta_0, \dots \theta_{T-1}]$ to identify the set of all parameters of the network.

If we consider no regularization of the parameters, i.e., $L = 0$, and a quadratic loss function in terms of Φ, then Equation (2) reads

$$J^{ERM}(\theta) = \min_{\theta \in \mathcal{U}} \left[\frac{1}{K} \sum_{i=1}^{N} ||x_T^i - y^i||^2 \right] = \min_{\theta \in \mathcal{U}} \left[\frac{1}{K} \sum_{i=1}^{N} ||f(x^i, \theta) - y^i||^2 \right] \tag{3}$$

being $x^i = [x_0, \dots, x_{T_1}]$ the discrete state process defined in Equation (1).

Optimizing J^{ERM} by computing its gradient is computationally expensive, especially if the amount of data K is very large.

To handle the curse of dimensionality, it is usually common to initialize parameters from a θ^0 from a probability distribution, to then optimize their choice inductively according to a Stochastic Gradient Descent scheme

$$\theta^{k+1} = \theta^k - \eta_t \frac{1}{K} \sum_{i=1}^{K} ||f(x^i, \theta) - y^i|| \nabla_\theta f(x^i, \theta) \tag{4}$$

with learning rate η_t over K optimization steps.

For the sake of completeness, before going to the limit (we pass from a discrete set of training data to the corresponding distribution), we point out in the following remark that it is also possible to associate a measure corresponding to the empirical distribution of the parameters when the number of neurons goes to infinity.

Remark 2. *A different approach, as illustrated, e.g., by Sirignano and Spilipupouls in [13], consists of associating to each layer the corresponding empirical measure and building a measure to describe the whole network, hence working with the empirical measure of controls, rather than states, as presented in Section 4. Following the perspective of mean field term in controls, the SGD*

Equation (4) can be formalized as a minimization method over the set of probability distributions. Moreover, the training of the NN is based on the correspondence between the empirical measure of neurons μ_N and the function f_N that is approximated by the NN. Specifically, it has been proved that training via gradient descent of an over-parametrised one-hidden-layer NN with infinite width is equivalent to gradient flow in Wasserstein space [2,9,14,15]. Conversely, in the small learning rate regime, the training is equivalent to an SDE, see, e.g., [16].

From here on, we deal with empirical distribution and measures associated to the training data.

2.3. Population Risk Minimization as Mean Field Optimal Control Problem

In what follows, we move from the discrete setting to the corresponding continuous idealization by:

- Going from layer index T to continuous parameter t;
- Passing from a discrete set of inputs/output to a distribution μ that represents the joint distribution in $\mathcal{W}_2(\mathbb{R}^d \times \mathbb{R}^l)$, modeling the input label distribution;
- Passing from empirical risk minimization to population risk (i.e., minimization over expectation $\mathbb{E}$).

In particular, we pass to the limit in the number of data samples (number of input-target pairs), also assuming a continuous dynamic in place of layer discretization. The latter limit allows us to describe the dynamic of the state process x with the following Ordinary Differential Equation (ODE)

$$\dot{x}_t = f(x_t, \theta_t), \qquad t \in [0, T]\,, \tag{5}$$

in place of the finite difference Equation (1). We identify the input–target pairs as sampled from a given distribution μ allowing us to write the SL problem as a Population Risk Minimization (PRM) problem.

In summary, we aim at approximating the Oracle function $\mathcal{F}$ using a provided set of training data sampled by a (known) distribution μ_0 by optimizing weights θ_t to achieve maximal proximity between x_T (output) and y_T (target). Thus, we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we assume inputs x_0 in $\mathbb{R}^d$ to be sampled from a distribution $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$, with corresponding target y_T in $\mathbb{R}^l$ sampled from a distribution $\nu \in \mathcal{P}(\mathbb{R}^l)$, while the joint probability distribution μ, which models the distribution of the input–target pairs, defined by $\mu := \mathcal{P}(x_0, y_T)$, belongs to the Wasserstein space $\mathcal{W}_2(\mathbb{R}^{(d+l)})$ and has μ_0 and ν as its marginals. We recall that given a metric space (X, d), the p-Wasserstein space $\mathcal{W}_p(X)$ is defined as the set of all Borel probability measures on X with finite p-moments.

The marginal distributions are obtained by projecting the joint probability distribution μ over the subspaces of inputs and output, respectively. We identify the first marginal, i.e., the projection over $\mathbb{R}^l$, with the distribution of inputs

$$\mu_0 = \int_{R^l} \mu(x, y) dy\,,$$

while the distribution of targets reads

$$\nu = \int_{R^d} \mu(x, y) dx\,.$$

Moreover, we assume the controls θ_t depend on the whole distribution of input–target pairs capturing the mean field aspect of the training data. We consider a measurable set of admissible controls, i.e., training weights, $\Theta \subseteq \mathbb{R}^m$ and we state a Mean Field Optimal Control Problem (MFOCP) to solve the following PRM problem:

$$\begin{aligned} &\inf_{\theta \in L^{\infty}([0,T],\Theta)} J^{PRM}(\theta) := \mathbb{E}_{\mu}\left[\Phi(x_T, y_T) + \int_0^T L(x_t, \theta_t)dt\right] \\ &\dot{x}_t = f(x_t, \theta_t) \quad 0 \le t \le T \quad x_0 \sim \mu_0, \quad (x_0, y_T) \sim \mu \end{aligned} \tag{6}$$

We briefly report basic assumptions allowing us to have a solution for (6):

- $f : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$, $L : \mathbb{R}^d \times \Theta \to \mathbb{R}$, $\Phi : \mathbb{R}^d \times \mathbb{R}^l \to \mathbb{R}$ are bounded;
- f, L, Φ are Lipschitz-continuous with respect to x, with the Lipschitz constants of f and L being independent of parameters θ;
- μ has finite support in $\mathcal{W}_2(\mathbb{R}^{(d+l)})$.

Problem (6) can be approached through two different methods: the first one is based on the Hamilton–Jacobi–Bellman (HJB) equation in the Wasserstein space, while the second one is based on a Mean Field Pontryagin Principle. We refer to [17,18] for viscosity solutions to the HJB equation in the Wasserstein space of probability measures, and to [19] for solving the constrained optimal control problems via the Pontryagin Maximum Principle.

For the sake of completeness, let us also cite [20], where the authors introduce a BSDE technique to solve the related Stochastic Maximum Principle, allowing us to consider the uncertainty associated with NN. The authors employ a Stochastic Differential Equation (SDE) in place of the ODE appearing in (6) to continuously approximate a Stochastic Neural Network (SNN). We deepen this approach in the next paragraph.

3. Stochastic Neural Network as a Stochastic Optimal Control Problem

In this paragraph, we generalize the previous setting considering a noisy dynamic, namely adding a stochastic integral to the deterministic setting described by the ODE in Problem (6). The reference model corresponds to Stochastic NN whose discrete state process is described by the following equation

$$X_{n+1} = X_n + hF(X_n, \theta_n) + \sqrt{h}\sigma_n\omega_n\,, \qquad n = 0, 1, \ldots, N-1 \tag{7}$$

with $\{\omega_n\}$ being a sequence of i.i.d. standard Gaussian random variables. We refer to [4] for a theoretical and computational analysis of the SNN.

Equation (7) can be generalized in a continuous setting. To this end, we consider a complete filtered probability space $\left(\Omega, \mathcal{F}, \mathbb{F}^W, \mathbb{P}\right)$, and we introduce the following SDE

$$X_t = X_0 + \int_0^T F(X_s, \theta_s) + \int_0^T \sigma_s dW_s\,, \tag{8}$$

with standard Brownian motion $W := (W_t)_{0 \le t \le T}$ and diffusion term σ. Analogously to ResNets, the index $T > 0$ represents a continuous parameter modeling the width of the layer, with X_T being the output of the network.

Here, we report the theory developed in [3] to study Equation (8) in the framework of the SOC problem by introducing the control process $u = [\theta, \sigma]$. Thus, we also consider the diffusion σ as a trainable parameter of the model. We start by translating the SDE (8) into the following controlled process, written in differential form

$$dX_t = f(X_t, u_t)dt + g(u_t)dW_t, \qquad 0 \le t \le T, \tag{9}$$

where $f(X_t, u_t) = F(X_t, \theta_t)$ and $g(u_t) = \sigma_t$. As in classical control theory applied to ML, the aim is to select the control u that minimizes the discrepancy between the SNN output and the data. Accordingly, we define the cost function for our stochastic optimal control problem as

$$J(u) := \mathbb{E}[\Phi(X_T, \Lambda)]\,, \tag{10}$$

with Λ being a random variable that corresponds to the target of a given input, i.e., X_0. Then, the optimal control $u^\star$ is the one that solves

$$J(u^\star) = \inf_{u \in \mathcal{U}[O,T]} J(u)$$

above the class of measurable control $\mathcal{U}$.

At this point, we are able to write the optimization problem that represents the analogue of Equation (6) with stochastic evolution (where the diffusion is also considered as a model parameter) but without reference to the mean field aspect of the learning procedure.

$$\begin{aligned} &\inf_{u \in L^\infty([0,T],\mathcal{U})} J(u) := \mathbb{E}[\Phi(X_T, \Lambda)] \\ &dX_t = f(X_t, u_t)dt + g(u_t)dW_t, \qquad 0 \le t \le T \end{aligned} \tag{11}$$

Following [3], we address the Stochastic Maximum Principle approach to solve the stochastic optimal control problem stated in (11). Firstly, the functional J is differentiated with respect to the control with a derivative in Gateaux sense over $[0, T]$

$$J'_u(t, u_t) = \mathbb{E}\left[f^i_u(X_t, u_t)^T Y_t + g'_u(u_t)^T Z_t \right]. \tag{12}$$

Then, via the martingale representation of Y_t, the following backward SDE is introduced

$$dY_t = f^i_x(X_t, u^\star_t)^T Y_t + Z_t dW_t\,, \quad Y_T = \Phi'_x(X_T, \Lambda) \tag{13}$$

to model the back-propagation of the forward state process equation defined in (9) associated with the optimal control $u^\star$.

Finally, the problem is solved via the gradient descent method with step size η_k

$$u^{k+1}_t = u^k_t - \eta_k J'_u(t, u^k_t)\,, \quad k = 0, 1, 2, \dots, \quad 0 \le t \le T. \tag{14}$$

Also in [3], the authors provide a numerical scheme whose main benefit is to derive an estimate of the uncertainty connected to the output of this stochastic class of NNs.

We remark that for Equation (14) it is not possible to write the chain rule as previously performed for Equation (4) due to the presence of the stochastic integral term that, differently from classical ML theory, makes the back-propagation itself a stochastic process, see Equation (13). However, modern programming libraries (e.g., TensorFlow or PyTorch) may perform the computation (14) automatically, reducing the computational cost, hence allowing us to go towards a mean field formulation (in terms of multiple interacting agents) of previous problems.

4. Mean Field Neural Network as a Mean Field Optimal Transport

In this section, we focus on the connection between SOC and OT, highlighting potential symmetries specifically for a class of infinite-dimensional stochastic games.

4.1. Optimal Transport

As seen in Section 3, SOC deals with finding the optimal control policy for a dynamic system in the presence of uncertainty. Conversely, OT theory focuses on finding the optimal map to transport from one distribution to another. More precisely, given two marginal distributions $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, the classical OT problem in the Kantorovich formulation reads

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c(x, y)\pi(dx, dy) \tag{15}$$

where c is a cost function and $\Pi(\mu, \nu)$ corresponds to the set of couplings between μ and ν.

We focus on the setting where μ and ν are distributions computed on $\mathbb{R}^d$, i.e., $\mu \sim (X_1, \ldots, X_d)$ and $\nu \sim (Y_1, \ldots, Y_d)$. The Monge formulation reads

$$\inf_{T:T\#\mu=\nu} \int c(x, T(x))\mu(dx) \tag{16}$$

where the infimum is computed over all measurable maps $T : \mathbb{R}^d \to \mathbb{R}^d$ with the pushforward constraint $T\#\mu = \nu$.

The possibility to link a SOC problem, hence the related mathematical formulation of a specific learning procedure, to the corresponding OT formulation relies on lifting the SOC problem in a proper Wasserstein space. For example, considering the SOC problem introduced in (11), the stochastic process X_t described by Equation (9) can be viewed as a vehicle of mass transportation under an initial measure μ_0.

We mention that there are also specific scenarios where the dynamics of the stochastic control problem can be interpreted as a mass transportation problem, provided that certain assumptions of functionals and cost are guaranteed. For example, in [21,22] and similarly in [23], the authors focus on extending an OT problem into the corresponding SOC formulation for a cost, which depends on the drift and the diffusion coefficients of a continuous semimartingale and the minimization is run among all continuous semimartingales with given initial and terminal distributions.

For example, in [22], the authors consider a special form for the cost function, namely $c(x, y) = L(y - x)$ with $L(u) : \mathbb{R}^d \to [0, +\infty]$ convex in u proving its equivalence to a proper SOC problem based on the so-called graph property. Indeed, we can define an image measure as $\pi_g : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ mapping x into $(x, g(x))$. Thus, for any measurable map $g : \mathbb{R}^d \to \mathbb{R}^d$, the following equality between the two formulations holds:

$$\int_{\mathbb{R}^d} L(g(x) - x)\mu(dx) = \int_{\mathbb{R}^d \times \mathbb{R}^d} L(y - x)\pi_g(dxdy) \tag{17}$$

Thus, μ_g models a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν.

For the problem stated in (17), we know from [24] that an optimal measure $\pi^\star$ always exists. Moreover, if the optimal measure $\pi^\star$ is supported by the graph of a measurable map, we say that the graph property holds; that is, if for any $\pi^\star$ optimal for (15), there exists a set Γ satisfying $\pi^\star(\Gamma) = 1$ with $\Gamma = (x, \gamma(x))$ for some measurable mapping γ that resembles the NN parameters introduced in Section 2 and analogously $\gamma(x)$ represents the corresponding output according to Equation (1).

4.2. Mean Field Games

In the context of Mean Field Games (MFGs), i.e., stochastic games where a large number of agents interact and influence each other, the link between SOC and OT is particularly explicable, specifically according to the variational formulation of MFGs, which is directly linked to the dynamic formulation of OT by Benamou and Brenier, see, e.g., ref. [25] for an in-depth analysis.

In Section 2, we focus on deterministic evolution by means of Equation (5) with the mean field interactions captured by the loss function as an expectation given a known joint measure μ between the input and target in the corresponding Mean Field Optimal Problem (22). On the other hand, in Section 3, we introduce the stochastic process in Equation (8) and state the learning problem as an SOC as shown in Equation (10) without focusing on the interaction during the evolution but looking at just a single trajectory. Finally, the further natural step relies on extending the previous equation to a McKean–Vlasov setting where the dynamic of a random variable X depends on the other N random variables by the mean of the distribution in order to merge the two scenarios presented in Sections 2 and 3 while extending the problem stated in (10) by allowing the presence of a mean field term.

Indeed, instead of considering a single evolution as in Equation (9), we introduce the following McKean–Vlasov SDE for N particles/agents

$$X_t^i = X_0^i + \int_0^T b\Big(X_s^i, m_{X_s}^N, \theta_s\Big) + \int_0^T \sigma dW_s\,, i = 1, \ldots, N \tag{18}$$

with X_0^i being the initial states. We assume a measurable drift $b : [0, T] \times \mathcal{W}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$, a constant diffusion σ, and we define the empirical distribution $m_{X_s}^N$ as

$$m_{X_s}^N = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_s^i}\,. \tag{19}$$

The main idea would be to model multiple SNNs and generalize the dynamic defined in (9); including the dependence on a mean field term in the drift allows us to model the shared connections between the neurons of different SNNs.

At the limit $N \to \infty$, the *population of SNNs* corresponds to the evolution of a representative SNN, while the empirical measure m^N tends to the probability measure m belonging to the Wasserstein space $W_2(\mathbb{R}^d)$, i.e., the space of probability measures on $\mathbb{R}^d$ with a finite second-order moment that captures a measure of interactions among SSNs.

More precisely, we introduce the following settings, which we need to define the solution of an MFG.

- A finite time horizon $T > 0$;
- $\mathcal{Q} \subseteq \mathbb{R}^d$ is the state space;
- $\mathcal{W}_2(\mathcal{Q})$ is the space of probability measure over $\mathcal{Q}$;
- $(x, m, \alpha) \in \mathcal{Q} \times \mathcal{W}_2(\mathcal{Q}) \times \mathbb{R}^k$ describes the agent state, the mean field term, and the agent control;
- $f : \mathcal{Q} \times \mathcal{W}_2(\mathcal{Q}) \times \mathbb{R}^k \to \mathbb{R}$, $(x, m, \alpha) \mapsto f(x, m, \alpha)$ and $g : \mathcal{Q} \times \mathcal{W}_2(\mathcal{Q}) \to \mathbb{R}$, $(x, m) \mapsto g(x, m)$ provide the running and the terminal cost, respectively;
- $b : \mathcal{Q} \times \mathcal{W}_2(\mathcal{Q}) \times \mathbb{R}^k \to \mathbb{R}^d$ represents the drift function;
- $\sigma > 0$ is the volatility of the state.

Definition 1 (MFG equilibrium)**.** *We consider an MFG problem with a given initial distribution $m_0 \in \mathcal{W}_2(\mathcal{Q})$. A Nash equilibrium is a flow of probability measures $\hat{m} = (\hat{m}(t, \cdot))_{0 \le t \le T}$ in $\mathcal{W}_2(\mathcal{Q})$ plus a feedback control $\hat{\alpha} : [0, T] \times \mathcal{Q} \to \mathbb{R}^k$ satisfying the following two conditions:*

1. *$\hat{\alpha}$ minimizes J_m^{MFG}over α:*

$$\mathbb{E}\left[\int_0^T f(X_t^{m,\alpha}, m(t, \cdot), \alpha(t, X_t^{m,\alpha}))dt + g(X_T^{m,\alpha}, m(T, \cdot))\right]$$

where $(X_t^{m,\alpha})$ solves the SDE

$$dX_t^{m,\alpha} = b(X_t^{m,\alpha}, m(t, \cdot), \alpha(t, X_t^{m,\alpha}))dt + \sigma dW_t$$

with W being a d-dimensional Brownian motion and $X_0^{m,\alpha}$ having distribution m_0;

2. *For all $t \in [0, T]$, $\hat{m}$ is the probability distribution of $X_t^{\hat{m},\hat{\alpha}}$.*

4.3. Mean Field Control

Differently from MFG, where players are modeled as competitors, Mean Field Control (MFC) models a framework that considers a large population of agents aiming to cooperate and optimize individual objectives. In the MFC setting, each agent cost depends on a mean field term representing the average behavior of all agents. Accordingly, the solution of an MFC is defined in the following way:

Definition 2 (MFC optimum). *Given $m_0 \in \mathcal{W}_2(\mathcal{Q})$, a feedback control $\alpha^* : [0,T] \times \mathcal{Q} \to \mathbb{R}^k$ is an optimal control for the MFC problem if it minimizes over α J^{MFC} defined by*

$$\mathbb{E}\left[\int_0^T f(X_t^\alpha, m^\alpha(t,\cdot), \alpha(t, X_t^\alpha))dt + g(X_T^\alpha, m^\alpha(T,\cdot))\right] \tag{20}$$

where $m^\alpha(t,\cdot)$ is the probability distribution of the law of X_t^α, under the constraint that the process $(X_t^\alpha)_{t\in[0,T]}$ solves the following McKean–Vlasov-type SDE:

$$dX_t^\alpha = b(X_t^\alpha, m^\alpha(t,\cdot), \alpha(t, X_t^\alpha))dt + \sigma dW_t \tag{21}$$

with X_0^α having distribution m_0.

We refer to [26] for an extensive treatment of McKean–Vlasov control problems (20).

By considering the joint optimization problem of the entire population, MFC enables the analysis of large-scale systems with cooperative agents and provides insights into the emergence of collective behavior. One possibility relies on stating the dynamic in Equation (6) in terms of probability measures. For example, we can consider a continuity equation such as the Fokker–Planck equation to consider the evolution of the density function. Along this setting, we cite the measure theoretical approach for NeurODE developed in [1], where the authors introduced a forward continuity equation in the space of measures with a constrained dynamic in the form of an ODE. Conversely, within the cooperative setting, we can also rely on a novel approach, named Mean Field Optimal Transport, introduced in [5], which we explore in the next paragraph.

4.4. Mean Field Optimal Transport

Mean Field Optimal Transport deals with a framework where all the agents cooperate (such as in MFC) in order to minimize a total cost without terminal cost but with an additional constraint since also the final distribution is prescribed. We notice that the setting with fixed initial and terminal distributions resembles the one introduced in the Population Risk Minimization problem described in Section 2. We follow the numerical scheme introduced in Section 3.1 in [5] to approximate feedback controls, that is, we introduce the following model.

Definition 3 (Mean Field Optimal Transport). *Let $\mathbb{R}^d$, describe the state space and denote by $\mathcal{W}_2(\mathbb{R}^d)$ the set of square-integrable probability measures on $\mathbb{R}^d$. Let $f : \mathbb{R}^d \times \mathcal{W}_2(\mathbb{R}^d) \times \mathbb{R}^k \to \mathbb{R}$ be the running cost, $g : \mathbb{R}^d \times \mathcal{W}_2(\mathbb{R}^d) \to \mathbb{R}$ be the terminal cost, $b : \mathbb{R}^d \times \mathcal{W}_2(\mathbb{R}^d) \times \mathbb{R}^k \to \mathbb{R}^d$ the drift function, and $\sigma \in \mathbb{R}$ the non-negative diffusion. Given two distributions, ρ_0 and $\rho_T \in \mathcal{W}_2(\mathbb{R}^d)$, the aim of MFOT is to compute the optimal feedback control $v : [0,T] \times \mathbb{R}^d \to \mathbb{R}^m$ minimizing*

$$J^{MFOT} : v \mapsto \mathbb{E}\left[\int_0^T f(X_t^v, \mu^v(t), v(t, X_t^v))dt\right] \tag{22}$$

where $\mu^v(t)$ is the distribution of process X_t^v, whose dynamics is given by

$$\begin{cases} X_0^v \sim \rho_0 \qquad X_T^v \sim \rho_T \\ dX_t^v = b(X_t^v, \mu^v(t), v(t, X_t^v))dt + \sigma dW_t\,, \qquad t \in [0,T] \end{cases} \tag{23}$$

with ρ_0 and ρ_T the prescribed initial and terminal distributions.

This type of problem incorporates mean field interactions into the drift and the running cost. Furthermore, it encompasses classical OT as a special case by considering $b(x,\mu,a) = a$, $f(x,\mu,a) = \frac{1}{2}a^T a$, and $\sigma = 0$.

The integration of MFC and OT allows us to both tackle the weight optimization problem in NN and to model the flow of information or mass between layers of neurons,

while the optimal weights may be computed as the minimizers of the functional with respect to the controls v

$$v^\star = \min_{v \in \mathcal{U}} J^{MFOT}(v) \tag{24}$$

along all the trajectories X^v, where $\mathcal{U}$ is the set of admissible controls.

Thus, we look at the MFNN as a collection of identical, interchangeable, indistinguishable NNs where the dynamic of the representative agents is a generalization of an SNN (7), allowing a dependence on the term $\mu^v(t)$ modeling the mean field interactions. By considering the MFNN dynamic as a population of interconnected NNs, we can employ mean field control to analyze the collective behavior and interactions of networks, accounting for their impact on the overall network performance.

To summarize, we are looking at this novel class of NN, i.e., MFNN, as the asymptotic configuration of NNs in a cooperative setting.

We remark that the representative agent does not know the mean field interaction terms, since it depends on the whole population, but an approximated version can be recursively learned. For example, in [5], the authors present a different numerical scheme to solve MFOT:

1. Optimal control via direct approximation of controls v;
2. Deep Galerkin method for solving forward–backward systems of PDEs;
3. Augmented Lagrangian method with Deep Learning exploiting the variational formulation of MFOT and the primal/dual approach.

We briefly review the direct method (1) to approximate feedback-type controls via an optimal control formulation. The controls are assumed to be of feedback form and can be approximated by

$$g(x, \mu) = G(\mathcal{W}_2(\mu, \rho_T)), \qquad \mu \in \mathcal{W}_2(\mathbb{R}^d), \tag{25}$$

where $G : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function. The idea is to use the function in Equation (25) as a penalty for being far from the target distribution ρ_T as the terminal cost to embed the problem into the classical MFG/MFC literature. Intuitively, Equation (25) corresponds to the infinite dimensional analogue of the loss function of the leveraged NN algorithm, where μ is the final distribution that has to be as close as possible in the sense of the Wasserstein metric to the target distribution ρ_T.

In view of obtaining a numerically tractable version of the SDE (23), one may consider a classical discretization Euler–Maruyama scheme, also requiring the set of controls v to be restricted to the ones approximated by NNs v_θ with parameters θ. Moreover, approximating the mean field term m by its finite dimensional counterpart, see Equation (19), allows us to develop a stable numerical algorithm, see Section 3.1 in [5] for further details, particularly with respect to the linked numerical implementation.

4.5. Other Approaches for Learning Mean Field function

For the sake of completeness, we also mention two different methods to deal with the approximation of the mean field function that can be used in parallel with MFOT:

- The first data-driven approach, presented in [27], has been considered to solve a stochastic optimal control problem, where the unknown model parameters were estimated in real time using a *direct filter method*. This method involves transitioning from the Stochastic Maximum Principle to approximate the conditional probability density functions of the parameters given an observation, which is a set of random samples;
- In [28], the authors report a map that by operating over an appropriate classes of neural networks, specifically the *bin-density-based approximation* and *cylindrical approximation*, is able to reconstruct a mapping between the Wasserstein space of probability measures and an infinite dimensional function space on a similar setting to MFG.

5. Conclusions and Further Directions

In the present article, we provided a general overview of methods at the intersection of parametric ML, MFC, and OT. By assuming a dynamical system viewpoint, we considered the deterministic, ODE-based setting of the supervised learning problem, to then incorporate noisy components, allowing for the definition of stochastic NNs, hence introducing the MFOT approach. The latter, derived as the limit in the number of training data, recasts the classical learning process as a Mean Field Optimal Transport one. As a result, we gained a unified perspective on the parameter optimization process, characterizing ML models with a specified learning dynamic, within the framework of OT and MFC, which may allow high-dimensional data sets to be efficiently handled.

We empathise that the major limitation of MFOT (22) concerns the fact that many of its convergence results, such as those related to corresponding forward–backward systems, still need to be verified. Nevertheless, it represents an indubitably fertile and stimulating research ground that should be enhanced since it permits the derivation of techniques that may significantly improve the robustness of algorithms, particularly when dealing with huge sets of training data that are potentially perturbed by random noise components, while also allowing hidden symmetries within data to be highlighted. The latter aspect is particularly interesting when dealing with intrinsically structured problems as, e.g., in the case of NLP tasks, see, e.g., [29,30].

Author Contributions: Conceptualization, M.G.; methodology, M.G.; validation, M.G. and L.D.P.; formal analysis, M.G.; investigation, M.G.; resources, M.G.; writing—original draft preparation, M.G. and L.D.P.; writing—review and editing, M.G. and L.D.P.; supervision, L.D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to kindly thank Beatrice Acciaio for her valuable advice.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DL	Deep Learning
HJB	Hamilton–Jacobi–Bellman
MFC	Mean Field Control
MFG	Mean Field Games
MFOCP	Mean Field Optimal Control Problem
ML	Machine Learning
MFOT	Mean Field Optimal Transport
NN	Neural Network
ODE	Ordinary Differential Equation
OT	Optimal Transport
SDE	Stochastic Differential Equation
SNN	Stochastic Neural Network
SGD	Stochastic Gradient Descent

References

1. Bonnet, B.; Cipriani, C.; Fornasier, M.; Huang, H. A measure theoretical approach to the mean-field maximum principle for training NeurODEs. *Nonlinear Anal.* **2023**, *227*, 113161. [CrossRef]
2. E, W.; Han, J.; Li, Q. A mean-field optimal control formulation of deep learning. *Res. Math. Sci.* **2019**, *6*, 10. [CrossRef]

3. Archibald, R.; Bao, F.; Cao, Y.; Zhang, H. A backward SDE method for uncertainty quantification in deep learning. *Discret. Contin. Dyn. Syst.* **2022**, *15*, 2807–2835. [CrossRef]
4. de Bie, G.; Peyré, G.; Cuturi, M. Stochastic Deep Networks. In Proceedings of the 36th International Conference on Machine Learning, PMLR 97, Long Beach, CA, USA, 9–15 June 2019.
5. Baudelet, S.; Frénais, B.; Laurière, M.; Machtalay, A.; Zhu, Y. Deep Learning for Mean Field Optimal Transport. *arXiv* **2023**, arXiv:2302.14739.
6. Huang, H.; Yu, J.; Chen, J.; Lai, R. Bridging mean-field games and normalizing flows with trajectory regularization. *J. Comput. Phys.* **2023**, *487*, 112155. [CrossRef]
7. Cao, H.; Guo, X.; Laurière, M. Connecting GANs, MFGs, and OT. *arXiv* **2020**, arXiv:2002.04112.
8. Li, Q.; Lin, T.; Shen, Z. Deep Learning via Dynamical Systems: An Approximation Perspective. *arXiv* **2019**, arXiv:1912.10382v1.
9. Di Persio, L.; Garbelli, M. Deep Learning and Mean-Field Games: A Stochastic Optimal Control Perspective. *Symmetry* **2021**, *13*, 14. [CrossRef]
10. Li, Q.; Chen, L.; Tai, C.; E, W. Maximum principle based algorithms for deep learning. *J. Mach. Learn. Res.* **2017**, *18*, 5998–6026.
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
12. Wojtowytsch, S. On the Convergence of Gradient Descent Training for Two-layer ReLU-networks in the Mean Field Regime. *arXiv* **2020**, arXiv:2005.13530.
13. Sirignano, J.; Spiliopoulos, K. Mean Field Analysis of Deep Neural Networks. *Math. Oper. Res.* **2021**, *47*, 120–152. [CrossRef]
14. Chizat, L.; Colombo, M.; Fernández-Real, X.; Figalli, A. Infinite-width limit of deep linear neural networks. *arXiv* **2022**, arXiv:2211.16980.
15. Fernández-Real, X.; Figalli, A. The Continuous Formulation of Shallow Neural Networks as Wasserstein-Type Gradient Flows. In *Analysis at Large*; Avila, A., Rassias, M.T., Sinai, Y., Eds.; Springer: Cham, Switzerland, 2022.
16. Chizat, L.; Bach, F. On the global convergence of gradient descent for overparameterized models using optimal transport. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018; pp. 3040–3050.
17. Gangbo, W.; Mayorga, S.; Swiech, A. Finite Dimensional Approximations of Hamilton-Jacobi Bellman Equations in Spaces of Probability Measures. *SIAM J. Math. Anal.* **2021**, *53*, 1320–1356. [CrossRef]
18. Jimenez, C.; Marigonda, A.; Quincampoix, M. Dynamical systems and Hamilton-Jacobi-Bellman equations on the Wasserstein space and their L2 representations. *J. Math. Anal. (SIMA)* 2022, *preprint*. Available online: https://cvgmt.sns.it/media/doc/paper/5584/AMCJMQ_HJB_2022-03-30.pdf (accessed on 17 February 2023).
19. Benoît, B. A Pontryagin Maximum Principle in Wasserstein spaces for constrained optimal control problems. *ESAIM Control. Optim. Calc. Var.* **2019**, *25*, 52. [CrossRef]
20. Bao, F.; Cao, Y.; Archibald, R.; Zhang, H. Uncertainty quantification for deep learning through stochastic maximum principle. *arXiv* **2021**, arXiv:3489122.
21. Mikami, T. Two End Points Marginal Problem by Stochastic Optimal Transportation. *SIAM J. Control. Optim.* **2015**, *53*, 2449–2461. [CrossRef]
22. Mikami, T.; Thieullen, M. Optimal transportation problem by stochastic optimal control. *SIAM J. Control Optim.* **2008**, *47*, 1127–1139. [CrossRef]
23. Tan, X.; Nizar Touzi, N. Optimal transportation under controlled stochastic dynamics. *Ann. Probab.* **2013**, *41*, 3201–3240. [CrossRef]
24. Villani, C. *Topics in Optimal Transportation*; Grad. Stud. Math. 58; AMS: Providence, RI, USA, 2003.
25. Benamou, J.D.; Carlier, G.; Santambrogio, F. Variational Mean Field Games. In *Active Particles, Volume 1: Advances in Theory, Models, and Applications*; Bellomo, N., Degond, P., Tadmor, E., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 141–171.
26. Carmona, R.; Lauriere, M. Deep Learning for Mean Field Games and Mean Field Control with Applications to Finance. In *Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices*; Capponi, A., Lehalle, C., Eds.; Cambridge University Press: Cambridge, UK, 2023; pp. 369–392. [CrossRef]
27. Archibald, R.; Bao, F.; Yong, J. An Online Method for the Data Driven Stochastic Optimal Control Problem with Unknown Model Parameters. *arXiv* **2022**, arXiv:2208.02241.
28. Pham, H.; Warin, X. Mean-field neural networks: Learning mappings on Wasserstein space. *arXiv* **2022**, arXiv:2210.15179.
29. Mao, K.; Xu, J.; Yao, X.; Qiu, J.; Chi, K.; Dai, G. A text classification model via multi-level semantic features. *Symmetry* **2022**, *14*, 1938. [CrossRef]
30. Yoo, Y.; Heo, T.S.; Park, Y.; Kim, K. A novel hybrid methodology of measuring sentence similarity. *Symmetry* **2021**, *13*, 1442. [CrossRef]