Preprints.org

Article

Not peer-reviewed version

From Optimal Control to Optimal Transport via Stochastic Neural Networks in the Mean Field Setting

Matteo Garbelli * and Luca Di Persio

Posted Date: 20 June 2023

doi: 10.20944/preprints202305.0252.v3

Keywords: Neural Network; Machine Learning; Optimal Transport; Mean Field Control

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

From Optimal Control to Optimal Transport via Stochastic Neural Networks in the Mean Field Setting

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

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

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

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

1. Introduction

In recent years, parametric Machine Learning (ML) applications, such as Neural Networks (NNs), have become central tools for various applications, ranging from computer vision and natural language processing to robotics and reinforcement learning. NNs consist of interconnected artificial neurons able to capture complex patterns while making accurate predictions. As their popularity has grown, also a fine mathematical description of the training procedure is increasingly required.

Along this latter research line, we propose 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, as to derive a unified perspective for Mean Field Control (MFC) and Optimal Transport (OT) theory in an infinite dimension, allowing to get new insights about relationships between data in a concrete 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 is stated as a Mean Field Optimal Control Problem (MFOCP), and with a dynamic that evolves in terms of an ordinary differential equation (ODE). Moreover, the training problem of a ResNet is shown to be equivalent to a MFOCP of Bolza type, see [16] and [7] for further details.

Then we introduce a noise component into the dynamics of the ODE to derive a stochastic differential equation (SDE) that allows us to consider the inherent uncertainties connected to the variations in the real-world data thus allowing for the stochastic aspects of the learning process.

Finally, we consider the Supervised Learning problem analysis within the so-called Mean Field Optimal Transport (MFOT) formulation, recently introduced in [3]. We describe the MFC tools relevant to formalize the training process, hence we formulate the training problem as MFOT in an infinite-dimensional setting. It is worth mentioning that considering the collective interactions and distributions of the network's parameters may facilitate the analysis of the network behaviour on a macroscopic level, hence improving the interpretability, scalability, and robustness of NNs models. Let us underline that the connection between mean field models and ML algorithms is also studied in [20], 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 [9], the authors analyze Generative Adversarial Networks (GANs) from the perspectives of MFGs providing a theoretical connection between GANs, OT and MFG and numerical experiments.

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

The 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 a 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 rephrasing MFC problems 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 [23].

The main reference for this part is the well-known paper [16], where the authors introduce a continuous idealization of Deep Learning (DL) to study the Supervised Learning (SL) procedure which is stated as an optimal control problem by considering the associated population risk minimization problem.

2.1. The Supervised Learning paradigm

Following [15,24], 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;
- $y_T^i \in \mathbb{R}^l$ the corresponding targets.

We assume the same dimension of the Euclidean space for NN inputs and outputs allowing to explicitly write a dynamic in terms of a difference equation. Hence, for a ResNet, see [19] 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 layer t–th that belong to a measurable set $\mathcal{U}$ with values in a subset of the Euclidean space $\mathbb{R}^m$.

Remark 1. *Following [30], 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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

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 Eq. (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 Eq. (1).

Optimizing J^{ERM} by computing its gradient is computationally expensive, especially if the number 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 [28], consists in 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 we see in Section 4. Following this perspective, the SGD equation (4) can be formalized as a minimization method over probability distribution. Moreover, the training of NN is based on the correspondence between the empirical measure of neurons μ_N and the function f_N approximated by NN. Specifically, the training via gradient descent of a limit of an over-parameterized 1-hidden layer Neural Network with infinite width is equivalent to gradient flow in Wasserstein space [12,15–17]. Similarly, the convergence to solutions to SDEs is obtained as universal continuum objects in the small learning rate regime, see, e.g., [11].*

From here on, we work with empirical distribution and measures associated with 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 distribution μ that represents the joint distribution in $\mathcal{W}_2(\mathbb{R}^d \times \mathbb{R}^l)$ modelling the input-label distribution;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

- 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) assuming also a continuous dynamic in place of layers discretization. The latter limit allows us to describe the dynamic of the state process x by 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, that models the distribution of the input-target pairs, is 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:

$$\inf_{\theta \in L^\infty([0,T],\Theta)} J^{PRM}(\theta) := \mathbb{E}_\mu \left[\Phi\left(x_T, y_T\right) + \int_0^T L(x_t, \theta_t) dt \right] \tag{6}$$

$$\dot{x}_t = f(x_t, \theta_t) \qquad 0 \le t \le T \qquad x_0 \sim \mu_0, \,, \quad (x_0, y_T) \sim \mu$$

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, and
- 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 [18] and [21] for viscosity solutions to the HJB equation in the Wasserstein space of probability measures, and to [8] for solving the constrained optimal control problems via Pontryagin Maximum Principle.

For the sake of completeness, let us also cite [6] 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

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 (6). The reference model corresponds to Stochastic NN whose discrete state process is described by the following equation

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

$\{\omega_n\}$ being a sequence of i.i.d. standard Gaussian random variables. We refer to [14] for a theoretical and computational analysis of SNN.

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

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

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

Here, we report the theory developed in [2] to study Eq. (8) in the framework of SOC problem by introducing the control process $u = [\theta, \sigma]$. Thus we consider also 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

$$dX_t = f(X_t, u_t)dt + g(u_t)dW_t, \qquad 0 \le t \le T, (3) \tag{9}$$

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

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

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 $\mathcal{U}$. Always in [2], the authors deepen the Stochastic Maximum Principle approach to solve the latter problem. Firstly, the functional J is differentiated with respect to the control with a derivative in Gateaux sense over $[0, T]$

$$J'_u(t, u_t) = \mathbb{E}\left[f^i_u(X_t, u_t)^T Y_t + g'_u(u_t)^T Z_t\right] \,. \tag{11}$$

Then, by the martingale representation of Y_t, the following Backward SDE is introduced

$$dY_t = f^i_x(X_t, u^\star_t)^T Y_t + Z_t dW_t \,, \quad Y_T = \Phi'_x(X_T, \Lambda) \tag{12}$$

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

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

$$u^{k+1}_t = u^k_t - \eta_k J'_u(t, u^k_t) \,, \quad k = 0, 1, 2, \dots, \quad 0 \le t \le T1, . \tag{13}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

The authors also 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 here it is not possible to write the chain rule for Eq. (13) as previously done for Eq. (4) since the presence of the stochastic integral term that, differently from classical ML theory, makes the back-propagation itself a stochastic process, see Eq. (12). However, modern programming libraries (e.g., TensorFlow or PyTorch) perform the computation (13) automatically reducing the computational cost, hence allowing 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

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 way to transport from one distribution to another.

In what follows, we recall some basic definitions related to the OT problem. Given two marginal distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, the classical OT problem in the Kantorovich formulation reads

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c(x.y)\pi(dx, dy) \tag{14}$$

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. $\mu \sim (X_1, \dots, X_d)$ and $\nu \sim (Y_1, \dots, Y_d)$. The Monge formulation reads

$$\inf_{T:T\#\mu=\nu} \int c(x, T(x))\mu(dx) \tag{15}$$

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 (and the related mathematical formulation of a specific learning procedure) to the corresponding OT formulation, allows for a deeper understanding of the underlying learning dynamics as well as the use of specific optimal transport techniques to analyze and solve the combined problem.

In this section, we focus on the connection between SOC and OT. This equivalence is particularly evident in the context of Mean Field Games (MFG), a class of stochastic control problems where a large number of agents interact and influence each other, and in particular the variational formulation of MFG which is directly linked to the dynamic formulation of OT by Benamou and Brenier, see, e.g., [4] for further details.

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 on functionals and cost are guaranteed. For example in [25] and similarly in [29], the authors focus on an OT problem, whose cost depends on the drift and the diffusion coefficients of continuous semimartingale, and the minimization is run among all continuous semimartingales with given initial and terminal distributions.

4.1. Mean Field Game

Equations (5) and (8) look at the state process X_t as a random variable, while the functional (10) is computed over the realization of this random variable without focusing on the interaction during the evolution. One further 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

Adding the dependence on a mean-field term into the drift allows us to model the shared connections between the neurons.

We introduce the following McKean-Vlasov SDE for N particles/agents

$$X_t^i = X_0^i + \int_0^T b\left(X_s^i, m_{X_s}^N, \theta_s\right) + \int_0^T \sigma dW_s \,, i = 1, \dots, N \tag{16}$$

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

$$m_{X_s}^N = \frac{1}{N} \sum_{j=1}^N \delta_{X_s^i} \,. \tag{17}$$

Going at the limit $N \to \infty$, 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.

More precisely, we introduce the following setting, that we need to define the solution of a MFG.

- a finite time horizon $T > 0$;
- $\mathcal{Q} \subseteq \mathbb{R}^d$ is the state space;
- $\mathcal{W}_2(\mathcal{Q})$ 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 $g : \mathcal{Q} \times \mathcal{W}_2(\mathcal{Q}) \to \mathbb{R}$, $(x, m) \mapsto g(x, m)$ provides the running and, resp., the terminal cost;
- $b : \mathcal{Q} \times \mathcal{W}_2(\mathcal{Q}) \times \mathbb{R}^k \to \mathbb{R}^d$ represents the drift function ;
- $\sigma > 0$ is the volatility of the state.

Definition 1 (MFG equilibrium)**.** *We consider a 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

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

   W being a d-dimensional Brownian motion and $X_0^{m,\alpha}$ has distribution m_0;
2. *for all $t \in [0,T]$, $\hat{m}$ is the probability distribution of $X_t^{\hat{m},\hat{\alpha}}$.*

4.2. Mean Field Control

Differently from MFG where players are modelled as competitors, MFC is a framework that considers a large population of agents aiming to cooperate and optimize their individual objectives while taking into account the collective behaviour of the entire population. In MFC, each agent seeks to minimize its own cost function by considering the impact of the mean field representing the average behaviour of all agents, hence capturing cooperation spirit among agents. Accordingly, the solution of a MFC is defined in the following way:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

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*

$$\mathbb{E}\left[\int_0^T f(X_t^\alpha, m^\alpha(t,\cdot), \alpha(t, X_t^\alpha))dt + g(X_T^\alpha, m^\alpha(T,\cdot))\right] \tag{18}$$

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 SDE of McKean-Vlasov type

$$dX_t^\alpha = b(X_t^\alpha, m^\alpha(t,\cdot), \alpha(t, X_t^\alpha))dt + \sigma dW_t \tag{19}$$

with X_0^α having distribution m_0.

We refer to [10] for an extensive treatment of McKean-Vlasov control problems (18).

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 behaviour. One possibility relies on stating the dynamic in Eq. (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 [7] where the authors introduce 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 [3], that we explore in the next paragraph.

4.3. Mean Field Optimal Transport

The 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 [3] to approximate feedback controls, namely 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*

$$J^{MFOT} : v \mapsto \mathbb{E}\left[\int_0^T f(X_t^v, \mu^v(t), v(t, X_t^v))dt\right] \tag{20}$$

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

$$\begin{cases} X_0^v \sim \rho_0 \qquad X_T^v \sim \rho_T \\ dX_t^v = b(X_t^v, \mu^v(t), v(t, X_t^v))dt + \sigma dW_t\,, \qquad t \in [0,T] \end{cases} \tag{21}$$

with ρ_0 and ρ_T are the prescribed initial and terminal distributions.

This type of problem incorporates mean field interactions in 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 $\sigma = 0$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

The integration of MFC and OT allows 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 controls v

$$v^\star = \min_{v \in \mathcal{U}} J^{MFOT}(v) \tag{22}$$

along all the trajectories X^v,being $\mathcal{U}$ 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 a SNN (7) allowing a dependence on the term $\mu^v(t)$ modelling 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 asymptotically 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 [3] the authors present different numerical scheme to solve MFOT:

1. Optimal control via direct approximation of controls v;
2. Deep Galerkin Method for solving a 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 controls of feedback type by an optimal control formulation. The controls are assumed of feedback form and can be approximated by

$$g(x, \mu) = G(\mathcal{W}_2(\mu, \rho_T)), \qquad \mu \in \mathcal{W}_2(\mathbb{R}^d) \tag{23}$$

where $G : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function. The idea is to use the function in Eq. (23) as a penalty for being far from the target distribution ρ_T as the terminal cost to embed the problem into classical MFG/MFC literature. Intuitively, Eq. (23) corresponds to the infinite dimensional analogous of the loss function of the leveraged NN algorithm, being μ the final distribution that has to be close as possible in the sense of Wasserstein metric to the target distribution ρ_T.

In view of obtaining a numerically tractable version of the SDE (21), one may consider a classical discretization Euler-Maruyama scheme, also requiring the set of controls v to be restricted to the ones approximated by a NNs v_θ with parameters θ. Moreover, approximating the mean field term m by its finite dimensional counterpart, see Eq. (17), allows to develop a stable numerical algorithm, see Section 3.1 in [3] for further details, particularly w.r.t. the linked numerical implementation.

4.4. Other approaches for learning Mean Field function

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

- the first data-driven approach, presented in [1], 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 [26], 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

5. Conclusions and further directions

In the present article, we provided an all-around overlook 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 stochastic NNs definition, hence introducing the MFOT approach. The latter, derived as the limit in the number of training data, recasts the classical learning process in a mean field optimal transport one. As a result, we gained a unified perspective for the parameters optimization process characterizing ML models with a specified learning dynamic, within the framework of OT and MFC, which may concede to efficiently handle high dimensional data sets.

We empathise that the major limitation of MFOT (20) 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 potentially perturbed by random noise components.

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 advices.

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

Abbreviations

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. Archibald, R.; Bao, F.; Yong, J. An Online Method for the Data Driven Stochastic Optimal Control Problem with Unknown Model Parameters, arXiv e-prints, **2022**.
2. Archibald, R.; Bao, F.; Cao, Y.; Zhang, H.; A backward SDE method for uncertainty quantification in deep learning. Discrete and Continuous Dynamical Systems, **2022**, 15(10): 2807-2835. doi: 10.3934/dcdss.2022062

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

3. Baudelet, S.; Frénais, B.; Laurière, M.; Machtalay, A.; Zhu, Y. Deep Learning for Mean Field Optimal Transport. arXiv e-prints **2023**. doi:10.48550/arXiv.2302.14739
4. Benamou, J.D; Carlier, G.; Santambrogio, F; Variational Mean Field Games. Nicola Bellomo, Pierre Degond, Eitan Tadmor. Active Particles, Volume 1, Springer, pp.141-171, **2017**. ⟨hal-01295299⟩
5. Backhoff-Veraguas, J.; Bartl, D.; Beiglblock, M.; Wiesel, J. Estimating processes in adapted Wasserstein distance. Ann. Appl. Probab. 32 (1) 529 - 550, February 2022. https://doi.org/10.1214/21-AAP1687
6. Bao, F.; Cao, Y.; Archibald, R.; Zhang, H. Uncertainty quantification for deep learning through stochastic maximum principle. arXiv: 3489122, **2021**.
7. Bonnet, B.; Cipriani, C.; Fornasier, M.; Huang, H. A measure theoretical approach to the mean-field maximum principle for training NeurODEs, Nonlinear Analysis, Volume 227, **2023**, 113161, ISSN 0362-546X, https://doi.org/10.1016/j.na.2022.113161.
8. Benoît, B. A Pontryagin Maximum Principle in Wasserstein spaces for constrained optimal control problems. ESAIM: COCV, 25 **2019** 52 https://doi.org/10.1051/cocv/2019044
9. Cao, H., Guo, X., Laurière, M., Connecting GANs, MFGs, and OT, arXiv e-prints, 2020.
10. Carmona R., Lauriere, M. Deep Learning for Mean Field Games and Mean Field Control with Applications to Finance. In A. Capponi & C. Lehalle (Eds.), Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices (pp. 369-392). Cambridge: Cambridge University Press. **2023** doi:10.1017/9781009028943.021
11. Chizat, L.; Bach, F., On the global convergence of gradient descent for overparameterized models using optimal transport. In Advances in neural information processing systems, **2018**, pages 3040–305.
12. Chizat, L.; Colombo, M., Fernández-Real, X.; and Figalli, A. Infinite-width limit of deep linear neural networks, arXiv e-prints, **2022**. doi:10.48550/arXiv.2211.16980.
13. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in Neural Information Processing Systems, **2013** pages 2292–2300.
14. de Bie, G.; Peyré, G.; Cuturi, M. Stochastic Deep Networks, Proceedings of the 36th International Conference on Machine Learning, **2019**. Long Beach, California, PMLR 97.
15. Di Persio L., Garbelli M. Deep Learning and Mean-Field Games: A Stochastic Optimal Control Perspective. Symmetry **2021** 13(1):14.
16. E, W.; Han, J.; Li, Q. A mean-field optimal control formulation of deep learning. Res Math Sci 6, 10. **2019** https://doi.org/10.1007/s40687-018-0172-y.
17. Fernández-Real, X.: Figalli, A. The Continuous Formulation of Shallow Neural Networks as Wasserstein-Type Gradient Flows. In: Avila, A., Rassias, M.T., Sinai, Y. (eds) Analysis at Large. Springer, Cham. **2022** https://doi.org/10.1007/978-3-031-05331-3-3
18. Gangbo, W.; Mayorga, S.; Swiech, A. Finite Dimensional Approximations of Hamilton-Jacobi Bellman Equations in Spaces of Probability Measures. SIAM Journal on Mathematical Analysis **2021** 53:2, 1320-1356
19. He, K.; Zhang, X.; Ren S.; Sun, J.; "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
20. Han Huang, Jiajia Yu, Jie Chen, Rongjie Lai, Bridging mean-field games and normalizing flows with trajectory regularization, Journal of Computational Physics, Volume 487, 2023, 112155, ISSN 0021-9991, https://doi.org/10.1016/j.jcp.2023.112155.
21. Jimenez, C.; A. Marigonda A.; Quincampoix, M. Dynamical systems and Hamilton-Jacobi-Bellman equations on the Wasserstein space and their L2 representations, **2022**. Preprint at https://cvgmt.sns.it/media/doc/paper/5584/AMCJMQ_HJB_2022-03-30.pdf
22. Hashimoto, T.; Gifford, D.; Jaakkola, T. Learning population-level diffusions with generative rnns. International Conference on Machine Learning, **2016**, pages 2417–2426.
23. Li Q.; Lin T.; Shen, Z.. Deep Learning via Dynamical Systems: An Approximation Perspective. **2019** Published in arXiv:1912.10382v1.
24. Li, Q., Long, C., Cheng, T.; E, W. Maximum principle based algorithms for deep learning. J. Mach. Learn. Res. **2017** 18, 1, 5998-6026.
25. Mikami, T.; Two End Points Marginal Problem by Stochastic Optimal Transportation SIAM Journal on Control and Optimization **2015** 53:4, 2449-2461

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 June 2023 doi:10.20944/preprints202305.0252.v3

26. Pham, H.; Warin, X. Mean-field neural networks: learning mappings on Wasserstein space, **2022** arXiv e-prints.
27. Peyré G.; Cuturi M. Computational Optimal Transport: With Applications to Data Science, Foundations and Trends in Machine Learning, **2019**. Vol. 11: No. 5-6, pp 355-607. http://dx.doi.org/10.1561/2200000073
28. Sirignano, J.; Spiliopoulos, K. Mean Field Analysis of Deep Neural Networks. Mathematics of Operations Research **2021**, 47(1):120-152.
29. Tan, X.; Nizar Touzi, N.; Optimal transportation under controlled stochastic dynamics." The Annals of Probability, 41(5) 3201-3240 **2013**. https://doi.org/10.1214/12-AOP797
30. Wojtowytsch, S., On the Convergence of Gradient Descent Training for Two-layer ReLU-networks in the Mean Field Regime, arXiv e-prints, **2020**. doi:10.48550/arXiv.2005.13530.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.