Neural Networks and Adapted Optimal Transport

Matteo Garbelli * and Luca Di Persio

Posted Date: 9 May 2023

doi: 10.20944/preprints202305.0252.v2

Keywords: Neural Network; Machine Learning; Adapted Optimal Transport; Mean Field function

Article

Neural Networks and Adapted Optimal Transport

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 Data Science applications, Machine Learning (ML) and Neural Networks (NNs) are widely used in a plethora of fields spanning from engineering to biology, from finance to medicine. Meanwhile, the pure analytical analysis of the related models often lacks detailed description. Trying to close this gap, during recent years the theory of Optimal Transport has become a popular tool within the ML community, since it allows the development of efficient algorithms via a linear programming approach to provide scalable solutions. Furthermore, many ML and statistics tasks require optimizing an objective function over the space of probability measures. Along this line, the training of a Neural Network through Stochastic Gradient Descent has been shown to be equivalent to a gradient flow in Wasserstein space. Starting from these considerations, we review the theoretical aspects of some essential references trying to highlight potential symmetries between game theory, applied probability and ML. Moreover, in view of providing a possibly unifying and efficient approach to insert the temporal structure of the data into the training process, we focus on the Adapted Optimal Transport (AOT) method. AOT encodes an adapted (non-anticipative) constraint into the allocation of mass of the classical OT problem while being particularly useful to define laws of stochastic processes. Accordingly, we provide an in-depth exploration of AOT, providing theoretical insights into the benefits of including causality constraints for the development of robust and scalable algorithms.

Keywords: neural network; machine learning; adapted optimal transport; mean field function

1. Introduction

The present paper aims at presenting a baseline for the problem of including causality constraint into Machine Learning (ML) tasks as to increase algorithms robustness, scalability and accuracy, particularly when dealing with structured time series-based applications. The analysis starts from a preliminary study of a Supervised Learning problem in terms of a Mean Field Optimal Control Problem (MFOCP), hence providing an overview about the most relevant theoretical aspects that can be useful for a generalization of MFOCP to include the temporal structure of the training data. Accordingly, we provide a complete review about Adapted Optimal Transport (AOT), that is a version of Optimal Transport (OT) enforcing an adapted (with respect to a filtration) constraint. Furthermore, we summarizes state-of-the-art works at the crossroad of probability theory, ML and OT methods. We deepen the introduction of probability measures into ML tasks specifically for Neural Networks (NNs) applications in the context for explainable AI. On the other hand, the use of probability measures can also help to analyze the behavior of specific algorithm, while NNs can be used to approximate functions depending on unknown distributions. Trying to clarify the interplay between probability measures and Neural Networks, we recall possible cross-related cases:

1. *Measure approach for NN training*: Probability measures can be used to analyze the behavior of neural networks, both during training and inference. For example, the loss function used to train a neural network can be seen as a divergence between the model distribution and the target distribution, which can be expressed in terms of probability measures, [17]. Similarly, the output of a neural network can be interpreted as a probability distribution over the output space, which can be analyzed using probability theory. This can be useful for understanding the uncertainty of the predictions made by the neural network, such as the Stochastic Deep Networks (SDN)

introduced in [25]. The design of a SDN involves a deep architecture that can handle probability distribution inputs and outputs and uses the classical Wasserstein distance as the loss function. Moreover, layers correspond to a sequence of elementary blocks that maps random vectors to random vectors to capture multiple interactions between samples from the distributions.

2. Probability measures as a tool to *regularize neural networks*: Probability measures can be used as a tool to regularize neural networks, by adding a regularization term to the loss function that encourages the output of the neural network to match a prior distribution by the so called entropic measure, such as in Eq. (26). Within this case, we mention the entropic version of Sinkhorn algorithm where an entropic regularization is added, developed, e.g., in [40]. This algorithm can be useful for scaling OT problem in high dimensional setting or for generative modeling tasks, where the goal is to learn a distribution that can generate new data samples.
3. Neural networks to *approximate empirical measures*: Neural networks can be used to approximate probability measures, such as empirical distributions over a state space for mean field function. We specifically consider this feature in Section 3. Analogously, one can consider NN architecture for learning population dynamics. For example, in [33], the authors focus on a recurrent architecture to model the diffusion of a population by means of a NN by injecting random noise.
4. Neural Network to *learn unknown probability measure* for unsupervised learning tasks such as density estimation, where the goal is to learn the underlying distribution of the data. For example, in [21], the authors try to build an algorithm to directly learn a probability measure via an optimal transport metric, also providing some convergence results for the learning algorithm.

The paper is organized in the following paragraphs. In Section 2, we review some basic works dealing with the mathematical formalism and definitions needed to precisely address supervised learning problems within the context of NNs. In Section 3, we review some developed methods to *learn*, i.e. approximate, mean field functions that depend on probability distribution obtained as limit object of empirical measures. In 4, we deal with the problem of optimizing over the set of probability measures with finite 2-moment focusing on JKO-net, an efficient and scalable method to optimize gradient flow in Wasserstein space. In Section 5, we conclude studying variants of the classical Optimal Transport problem, specifically focusing on AOT by also considering the class of equivalence of filtered processes and associated adapted empirical measures.

2. A continuous idealization of NN for a Mean Field Optimal Control Problem

In this section, we present the basic workflow to treat a feed-forward NN as a dynamical system in order to derive the associated population risk minimization problem.

Indeed, in [35], a continuous idealization of Deep Learning (DL) is introduced, to study the Supervised Learning (SL) problem as am Optimal Control Problem, see, e.g. [27] and [36] for more details.

Following [26], SL aims at estimating the function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$, commonly known as the Oracle, where $\mathcal{X}$ is a subset of $\mathbb{R}^d$ that contains input arrays such as images, texts, and time series, and $\mathcal{Y}$ is the corresponding target set. To achieve this, training begins with a set of K input-target pairs $\{x_0^i, y_T^i = \mathcal{F}(x_T^i)\}_{i=1}^K$ where

- $x_0 \in \mathbb{R}^d$ denotes the input of the NN;
- $x_T \in \mathbb{R}^d$ denotes the output of the NN:
- $y_T \in \mathbb{R}^l$ the corresponding target.

For a T-layers network, the Feed-Forward propagation is given by

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

θ_t being the trainable parameters (e.g. bias, weights) in layer t that belong in to a measurable set $\mathcal{U}$ with values in subset of the euclidean space $\mathbb{R}^m$.

We aim at minimizing over the set of measurable parameters $\mathcal{U}$ a *terminal loss function* $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ plus a regularization term L over a set of measurable parameters $\mathcal{U}$ to control the *size* of parameters. In this way, we can state a Supervised Learning problem as an Empirical Risk Minimization problem, namely

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

2.1. Mean Field Optimal Control Problem

The next steps involve moving from the discrete setting to the corresponding continuous idealization by:

- going from layer index T to continuous parameter t;
- passing from discrete set of inputs/output to distribution μ_0 that represents the joint distribution in $\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^l)$ modelling the input-label distribution;
- identifying targets y as random variables sampled from the projection over $\mathcal{Y}$ of the distribution μ_0;
- passing from empirical risk minimization to population risk (i.e. minimization over expectation $\mathbb{E}$).

Specifically, we pass to the limit both in the layer discretization to obtain a dynamic described by an Ordinary Differential Equation (ODE) instead of finite difference equation (1). Also we identify the input-target pairs as samples from a given distribution μ_0 allowing to write a SL problem as a population risk minimization problem.

In summary, the problem aims 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).

More specifically, we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we assume inputs x_0 in $\mathbb{R}^d$ being sampled from a distribution $\mathcal{P}(\mathbb{R}^d)$ and the corresponding target y_T in $\mathbb{R}^l$ belongs to $\mathcal{P}(\mathbb{R}^l)$. These variables are considered to be random variables that are jointly distributed according to μ_0, which is a distribution in the Wasserstein space $\mathcal{W}_2(\mathbb{R}^{(d+l)})$. 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, i.e., measures μ such that

$$\int_X d(x, y)^p d\mu(x) < \infty \tag{3}$$

for some $y \in X$. Here, $d(x, y)$ is the metric between x and y, and p is a positive real number. The Wasserstein space can be endowed with the Wasserstein distance, also known as the p-th Wasserstein distance, which is a metric that measures the distance between two probability measures μ and ν in $\mathcal{W}_p(X)$ as the infimum of the expected cost of transporting the mass from μ to ν:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X \times X} d(x, y)^p d\gamma(x, y) \right)^{1/p} \tag{4}$$

where $\Gamma(\mu, \nu)$ is the set of all probability measures on $X \times X$ with marginals μ and ν.

In our setting, the joint measure μ_0 that models the distribution of the input-target pairs and is defined by $\mu_0 := \mathcal{P}(x_0, y_T)$ is studied as an element of $\mathcal{W}_2$.

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

$$\inf_{\theta \in L^\infty([0,T],\Theta)} J(\theta) := \mathbb{E}_{\mu_0} \left[\Phi\left(x_T, y_T\right) + \int_0^T L(x_t, \theta_t) dt \right] \tag{5}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2023 doi:10.20944/preprints202305.0252.v2

Peer-reviewed version available at Symmetry 2023, 15, 1724; doi:https://doi.org/10.3390/sym15091724

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

We briefly report basic assumptions we need to have a solution for (5):

- $f : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$, $L : \mathbb{R}^d \times \Theta \to \mathbb{R}$, $\Phi : \mathbb{R}^d \times \mathbb{R}^l \to \mathbb{R}$ are bounded;
- f, L, Φ are Lipschitz continuous with respect to x with the Lipschitz constants of f and L being independent from parameters θ;
- μ_0 has finite support in $\mathcal{W}_2(\mathbb{R}^{(d+l)})$

Problem (5) 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 [30] and [32] for a viscosity solutions to the HJB equation in the Wasserstein space of probability measures and to [19] for a Pontryagin Maximum Principle for constrained optimal control problems in the Wasserstein space of probability measures.

About some generalizations of the presented result, we cite the article [16] where the authors introduce a BSDE technique to solve the Stochastic Maximum Principle to examine the uncertainty associated with DL. The authors employ an Stochastic Differential Equation (SDE) in place of the ODE appearing in (5) to continuously approximate a Stochastic Neural Network (SNN).

2.2. Empirical measures over controls

For the sake of completeness, we mention also a completely different perspective to describe the learning process in a NN, that deals with the introduction of the empirical distribution of the parameters of neurons (instead of the input-target pairs). Indeed, as illustrated, e.g., by Sirignano and Spilipupouls in [43], it is possible to associate to each layer the corresponding empirical measure and build a measure to describe the whole network that takes values in the space of the called nested measure (also known as adapted Wasserstein distance), that we deepen in Section 5. Following this perspective, the method of Stochastic Gradient Descent (SGD) can be formalized as minimization over probability distribution. Moreover, the training of NN is based on the correspondence between empirical measure of neurons μ_N and the approximated function f_N. 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 [23,26,27,29]. Similarly, the convergence to solutions to SDEs have been obtained as universal continuum objects in the small learning rate regime as proved in, e.g., [22].

3. Learning Mean Field Function

In this section, we study some tools in order to learn functions showing a dependency on a mean field term, such as an empirical distribution over a given state space. We start with a brief description of a dynamica of McKean-Vlasov type.

3.1. The McKean-Vlasov SDE

The McKean-Vlasov dynamic involves a SDE that models the behavior of a large population of interacting particles. In this context, the optimal control problem for this system involves finding a control strategy that minimizes a certain cost function while satisfying the dynamics of the system. In particular, the McKean-Vlasov dynamics can be written as:

$$dX_t = b(X_t)dt + \sqrt{\epsilon}dW_t + \frac{1}{N}\sum_{i=1}^{N}\int K(X_t - X_i(t))d\mu_i(t)dt, \tag{6}$$

where X_t is the state of the system at time t, $b(X_t)$ is the drift term, ϵ is the diffusion coefficient, W_t is a standard Brownian motion, N is the number of particles in the population, $\mu_i(t)$ is the empirical

measure of the particles at time t, and K is a kernel function. Accordingly, the optimal control problem for the McKean-Vlasov dynamics can be written as:

$$J(u) = \mathbb{E}\left[\int_0^T L(X_t, u_t)dt + \phi(X_T)\right], \tag{7}$$

subject to the dynamics of the system:

$$dX_t = b(X_t)dt + \sqrt{\epsilon}dW_t + \frac{1}{N}\sum_{i=1}^{N}\int K(X_t - X_i(t))d\mu_i(t)dt + u_t dt, \tag{8}$$

where u_t is the control input, $L(X_t, u_t)$ is the instantaneous cost, and $\phi(X_T)$ is the terminal cost.

A possible approach to the optimal control problem consists in using the Pontryagin Maximum Principle, which gives necessary conditions for optimality. The Hamiltonian of the system is defined as:

$$H(X_t, u_t, p_t) = L(X_t, u_t) + p_t \cdot \left[b(X_t) + \frac{1}{N}\sum_{i=1}^{N} K(X_t - X_i(t))\right] + \frac{\epsilon}{2}|p_t|^2, \tag{9}$$

where p_t is the adjoint variable. The optimal control is given by:

$$u_t^* = \arg\min_u H(X_t, u, p_t). \tag{10}$$

The adjoint variable satisfies the following (backward) differential equation:

$$dp_t = -\frac{\partial H}{\partial X}(X_t, u_t^*, p_t)dt, \quad p_T = \frac{\partial \phi}{\partial X}(X_T). \tag{11}$$

3.2. Mean Field Optimal Transport

In the recent work [12], the authors introduce the Mean Field Optimal Transport (MFOT) problem in the field of Mean Field Control, namely optimal control for McKean-Vlasov dynamics. Differently from classical Mean Field Game (MFG) theory, the final distribution is prescribed while all the agents cooperate in order to minimize a total cost without terminal cost. Indeed we follow the numerical scheme introduced in section 3.1 in [12] to approximate feedback controls.

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

$$J^{MFOT} : v \mapsto \mathbb{E}\left[\int_0^T f(X_t^v, \mu^v(t), v(t, X_t^v))dt\right] \tag{12}$$

being $\mu^v(t)$ the distribution of X_t^v under the constraint X^v solves the SDE

$$\begin{cases} X_0^v \sim \rho_0 \qquad X_T^t \sim \rho_T \end{cases} dX_t^v = b(X_t^v, \mu^v(t), v(t, X_t^v))dt + \sigma dW_t\,, \qquad t \in [0, T] \tag{13}$$

Always in [12], the authors present different numerical methods 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.

Peer-reviewed version available at Symmetry 2023, 15, 1724; doi:https://doi.org/10.3390/sym15091724

In this section, we focus on the direct method (1) to approximate controls of feedback type by an optimal control formulation. The controls are assumed of feedback form and they are approximated by

$$g(x,\mu) = G(\mathcal{W}_2(\mu,\rho_T)), \qquad \mu \in \mathcal{P}_2(\mathbb{R}^d) \tag{14}$$

being $G : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function.

The idea is to use the function in Eq. (14) as a penalty for being far from the target distribution ρ_T instead of the classical terminal cost typical of the MFG/MFC literature. Moreover, discretizing the SDE (13) by an Euler-Maruyama scheme as well as restricting the set of controls v to ones approximated via NNS v_θ with parameters θ helps the development of numerical algorithms. See Section 3.1 in [12] for further details.

3.3. Alternative methods

Another proposed data-driven approach, presented in [8], 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 [37], 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.

4. Learning Gradient Flow

Numerous problems in ML and statistics deal with minimizing some objective function over probability distributions. A powerful tool for optimizing functionals relies on gradient flow in the Wasserstein space.

4.1. Gradient Flow in Wasserstein space

Beyond OT problems, a popular use of Wasserstein distances is to study gradient flows, following the formalism of minimizing movements studied in [3]. This corresponds to discrete implicit stepping (i.e. proximal maps) for the Wasserstein distance (instead of more common Euclidean or Hilbertian metrics) functionals, for instance to model crowd motions.

In the discrete-time setting, gradient flow in Wasserstein space can be used to study the evolution of probability measures over a sequence of discrete time steps. Consider a probability measure ρ^t over a finite Euclidean space. The gradient flow of the Wasserstein distance can then be used to find a sequence of probability measures ρ^t, representing, e.g., the evolution of a distribution of a cloud of points over a finite set $X \in \mathcal{R}^n$. The Wasserstein distance between two consecutive measure ρ^t and ρ^{t+1} is minimized along this sequence of measures, which can be seen as a discrete-time analogue of the continuous-time gradient flow.

The dynamic formulation of OT can be solved by means of the so-called Jordan-Kinderlehrer-Otto (JKO) scheme, which consists in an implicit Euler discretization of a Wasserstein gradient flow

$$\rho^k = argmin_{\rho\in\mathcal{P}_2(\mathbb{R}^n)} \left[\mathcal{F}(\rho) + \frac{1}{2\tau}\mathcal{W}_2^2(\rho^{k-1},\rho)\right], \tag{15}$$

with $\mathcal{F}$ being an energy functional and τ the time-discretization update. The solution of this consecutive minimization problems result in a sequence of probability measures in $P_2(\mathbb{R}^n)$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2023 doi:10.20944/preprints202305.0252.v2

Peer-reviewed version available at Symmetry 2023, 15, 1724; doi:https://doi.org/10.3390/sym15091724

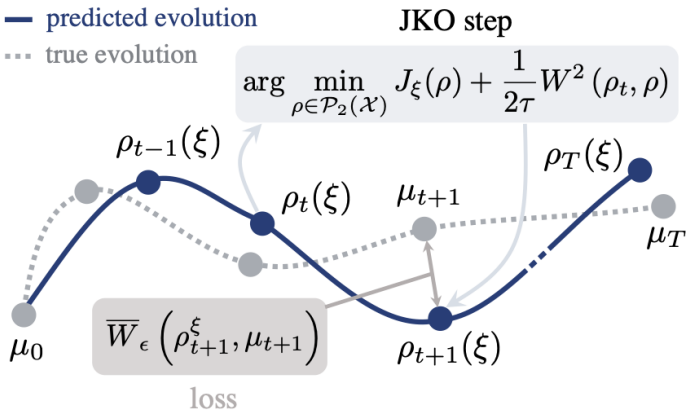

Figure 1. Proximal OT for a population dynamics (from [18]).

The idea is to move from optimization over probability measure of Eq. (15) to an optimization over convex function, which is still challenging, especially in high dimensions. From Brenier's theorem, for any $\rho \in P_{2,ac}$, there exists a unique ρ^{k-1}-measurable gradient $\nabla\psi : \mathbb{R}^D \to \mathbb{R}^D$ of a convex function ψ s.t. $\rho = \nabla\psi\#\rho^{(k-1)}$

4.2. Stochastic Optimization with ICNN and JKOnet

A large variety of dedicated numerical schemes has been proposed for spatial discretization and solving of these time-discrete flows, such as for instance finite differences and Lagrangian schemes. A bottleneck of these schemes is the high computational complexity due to the resolution of an OT-like problem at each time step.

In [7], the authors propose an approach that relies on Input-Convex Neural Networks (ICNN), a class of deep models with a convex output with respect to their inputs introduced in [13], to parameterize the space of convex functions $\{\psi_\theta\}$in order to approximate the JKO scheme (15) through

$$\theta^* = argmin_\theta \left[\mathcal{F}(\nabla\psi_\theta\#\rho^{(k-1)}) + \frac{1}{2h}\int_{\mathbb{R}^D} ||x - \nabla\psi_\theta(x)||_2^2 d\rho^{(k-1)}(x)\right]$$

The authors in [18] investigate the dynamics of a diverse set of points within the framework of Proximal OT using a Jordan-Kinderlehrer-Otto (JKO) flow of measures. These points are observed periodically at different timestamps. The authors propose a 2-level algorithm to capture the trajectories of this population. Specifically, at time $t+1$, a new configuration is generated that reduces the energy while remaining close to the previous configuration observed at time t, as measured by the Wasserstein distance.

To learn an energy function only through snapshots, the *JKOnet* algorithm is proposed to compute the JKO flow by a 2-time scale learning algorithm:

- an *upper level* optimization aiming at minimizing the Wasserstein distance w.r.t. to an energy functional J_ψ.

- a *lower level optimization* of the constrained parameters θ of a ICNN, [13].

In Figure 2, we can see an observed trajectory $(\mu_0, ..., \mu_T)$ of point clouds, we seek parameters ψ for the energy J_ψ such that the predictions $\rho_1, ..., \rho_T$ following a JKO flow from $\rho_0 = \mu_0$ are close the observed trajectory, by minimizing (as a function of ψ) the sum of Wasserstein distances between ρ_{t+1}, the JKO step from ρ_{t-1} using J_ψ, and data μ_{t+1}.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2023 doi:10.20944/preprints202305.0252.v2

Peer-reviewed version available at Symmetry 2023, 15, 1724; doi:https://doi.org/10.3390/sym15091724

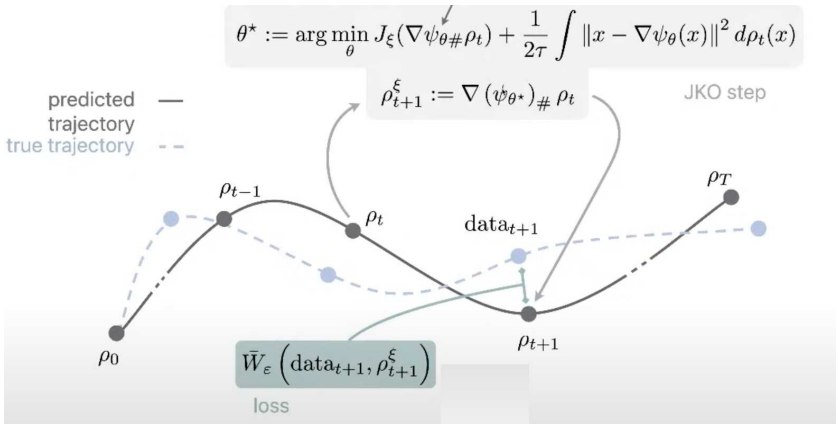

Figure 2. JKOnet 2-level optimization process (from [18]).

In [31], the authors deal with deep network generative models focusing on the problem of comparing two probability distributions that are degenerate, i.e. distribution in higher-dimensional spaces supported on low-dimensional manifolds, while estimating the parameters of a chosen model that fits observed data.

5. From OT to AOT

After reviewing some variants of the classical OT problem, we present some definitions concerning AOT, i.e. an adapted version of OT.

5.1. OT variants

We start with a brief review of classical OT. Given two marginal distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, the classical OT problem reads

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c(x.y)\pi(dx, dy) \tag{16}$$

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

We focus on the setting where μ and ν are distributions computed, e.g., from time series 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

$$\inf_{T:T\#\mu=\nu} \int c(x, T(x))\mu(dx) \tag{17}$$

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

Some generalizations of the classical 2-marginal OT problem concern:

- *Martingale Optimal Transport.* A first model to encode a temporal adapted structure into OT problems moves within the theory Martingale Optimal Transport (MOT) dealing with transport plans π with martingale couplings. The stability of a sequence of transport plans $\{\pi^1, \ldots, \pi^N\}$ with martingales marginals has been proved to hold in [10], also for the Weak Optimal Transport.
- *Multi marginal optimal transport problem.* Multi-marginal optimal transport problems introduced in [4] are formulated as follows. Consider $M \in \mathbb{N}$ probability measures $\mu_1, \ldots, \mu_M$ on $\mathbb{R}^d$ and consider the optimization problem

$$\inf_{\pi \in \Pi(\mu_1, \ldots, \mu_M)} \int_{\left(\mathbb{R}^d\right)^M} c\left(x_1, \ldots, x_M\right) d\pi\left(x_1, \ldots, x_M\right) \tag{18}$$

with c being a lower semi-continuous cost function defined on $\left(\mathbb{R}^d\right)^M$ with marginal laws $\mu_1, ..., \mu_M$ and $\Pi(\mu_1, \dots, \mu_M) = \{\pi \in \mathbb{R}^{\lceil} \times \dots \times \mathbb{R}^{\lceil} \mid \quad \forall 1 \leq i \leq M, \quad \int_{\mathbb{R}^d} d\pi = \mu_i\}$

Trying to approximate this problem by discretizing the state space through N points, $x^1, \dots, x^N \in \mathbb{R}^d$ chosen and a priori fixed, lead to a linear programming problem of size N^M.

- *From Martingale-constrained to Moment Constrained Optimal Transport* . In [5,6], the authors develop algorithms to sample from probability distribution while preserving the convex order enabling to use linear programming solver.

  The same authors study the same problem also in [4] following another perspective by relaxing the martingale constraint into a finite number of moment constraints by means of N real-valued bounded (test) function $\phi_1, \dots, \phi_N$ defined on $\mathbb{R}^d$.

  This method is called Moment Constrained Optimal Transport (MCOT). Differently from multi marginal setting, it is not the state space to be discretized but the marginal laws constrained of Eq. (18) are relaxed into a finite number of moment constraints via some given $\{\phi_N\}_{1,\dots,N}$ It solves the following:

$$\inf \int_{\mathbb{R}^d \times \mathbb{R}^d} c\,(x, y)\, d\pi\,(x, y) \tag{19}$$

  where the infimum is computed over the set of probability measures π with values in $\mathbb{R}^d \times \mathbb{R}^d$ satisfying for all $1 \leq i, j \leq N$

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \phi_i(x) d\pi(x, y) = \int_{\mathbb{R}^d} \phi_i(x) d\mu(x) \tag{20}$$

  and

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \phi_j(x) d\pi(x, y) = \int_{\mathbb{R}^d} \phi_j(y) d\nu(y) \tag{21}$$

  Always in [4], they prove that MCOT is a generalization of Martingale constrained OT when by means of L^1-valued test functions and characteristic functions to enforce the martingale constraint (see Equation (3.3.11) from [4]).

5.2. Adapted Optimal Transport

Adapted Optimal Transport (AOT) is the natural problem to deal with laws of stochastic processes, as shown in e.g. [1,14,15,34,38]. Intuitively, causality encodes an adapted (nonanticipative) constraint into the allocation of mass of the classical optimal transport problem. The relationship between causal plans and adapted processes is the adapted analogous of the one between transport plans (Kantorovich) and classical transport maps (Monge).

The basic idea of COT relies on the assumption that the transport maps T introduced in (17)have to be adapted.

If, in the classical setting one can split T component-wise by $T(x) = (T_1, (x), \dots, T_d(x)$, in the COT setting the adapted maps are defined in the form $T(x) = (T_1, (x_1), T_2(x_1, x_2) \dots, T_d(x)$ where T_t depends only on the past component of x, namely $x_1, \dots, x_t$. Thus, the adapted analogous of (17) is defined as

$$\inf_{T:T\#\mu=\nu,,\, T \text{ adapted}} \int c(x, T(x))\mu(dx) \tag{22}$$

Concerning coupling, the adaptness of T means that under π, the first component of Y until component t should be independent of the component of X after t. This construction allows to include the conditional independence of Y into the set Π_c of arbitrary causal couplings that is defined by

$$\Pi_c(\mu,\nu) = \{\pi \in \Pi(\mu,\nu) \,:\, \textit{if}\,(X,Y) \sim \pi, \\ \textit{then}\,(Y_1,\ldots,Y_t) \perp\!\!\!\perp (X_{t+1},\ldots,X_d)\,|\,(X_1,\ldots,X_t)\}$$

Optimizing over the set of causal coupling let us formulate the first *AOT* problem

$$\inf_{\pi\in\Pi_c(\mu,\nu)} \int c(x.y)\pi(dx,dy) \tag{23}$$

that is called Causal OT.

In order to provide a symmetrical property that is desirable to give a more robust structure to the problem and also necessary to introduce a proper notion of adapted Wasserstein distance, one may introduce the set of bicausal couplings Π_{bc}

$$\Pi_{bc}(\mu,\nu) = \{\pi \in \Pi(\mu,\nu) \,:\, \textit{if}\,(X,Y) \sim \pi, \\ \textit{then}\,(Y_1,\ldots,Y_t) \perp\!\!\!\perp (X_{t+1},\ldots,X_d)\,|\,(X_1,\ldots,X_t)\} \\ \textit{and}\,(X_1,\ldots,X_t) \perp\!\!\!\perp (Y_{t+1},\ldots,Y_d)\,|\,(Y_1,\ldots,Y_t)\} \tag{24}$$

and the equivalent Bicausal-OT formulation

$$\inf_{\pi\in\Pi_{bc}(\mu,\nu)} \int c(x.y)\pi(dx,dy) \tag{25}$$

To fix nomenclature, we call Causal OT (23) and Bicausal OT (25), AOT problems. As recently proved in [10], it is possible to write a Linear Program formulation for AOT problems (23) and (25) by discretizing the marginal distributions, see Lemma 3.11 in [10] for further details.

As shown in several works, e.g. [20,24,40], to compute efficiently a solution of OT problems, it is convenient to work with a regularized version of (16) by introducing an entropic regularization in order to apply Sinkhorn's algorithm [41]. We refer to the landmark book [40] for a complete overview with a detailed treatment of Sinkhorn application in Section 4, [40].

In the same fashion, we provide to Eq. (25) an entropic regularization term to study the following Entropic-AOT

$$\inf_{\pi\in\Pi_{bc}(\mu,\nu)} \int c(x.y)\pi(dx,dy) + \varepsilon D_{KL}(\pi,\mu\otimes\nu)\,, \tag{26}$$

being ε a positive positive constant, D_{KL} the Kullback-Leibler divergence and $\mu\otimes\nu$ the product measures between marginals.

In [10] the authors construct an adapted version of Sinkhorn algorithm allowing to handle computations by constructing a sequence of transport plans π^k that converges (lineraly) to the optimal value of Eq. (26).

In next paragraphs, we review some of the basic tools of the causal framework: the Adapted Wasserstein distance inherited from AOT, the class of equivalence of filtered processes and the Adapted empirical measure.

5.3. Adapted Wasserstein Distance

The starting point is to consider stochastic processes on the canonical space in finite discrete time with values in $\mathbb{R}^d$. We can represent the law of a stochastic process as a probability measure on $\mathcal{P}(\mathbb{R}^d)$. Even at this very early stage, it is clear that the usual weak topology on $\mathcal{P}(\mathbb{R}^d)$ is not enough for encoding the temporal structure of the process that is crucial for stochastic optimization problem.

Thus, different researchers have introduced topological structure on the set of stochastic processes in order to capture the temporal structure. Among different approaches on defining a distance for multistage stochastic optimization, we focus on nested distance also known as adapted Wasserstein distance.

The notion has been independently defined by Pflug and Pichler [38] in a pure distribution setting, i.e. without reference to probability spaces, by means of particular stochastic processes, called tree processes, and by Lassalle [34] that introduces the so-called causal transport plans in Polish spaces.

Following a dynamical setting, adapted Wasserstein distance can model a distance between laws of stochastic processes integrating also the temporal structure of the data.

Recalling notation and definition from Eq. (25), we define the adapted Wasserstein distance as

$$\mathcal{AW}_p^p(\mu,\nu) \;=\; \inf_{\pi\in\Pi_{bc}(\mu,\nu)} \mathbb{E}^{\pi}[||X-Y||_p^p]\,, \qquad p\in[1,+\infty] \tag{27}$$

with Π_{bc} being the set of bicausal couplings defined in Eq. (24).

The nested distance represents a compatible metric for the weak adapted topology. More precisely, as proven in [9], the convergence $\mathcal{AW}_p$ is equivalent to the convergence in the weak topology plus convergence of the p^{th} moment. On the other hand, a possible limitation is due to the fact that $\left(\mathcal{P}_p(\mathbb{R}^d),\mathcal{AW}_p\right)$ is not complete since $\mathcal{P}_p(\mathbb{R}^d)$. According to [11], a possible method to store the extra information needed for achieving the completion of $\left(\mathcal{P}_p(\mathbb{R}^d),\mathcal{AW}_p\right)$ is by introducing the following space of filtered stochastic processes.

5.3.1. The class of filtered processes

We recall Definition 1.1 from [11]: a five-tuple

$$\mathbb{X} \coloneqq \left(\Omega,\mathcal{F},\mathbb{P},(\mathcal{F}_t)_{i=1}^N,(X_t)_{i=1}^N\right),$$

is called a filtered stochastic processes if $(\mathcal{F}_t)_{i=1}^N$ is adapted to $(X_t)_{i=1}^N$.

It is straightforward to extend the notion of adapted Wasserstein distance defined in (27) to filtered processes by the following

$$\mathcal{AW}_p^p(\mathbb{X},\mathbb{Y}) \;=\; \inf_{\pi\in\Pi_{bc}(\mathbb{X},\mathbb{Y})} \mathbb{E}^{\pi}[||X-Y||_p^p]\,, \qquad p\in[1,+\infty]\,; \tag{28}$$

We report some properties about the Wasserstein space of stochastic processes:

- $(\mathcal{FP}_p,\mathcal{AW}_p)$ is a geodesic space (assuming $p>1$) and it is the completion of $\left(\mathcal{P}_p(\mathbb{R}^d),\mathcal{AW}_p\right)$. The first property suggests the possibility an interpolation, in the sense of McCann, also of stochastic processes that is not possible if one considers the usual Wasserstein space $\mathcal{P}_p(\mathbb{R}^d)$;
- two filtered processes $\mathbb{X}$ and $\mathbb{Y}$ are identified if $\mathcal{AW}_p(\mathbb{X},\mathbb{Y})=0$. Moreover, this equivalence relation encodes also information about their filtration.
- the space of filtered stochastic processes is defined as a set of equivalence classes as in the case of L^p spaces. An equivalence class of $\mathcal{FP}_p$ collects all representatives processes that are identical from a probabilistic perspective.

5.3.2. Adapted Empirical Measure

Another key concept in AOT theory concerns constructing empirical measures of stochastic processes also including a dependence on the flow of time.

We report three possible methods to construct adapted empirical measures:

- in [38], adapted empirical distribution is constructed via convolution of a smooth kernel with the empirical measure on paths;
- in [14,15] the authors project paths on a grid by means of a disjoint union of small cubes projecting the stochastic process into the center of these small cubes;
- in [1] the authors show the convergence of adapted empirical measure in $\mathbb{R}^d$, removing the assumption of compact support for the underlying distribution. Moreover, the authors introduce

Peer-reviewed version available at Symmetry 2023, 15, 1724; doi:https://doi.org/10.3390/sym15091724

the non-uniform adapted empirical measures by defining a non-uniform grid on $\mathbb{R}^{dT}$, dealing with cubes of different sizes, i.e. of lower dimension near the origin, bigger far in the margins.

In all works, they establish convergence of the empirical measure to the true distribution.

6. Conclusions

The present review article provides an all-around and diversified overlook of methods at the overlapping of ML, game theory and probability theory. In doing this, we tried to provide a concrete outlook to possible developments of a challenging research path focused on developing robust algorithms for Adapted optimal transport. Indeed, employing the Sinkhorn algorithm with the addition of the entropic regularization potentially allows to derive a method that can handle the curse of dimensionality. On the other hand, OT problems are sensitive to small perturbations in the input data, which can make the solutions unstable. The application of their convergence results in a causal setting could lead to improved training algorithms and more stable models in the field of causal inference for times series.

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.

Sample Availability: Samples of the compounds ... are available from the authors.

Abbreviations

The following abbreviations are used in this manuscript:

OT	Optimal Transport
COT	Causal Optimal Transport
AOT	Adapted Optimal Transport
NN	Neural Network
MFOCP	Mean Field Optimal Control Problem
MFG	Mean Field Games
MFC	Mean Field Control
ML	Machine Learning
DL	Deep Learning
MFOT	Mean Field Optimal Transport
SDE	Stochastic Differential Equation
ICNN	Input-Convex Neural Networks
JKO	Jordan-Kinderlehrer-Otto
SNN	Stochastic Neural Network
SGD	Stochastic Gradient Descent
HJB	Hamilton-Jacobi-Bellman
ODE	Ordinary Differential Equation
MCOT	Moment Constrained Optimal Transport
MOT	Martingal Optimal Transport

References

1. Acciaio, B.; Hou, S. Convergence of Adapted Empirical Measures on $\mathbb{R}^d$. **2022** arXiv e-prints. doi:10.48550/arXiv.2211.10162
2. Acciaio B.; Kratsios A.; Pammer G. Metric hypertransformers are universal adapted maps. **2022** Preprint, arXiv:2201.13094.
3. Ambrosio L.; Gigli N.; Savaré, G- Gradient flows: in metric spaces and in the space of probability measures, Springer Science & Business Media, **2008**.
4. Alfonsi, A.; Coyaud, R.; Ehrlacher, V.; Lombardi, D. Approximation of Optimal Transport problems with marginal moments constraints. Mathematics of Computation, **2020**, ⟨10.1090/mcom/3568⟩. ⟨hal-02128374⟩
5. Alfonsi A.; Corbetta, J.; Jourdain, B. Sampling of one-dimensional probability measures in the convex order and computation of robust option price bounds. International Journal of Theoretical and Applied Finance, **2019**0(0):1950002, 0.
6. Alfonsi A.; Corbetta, J.; Jourdain, B. Sampling of probability measures in the convex order by Wasserstein projection. arXiv e-prints, page arXiv:1709.05287, Sep **2017**.
7. Alvarez-Melis, D.; Schiff, Y.; Mroueh, Y. Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks, arXiv e-prints, **2021**.
8. 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**.
9. Backhoff-Veraguas, J.; Bartl, D.; Beiglblock, M.; and Eder, M. All adapted topologies are equal. Probab. Theory Related Fields, 178(3-4):1125–1172, **2020**.
10. Backhoff-Veraguas, J.; Pammer, G. Stability of martingale optimal transport and weak optimal transport. Ann. Appl. Probab. 32 (1) 721 - 752, February **2022**. https://doi.org/10.1214/21-AAP1694
11. Bartl, D.; Beiglböck, M.; Pammer, G. The Wasserstein space of stochastic processes. arXiv e-prints, **2021**. doi:10.48550/arXiv.2104.14245.
12. 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
13. Brandon A.; Xu, L.; Kolter, J.Z. Input Convex Neural Networks Proceedings of the 34th International Conference on Machine Learning, **2017**, PMLR 70:146-155.
14. Backhoff-Veraguas J.; Beiglboeck, M.; Lin, Y.; Zalashko, A. Causal transport in discrete time and applications. SIAM Journal on Optimization, **2017**, 27(4):2528–2562.
15. 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
16. Bao, F.; Cao, Y.; Archibald, R.; Zhang, H. Uncertainty quantification for deep learning through stochastic maximum principle. arXiv: 3489122, **2021**.
17. 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.
18. Bunne, C.; Meng-Papaxanthos, L.; Krause, A.; Cuturi, M. Proximal Optimal Transport Modeling of Population Dynamics, arXiv e-prints, **2021**.
19. 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
20. Carlier, G. On the linear convergence of the multi-marginal Sinkhorn algorithm. SIAM Journal on Optimization, **2022**, 32 (2), pp.786-794.
21. Canas, G.; Rosasco, L. Learning Probability Measures with respect to Optimal Transport Metrics, Advances in Neural Information Processing Systems 25, **2012**
22. 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.
23. 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.
24. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in Neural Information Processing Systems, **2013** pages 2292–2300.

25. 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.
26. Di Persio L., Garbelli M. Deep Learning and Mean-Field Games: A Stochastic Optimal Control Perspective. Symmetry **2021** 13(1):14.
27. 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.
28. Feydy, J.; Séjourné, T.; Vialard, F.X.; Amari, S.; Trouve, A.; Peyré, G. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, **2019**, PMLR 89:2681-2690.
29. 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
30. 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
31. Genevay, A.; Peyre, G.; Cuturi, M. Learning Generative Models with Sinkhorn Divergences. Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, **2018**, PMLR 84:1608-1617.
32. 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
33. Hashimoto, T.; Gifford, D.; Jaakkola, T. Learning population-level diffusions with generative rnns. International Conference on Machine Learning, **2016**, pages 2417–2426.
34. Lassalle R. Causal transport plans and their Monge–Kantorovich problems, Stochastic Analysis and Applications **2018**, 36:3, 452-484, DOI: 10.1080/07362994.2017.1422747.
35. Li Q.; Lin T.; Shen, Z.. Deep Learning via Dynamical Systems: An Approximation Perspective. **2019** Published in arXiv:1912.10382v1.
36. Li, Q., Long, C., Cheng, T.; E, W. Maximum principle based algorithms for deep learning. J. Mach. Learn. Res. **2017** 18, 1, 5998-6026.
37. Pham, H.; Warin, X. Mean-field neural networks: learning mappings on Wasserstein space, **2022** arXiv e-prints.
38. Pflug G.; Pichler, A.. A distance for multistage stochastic optimization models. SIAM Journal on Optimization, **2012**, 22(1):1-23.
39. Pichler, A.; Weinhardt, M. The nested Sinkhorn divergence to learn the nested distance. Comput Manag Sci **2022** 19, 269–293.
40. 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
41. Sinkhorn, R. Diagonal equivalence to matrices with prescribed row and column sums. The American Mathematical Monthly, **1967**, 74(4):402–405.
42. Seguy, V.; Cuturi, M., Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric, **2015** arXiv e-prints.
43. Sirignano, J.; Spiliopoulos, K. Mean Field Analysis of Deep Neural Networks. Mathematics of Operations Research **2021**, 47(1):120-152.
44. Xu, T.; Li K. W.; Munn, M.; Acciaio, B.; COT-GAN: Generating Sequential Data via Causal Optimal Transport. Neural Information Processing Systems (NeurIPS), **2020**.
45. Zalashko, A.; Causal optimal transport: theory and applications. PhD Thesis, **2017** University of Vienna.