universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"On the numerical solution of linearly constrained transport problems"

verfasst von / submitted by

Fabian Pfurtscheller

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 821
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Mathematik
Betreut von / Supervisor:	Ass.-Prof. Dr. Julio Daniel Backhoff-Veraguas

Abstract

In this thesis, we propose a novel algorithm to solve the Causal Optimal Transport problem in N discrete time steps, an example of a linearly constrained transport problem. This algorithm is based on a similar algorithm used to solve the classical Optimal Transport problem. Firstly, the theoretical approximation capabilities of a class of approximative problems based on neural networks are studied. Secondly, further conditions are established, under which theoretical convergence of an approximative problem can be shown. Finally, after a detailed examination of the algorithm, the results of some numerical experiments to solve the Causal Optimal Transport for some commonly occurring examples are presented.

Abstrakt

Diese Masterarbeit stellt einen neuen Algorithmus zur Lösung des kausalen optimalen Transportproblems in N diskreten Zeitschritten, ein Beispiel für ein Transportproblem mit linearen Nebenbedingungen, vor. Dieser Algorithmus basiert auf einem ähnlichen Algorithmus für das klassische optimale Transportproblem. Zunächst wird die theoretische Fähigkeit zur Approximation des Problems einer Klasse an approximativen Problemen basierend auf neuronalen Netzwerken untersucht. Anschließend werden weitere Bedingungen eingeführt, mit Hilfe derer die theoretische Konvergenz eines approximativen Problems gezeigt werden kann. Nach einer detaillierten Untersuchung des Algorithmus werden schließlich die Ergebnisse einiger numerischer Experimente zur Lösung des kausalen optimalen Transportproblems für einige häufig auftretende Beispiele illustriert.

Acknowledgements

I would like to thank my supervisor, Julio Backhoff-Veraguas, for his support and advice during my work on this thesis. I thank my friends and my family for their motivation and support. Most of all, I take this opportunity to thank my parents for their encouragement and unwavering support.

Contents

Chapter 1

Introduction

In 1781's France, Gaspard Monge posed to the Royal Academy of Science the seemingly innocent question on how to most efficiently transport a pile of sand into a different location, [30]. While a price was offered for its solution, it took until the 1940s until a Soviet mathematician, Leonid Kantorovich, reformulated Monge's problem using probability distributions and connecting the problem with his recently developed method of linear programming, which would finally make it possible to answer some questions on existence, uniqueness and characterisation on the most efficient transport, the optimal **transport plan**, in a general fashion, [27].

In the time since, Monge's and Kantorovich's problem of **Optimal Transport** has seen solutions to its main questions - When does a solution exist? Is this solution unique? How can we characterize it? - being found. At the same time, it has entered the domain of many, seemingly unconnected applications, ranging from geometry and analysis of PDEs, [42], probability theory, [11], economy and mathematical finance, [19], [8], [39], to most recently a range of applications in the data sciences, image processing and machine learning, [31].

Concerning the numerical investigation of the Optimal Transport problem, multiple schemes are discussed in [31]. For this thesis, we took in particular note of a scheme proposed in [15]. This paper gives a way to numerically solve a number of optimization problems on probability densities, among which the Optimal Transport problem, by reformulating the problem as a simultaneous maximization and minimization of an objective function - a so-called **Min-Max**-problem. The authors then approximate this problem with the help of two competing neural networks. This neural network approximation can then be solved numerically, using an algorithm provided in the paper.

In part connected to some applications, interest has also been cast on the study of Optimal Transport with some additional constraints. As the name of this thesis suggests, we

will closely investigate a particular linear constraint, namely the Causal Optimal Transport in discrete time. Introduced in a more general setting in [29] and in the discrete setting in [5], it studies Optimal Transport with an additional time structure subject to a **causality** constraint. They play a role whenever special interest is given to this time structure, as then, some methods and tools developed for the classical Optimal Transport problem give very unintuitive results.

To state a brief, rough interpretation of this constraint and to keep with Monge's picture of moving piles of sand, solving the Causal Optimal Transport problem in discrete time can be interpreted as finding the most efficient way to model the displacement of sand over N distinct time steps. Interpreting the distribution of the pile of sand at times $t = 1$ to N at the original location as a stochastic process $\{X_t\}_{t=1}^N$ taking values in $\mathbb{R}^2$ and the distribution at the destination as a stochastic process $\{Y_t\}_{t=1}^N$, a transport plan is called **causal**, if at every time step t, the distribution of the pile of sand at the destination, that is, Y_t, given the past distribution of the sand at the origin - $X_1, \dots, X_t$ - is independent of the future distribution of sand at the origin, $X_{t+1}, \dots, X_N$.

An exact and rigorous mathematical formulation, beyond these toy interpretations, of both the standard Optimal Transport problem and its causally constrained version will be given in the next chapter. Despite interest on the subject having started only in the last decade, Causal Optimal Transport has seen applications in mathematical finance, [3], image processing, [45], and stochastic optimization, [2].

The Causal Optimal Transport problem in discrete time does not fall under the class of problems for which approximation by neural networks was shown in [15]. However, we will show in this thesis that by using similar tools and arguments, an approximation by neural networks also exists for this constrained Optimal Transport problem, and, under some extra conditions, converges to the Causal Optimal Transport problem. Further, we adapted the algorithm proposed in [15] in such a way that it can be used to numerically solve the Causal Optimal Transport problem in discrete time. This thesis is among the first works that study a numerical solution scheme to this problem. Independently and at the same time as the writing of this thesis, a related problem using similar techniques, also basing their method on neural network approximations, was proposed in [24]. In contrast to [24], we do not view the problem from the angle of applications in mathematical finance and do not include any considerations from the point of view of learning theory or any quantitative statements on convergence rates. Instead, we more generally describe the approximation process, give only theoretical statements about approximation, and then more closely inspect the chosen algorithm.

The thesis is structured as follows. In the second chapter, we will investigate in detail some background on the theory of Optimal Transport, name some particularly important results and concepts, and finally closely inspect the constrained problem we are interested in solving numerically, the Causal Optimal Transport problem in discrete time. In Chapter 3, we take note of the reformulation for the Optimal Transport problem proposed in [15], and show that a similar reformulation can be done for the Causal Optimal Transport problem. In the fourth chapter, we will give an introduction to neural networks, the space of functions on which we will later construct a numerical scheme. In the fifth chapter, we show that indeed, we can approximate the Causal Optimal Transport problem with neural network functions, and when moving to a slight relaxation, find approximative solutions using neural networks. The algorithm we use to do so will be described in Chapter 6, while the last chapter is dedicated to present the results of some numerical experiments.

Chapter 2

Transport problems

In this chapter, we will give a brief overview over the field of Optimal Transport, attempt to motivate the problem and investigate it from an angle helpful for the later numerical considerations. The second part of the chapter then enters the domain of linearly constrained Optimal Transport in dealing with the concept of Causal Optimal Transport. Principally, this chapter relies on the two monographs of Villani, [43] and [42], and the book of Santambrogio, [39]. For the computational point of view, the book of Peyré and Cuturi [31], has been helpful. The section regarding the causally constrained problem is then based on [5] and [4].

2.1 Optimal Transport

2.1.1 Monge problem

In the original formulation of Monge,[30], and translated into modern mathematical language by [39], the Optimal Transport problem is stated as follows:

Problem 2.1 (Optimal transport of a pile of sand)**.** Given two mass densities f, g on $\mathbb{R}^d$, that is, measurable, non-negative functions f and g such that $\int f(x)\,dx = \int g(y)\,dy = 1$, we want to find a map T pushing f into g, meaning that for all Borel sets A

$$\int_A g(y)\,dy = \int_{T^{-1}(A)} f(x)dx, \tag{2.1}$$

that simultaneously minimizes the quantity

$$\int_{\mathbb{R}^d} |T(x) - x| f(x)\,dx. \tag{2.2}$$

In Monge's description, $f(x)$ was the starting density of the pile of sand, $g(y)$ the final destination, and the function T represented the movement of each sand particle, which was to be chosen in an optimal way, that is, preferring smaller distances over larger distances. This constraint is modelled by the function $(x, y) \mapsto |x - y|$ representing the **cost** of movement.

Monge's original formulation can be generalized into a broader setting, for this we follow [39]. We take $\mathcal{X}$ and $\mathcal{Y}$ to be general completely metrizable and separable spaces - known as **Polish spaces** - and allow arbitrary cost functions $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$, only requiring them to be measurable. Further, instead of specifying the distributions via the densities f and g, we do so by measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, where $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ are the sets of probability measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Note that unlike in the first formulation, these are not necessarily absolutely continuous with regard to the Lebesgue measure. Finally, for a Borel-measurable map $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$, where $\mathcal{B}(\mathcal{X}, \mathcal{Y})$ is the set of all Borel-measurable maps from $\mathcal{X}$ to $\mathcal{Y}$, and a probability measure μ, we denote for measurable $A \subset \mathcal{Y}$ by

$$T_{\#}\mu(A) := \mu(T^{-1}(A))$$

the **push-forward** of μ under T. We call such a map T a **transport map**. Then, Monge's Optimal Transport problem reads as follows:

$$\inf_{T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})} \int c(x, T(x))\, \mu(dx), \text{ such that } T_{\#}\mu = \nu. \tag{MP}$$

In general, solving the problem (MP) is not easy, and may not always be be possible. Consider the following example, taken from [39]:

Example 2.2. Take $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $\mu = \delta_0$. Then, for every T measurable,

$$T_{\#}\mu = \delta_{T(0)}.$$

But if $\nu \neq \delta_a$ for all $a \in \mathbb{R}$, then no solution to (MP) exists.

Even in more regular cases, such as when a Lebesgue-density for the measures μ and ν exists, the problem is not easy to solve. Denoting by λ the Lebesgue measure and using a change of variables, we end up with a partial differential equation for T, the **Monge-Ampere-Equation**, see [43] for details:

$$\left(\frac{d\nu}{d\lambda} \circ T\right)(x) |\det(\nabla T)| = \frac{d\mu}{d\lambda}(x).$$

This PDE is non-linear in T, presenting substantial difficulties in finding solutions.

Furthermore, the problem does not behave well with some very natural notions of convergence, such as weak convergence on a Hilbert space. We say a sequence f_n in some Hilbert space $\mathcal{H}$ with a scalar product $(\cdot, \cdot)$ converges weakly to $f \in \mathcal{H}$ if for all $g \in \mathcal{H}$, $(f_n, g) \to (f, g)$. Even if we have weak convergence of a sequence of candidate minimizers $T_n \in \mathcal{L}^2(\lambda)$ equipped with the scalar product $(f, g) := \int f g d\lambda$, to a function T, there is no guarantee that through this convergence, the push-forward constraint is preserved. Consider the following example from [7].

Example 2.3. Let $\mathcal{X} = [0,1], \mathcal{Y} = [-1,1]$ and $\mu = \lambda_{[0,1]}, \nu = \frac{1}{2}(\delta_{-1} + \delta_1)$. We consider the function $t : [0,1] \to [-1,1]$ mapping the interval $[0, \frac{1}{2})$ to -1 and $[\frac{1}{2}, 1]$ to 1. We extend t periodically to $\mathbb{R}$ and set

$$T^{(n)}(x) := t(nx).$$

Then, for all n, $T^{(n)}_{\#}\mu = \nu$. We now observe that $T^{(n)}$ converges weakly to $T \equiv 0$. Indeed, it is enough to consider functions $g \in \mathcal{C}([0,1])$ and $n = 2^k$, for $k \in \mathbb{N}$, as those sets are dense in L^2 and $\mathbb{N}$, respectively. Then,

$$\begin{aligned} I^k := \int_{[0,1]} T^{(2^k)}(x) g(x)\, dx &= \int_{A_k} T^{(2^k)}(x) g(x)\, dx + \int_{B_k} T^{(2^k)}(x) g(x)\, dx \\ &= \int_{A_k} g(x)\, dx - \int_{B_k} g(x)\, dx, \end{aligned}$$

where we divide the interval $[0,1]$ according to the k-th digit of the binary expansion, so that for $x = \sum_{l \in \mathbb{N}} a_l 2^{-l}$ with $a_l \in \{0,1\}$, we have $x \in A_k \Leftrightarrow a_k = 0$, and $B_k = [0,1] \setminus A_k$. As g is continuous, it is in particular uniformly continuous on the compact interval $[0,1]$ so for all $\varepsilon > 0$, there exists a $K \in \mathbb{N}$ such that if $x = \sum_{l \in \mathbb{N}} a_l 2^{-l}$ and $\bar{x} = \sum_{l \in \mathbb{N}} \bar{a}_l 2^{-l}$ with $a_l = \bar{a}_l \ \forall l < K$, then $|g(x) - g(\bar{x})| < \varepsilon$. For this reason, $|I^k| \leq \varepsilon$. We conclude that $\lim_{k \to \infty} I^k = \int_{[0,1]} Tg(x)\, dx = 0$, so $T \equiv 0$. But, $T_{\#}\mu = \delta_0 \neq \nu$, so while for every n, $T^{(n)}$ pushes μ to ν, this does not hold true for the weak limit.

2.1.2 Kantorovich problem

Noting these difficulties in solving problem (MP), we move to a relaxation of Monge's problem first published by the Soviet mathematician Leonid Kantorovich in 1942, [27]. In order to inspect this, we first need to consider the following definition.

Definition 2.4 (Coupling)**.** Given two Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, two probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, and a probability measure $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ we denote by $\gamma_{1,2}$ the first or second marginal of γ, respectively, that is

$$\gamma_1 := (p_1)_{\#}\gamma,$$

where (p_1) is the projection from $(\mathcal{X} \times \mathcal{Y})$ to $\mathcal{X}$, and respectively for the second marginal,

$$\gamma_2 := (p_2)_{\#}\gamma,$$

with (p_2) the projection from $(\mathcal{X} \times \mathcal{Y})$ to $\mathcal{Y}$. Then,

$$\Pi(\mu, \nu) := \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \gamma_1 = \mu, \gamma_2 = \nu\}$$

is called the set of **couplings** or **transport plans** between μ and ν.

This allows us to formulate Kantorovich's version of the Optimal Transport problem. Given two Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, two measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, and a cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$, the Kantorovich problem is defined as

$$\inf_{\gamma} \int_{\mathcal{X}\times\mathcal{Y}} c \, d\gamma, \text{ such that } \gamma \in \Pi(\mu, \nu). \tag{KP}$$

We immediately note that any transport map gives rise to a transport plan. In fact, take any transport map T mapping μ to ν, that is, T measurable such that $T_{\#}\mu = \nu$. Then, $\gamma := (\mathrm{id}, T)_{\#}\mu \in \Pi(\mu, \nu)$. On the other hand, if $(\mathrm{id}, T)_{\#}\mu \in \Pi(\mu, \nu)$, then also $T_{\#}\mu = \nu$.

Furthermore, we note further additional benefits of the formulation (KP) over (MP). Firstly, as the **product coupling** $\mu \otimes \nu \in \Pi(\mu, \nu)$, the situation of Example 2.2 cannot occur, where the set over which to minimize was potentially empty. Secondly, we note that the function $\gamma \mapsto \int c \, d\gamma$ is linear in γ, quite in contrast to the Monge problem, which was highly non-linear in T.

Finally, given some constraints on the cost function c it is comparatively easy to show results on existence of minimizers, as the set $\Pi(\mu, \nu)$ is both convex and compact. Convexity is immediate. Take $\gamma^{(1)} \neq \gamma^{(2)} \in \Pi(\mu, \nu)$ and $\rho \in [0, 1]$. Then, for any $A \subset \mathcal{X}$,

$$\begin{aligned}\left(\rho\gamma^{(1)} + (1-\rho)\gamma^{(2)}\right)(A \times \mathcal{Y}) &= \rho\gamma^{(1)}(A \times \mathcal{Y}) + (1-\rho)\gamma^{(2)}(A \times \mathcal{Y}) \\ &= \rho\mu(A) + (1-\rho)\mu(A) = \mu(A),\end{aligned}$$

and similarly for $\mathcal{X} \times B$ for any $B \subset \mathcal{Y}$, so any convex combination of elements in $\Pi(\mu, \nu)$ has the correct marginals, and is itself in $\Pi(\mu, \nu)$. Compactness is a consequence of Prokhorov's theorem, which we will state below. First, let us introduce a topology on the space of measures $\mathcal{P}(\mathcal{Z})$, for any Polish space $\mathcal{Z}$.

Definition 2.5 (Weak convergence)**.** We say measures $\gamma_n \in \mathcal{P}(\mathcal{Z})$ converge **weakly** to a measure $\gamma \in \mathcal{P}(\mathcal{Z})$, if for all functions $f \in \mathcal{C}_b(\mathcal{Z})$, the set of continuous and bounded functions from $\mathcal{Z}$ to $\mathbb{R}$, we have

$$\int f \, d\gamma_n \to \int f \, d\gamma.$$

We call the topology induced by this convergence the **weak topology** on $\mathcal{P}(\mathcal{Z})$.

Theorem 2.6 (Prokhorov's theorem)**.** *Let $\mathcal{X}$ be a Polish space. A family $\Gamma \subset \mathcal{P}(\mathcal{X})$ is relatively compact with regard to the weak topology iff it is tight, that is, if for every $\varepsilon > 0$ there exists a compact $K_\varepsilon \subset \mathcal{X}$ such that*

$$\sup_{\gamma \in \Gamma} \gamma(\mathcal{X} \setminus K_\varepsilon) < \varepsilon.$$

A proof can be found in [9].
We now follow [39] in concluding the following fact.

Corollary 2.7. *The set $\Pi(\mu, \nu)$ is compact.*

Proof. By Theorem 2.6, we first need to show that $\Pi(\mu, \nu)$ is tight. Let $\gamma \in \Pi(\mu, \nu)$ and $\varepsilon > 0$ be given. Now, the singleton sets $\{\mu\}$ and $\{\nu\}$ are tight as these are Borel measures on a Polish space, so we can find $K_1 \subset \mathcal{X}$, $K_2 \subset \mathcal{Y}$ compact such that $\mu(\mathcal{X} \setminus K_1) < \varepsilon$, and $\nu(\mathcal{Y} \setminus K_2) < \varepsilon$. But $K_1 \times K_2$ is compact in $\mathcal{X} \times \mathcal{Y}$, so as

$$\gamma(\mathcal{X} \times \mathcal{Y} \setminus K_1 \times K_2) \leq \gamma((\mathcal{X} \setminus K_1) \times \mathcal{Y}) + \gamma(\mathcal{X} \times (\mathcal{Y} \setminus K_2)) = \mu(\mathcal{X} \setminus K_1) + \nu(\mathcal{Y} \setminus K_2) \leq 2\varepsilon,$$

$\Pi(\mu, \nu)$ is tight. By Theorem 2.6, it is thus relatively compact. Further, we argue why $\Pi(\mu, \nu)$ is closed in the weak topology. We take a sequence $\gamma_n \in \Pi(\mu, \nu)$ such that $\gamma_n \to \gamma$ weakly, for some $\gamma \in \mathcal{P}(\mathcal{X}, \mathcal{Y})$. We want to show that also $\gamma \in \Pi(\mu, \nu)$, i.e., that γ has the correct marginals, μ and ν. We thus take a function $f \in \mathcal{C}_b(\mathcal{X})$ and write $F(x, y) := f(x)$ for the interpretation of f as a function in the space $\mathcal{C}_b(\mathcal{X} \times \mathcal{Y})$. Then,

$$\int f \, d\gamma = \int F \, d\gamma = \lim_{n\to\infty} \int F \, d\gamma_n = \lim_{n\to\infty} \int f \, d\gamma_n = \int f \, d\mu,$$

so as f was arbitrary, $\gamma_1 = \mu$. The same argument for the second coordinate gives us that $\Pi(\mu, \nu)$ is also closed, and thus compact. □

We now introduce the following definition, giving us a constraint on the cost function c that will allow us to show existence of optimizers.

Definition 2.8 (Lower semi-continuity)**.** A function $f : \mathcal{Z} \to [0, \infty]$ is called **lower semi-continuous** if for all $C \geq 0$, the set $\{z \in \mathcal{Z} : f(z) \leq C\}$ is closed in the topology of $\mathcal{Z}$. If $\mathcal{Z}$ is equipped with a metrizable topology - for instance, if $\mathcal{Z}$ is a Polish space - then this is equivalent to

$$\liminf_{z_n \to z} f(z_n) \geq f(z).$$

Lastly, we will need the following theorem, for a proof of which we refer to [39].

Theorem 2.9 (Weierstrass criterion for the existence of optimizers)**.** *If $f : \mathcal{Z} \to [0, \infty]$ is lower-semi continuous and if $\mathcal{Z}$ is compact, then there exists an $\bar{z} \in \mathcal{Z}$ such that $f(\bar{z}) = \min_{z \in \mathcal{Z}} f(z)$.*

Putting all this together guarantees us the existence of an optimizer for (KP), provided our cost function c is lower semi-continuous. We show this following a proof from [7].

Corollary 2.10 (Existence of optimizers for (KP))**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces and $c : (\mathcal{X} \times \mathcal{Y}) \to [0, \infty]$ be lower semi-continuous. Then there exists an optimizer γ^* for (KP), that is, there exists a $\gamma^* \in \Pi(\mu, \nu)$ such that $\int c \, d\gamma^* \leq \int c \, d\gamma$ for all $\gamma \in \Pi(\mu, \nu)$.*

Proof. The only thing that remains to be shown is that if c is lower semi-continuous, the map $\gamma \mapsto \int c\, d\gamma$ is, too. Let us denote by d a compatible metric on the Polish space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We consider a family of continuous and bounded functions c_n increasing to c. For instance, we take $c_n(z) := (\inf_{\bar{z}}(c(\bar{z}) + nd(z, \bar{z}))) \wedge n$, then $c_n \in \mathcal{C}_b(\mathcal{Z}), c_n \leq c$ and $c \leq \liminf c_n$, so that $c = \sup_n c_n$. Now, if $\gamma_n \to \gamma$ weakly, we have

$$\liminf_{n\to\infty} \int c\, d\gamma_n \geq \sup_k \liminf_n \int c_k\, d\gamma_n = \sup_k \int c_k\, d\gamma = \int \sup_k c_k\, d\gamma = \int c\, d\gamma,$$

so $\gamma \mapsto \int c\, d\gamma$ is lower semi-continuous with respect to the weak topology. As $\Pi(\mu, \nu)$ is compact by Corollary 2.7, by Theorem 2.9 we have that there exists an optimizer γ^*. □

Finally, let us argue why it makes sense to consider both problems (MP) and (KP) under the same umbrella of Optimal Transport problems. Following [43], we say (KP) is the *relaxation* of (MP). This means that we extend the class of objects on which we take the infimum by embedding the smaller class - the set of transport maps in (MP) - into a larger class - the set of transport plans in (KP). In this case, this is given by the embedding $\mathcal{I}$

$$\mathcal{I} : T \mapsto [x \mapsto \delta_{T(x)}],$$

mapping from the space of functions $T : \mathcal{X} \mapsto \mathcal{Y}$ to the space of measurable functions from $\mathcal{X}$ to $\mathcal{P}(\mathcal{Y})$. In this way, the nonlinear constraints in (MP) turn into the linearly constrained problem (KP). Like this, we increase the number of cases for which solutions exist, while we already noted that all solutions to (MP) extend to a solution to (KP).

2.1.3 Duality

In the last subsection we saw that the problem (KP) has some nice properties - we try to minimize a linear functional over a convex and compact set. As such, it makes sense to consider a dual problem that we will briefly motivate in the following. The formulation of the dual problem will be particularly important for the numerical considerations in later chapters. This subsection follows for the most part [43].

Let us first try to rephrase the constraint $\gamma \in \Pi(\mu, \nu)$. We fix two Polish spaces $\mathcal{X}$ and Y and note that $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ if and only if

$$\int_{\mathcal{X}\times\mathcal{Y}} (\varphi(x) + \psi(y))\, \gamma(dx, dy) = \int_{\mathcal{X}} \varphi(x)\, \mu(dx) + \int_{\mathcal{Y}} \psi(y)\, \nu(dy) \tag{2.3}$$

for all $(\varphi, \psi) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y})$.

Before we use this to motivate the dual problem, let us briefly note that Equation (2.3) is important also from a numerical point of view, as it enables us to replace the constraint $\gamma \in \Pi(\mu, \nu)$, which is it not at all obvious how to check numerically, with a

family of unconstrained linear optimization problems for a - reasonably nice - family of test functions. Equation (2.3) and similar equalities for constrained Optimal Transport problems will thus play an important role in later chapters.

We are now ready to state the dual problem for (KP).

Theorem 2.11. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces and $c : (\mathcal{X} \times \mathcal{Y}) \to [0, \infty]$ be lower semi-continuous. Then,*

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c \, d\gamma = \sup_{\substack{\varphi \in \mathcal{C}_b(\mathcal{X}), \psi \in \mathcal{C}_b(\mathcal{Y}) \\ \varphi(x)+\psi(y) \leq c(x,y)}} \int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{Y}} \psi \, d\nu. \tag{2.4}$$

For two different rigorous proofs of the above, we consult [43] or [39]. We will only give a motivation adapted from [39].

Let us denote by $\mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ the set of positive measures on $\mathcal{X} \times \mathcal{Y}$. Then, taking $\gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ and $\varphi \in \mathcal{C}_b(\mathcal{X}), \psi \in \mathcal{C}_b(\mathcal{Y})$, we see that

$$\sup_{\varphi,\psi} \int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{Y}} \psi \, d\nu - \int_{\mathcal{X}\times\mathcal{Y}} (\varphi(x) + \psi(y)) \, d\gamma = \begin{cases} 0 & \text{if } \gamma \in \Pi(\mu,\nu) \\ \infty & \text{else} \end{cases} \tag{2.5}$$

If we thus add (2.5) to (KP) we can discard the constraint $\gamma \in \Pi(\mu, \nu)$ and observe

$$(\mathrm{KP}) = \inf_{\gamma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c \, d\gamma + \sup_{\varphi,\psi} \int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{Y}} \psi \, d\nu - \int_{\mathcal{X}\times\mathcal{Y}} (\varphi(x) + \psi(y)) \, d\gamma. \tag{2.6}$$

Now, we would like to interchange infimum and supremum, relying on some kind of **minimax-theorem**. In the case above, restricting c to be lower semi-continuous, we are allowed to so. For this, we consult [39], following arguments from a theorem from [34]. Having reassured ourselves, we proceed

$$(\mathrm{KP}) = \sup_{\varphi,\psi} \int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{Y}} \psi \, d\nu + \inf_{\gamma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) - (\varphi(x) + \psi(y)) \, d\gamma, \tag{2.7}$$

and note, that in a similar fashion to (2.5) we can re-write the constraint on the functions φ, ψ by noting

$$\inf_{\gamma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) - (\varphi(x) + \psi(y)) \, d\gamma = \begin{cases} 0 & \text{if } \varphi(x) + \psi(y) \leq c(x,y) \\ -\infty & \text{else} \end{cases} \tag{2.8}$$

Now, using (2.8) to constrain the set of functions to maximize in (2.7) finally leads to

$$(\mathrm{KP}) = \sup_{\substack{\varphi \in \mathcal{C}_b(\mathcal{X}), \psi \in \mathcal{C}_b(\mathcal{Y}) \\ \varphi(x)+\psi(y) \leq c(x,y)}} \int_{\mathcal{X}} \varphi \, d\mu + \int_{\mathcal{Y}} \psi \, d\nu, \tag{2.9}$$

which is just (2.4). Let us then finally denote the dual Kantorovich problem by

$$\sup_{\substack{\varphi\in\mathcal{C}_b(\mathcal{X}),\psi\in\mathcal{C}_b(\mathcal{Y})\\ \varphi(x)+\psi(y)\leq c(x,y)}} \int_{\mathcal{X}} \varphi\, d\mu + \int_{\mathcal{Y}} \psi\, d\nu. \tag{DP}$$

An important consequence of the equivalence of the primal and dual problems that we will need often in the following chapters is this next proposition.

Proposition 2.12 (Kanotorovich-Rubinstein formula)**.** *Let $\mathcal{X} = \mathcal{Y}$ be Polish spaces and $c(x, y)$ a distance. Then,*

$$\inf_{\gamma\in\Pi(\mu,\nu)} \int c\, d\gamma = \sup_{\varphi\ 1-Lipschitz} \int \varphi\, d(\mu - \nu).$$

For a proof of this we will refer to [43].

2.1.4 Wasserstein distances

The theory we have discussed so far can further be used to define a topology on the space of probability measures, which we will do following [43]. Let $\mathcal{X}$ be a Polish space and d a distance on $\mathcal{X}$. For $p \geq 1$, we then denote the set of probability measures on $\mathcal{X}$ with finite p-th moment as

$$\mathcal{P}_p(\mathcal{X}) := \{\mu \in \mathcal{P}(\mathcal{X}) : \int d^p(x, x_0)\, \mu(dx) < \infty, \text{ for any } x_0 \in \mathcal{X}\}.$$

On this space, we define the following distance:

Definition 2.13 (Wasserstein distance)**.** For $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, we define

$$\mathcal{W}_p(\mu, \nu) := \left(\inf_{\gamma\in\Pi(\mu,\nu)} \int d^p(x, y)\, \gamma(dx, dy) \right)^{\frac{1}{p}}. \tag{2.10}$$

We refer to [43] to reassure ourselves that $\mathcal{W}_p$ as defined above is in fact a distance on $\mathcal{P}(\mathcal{X})$. We note that when setting $p = 1$, by Proposition 2.12, we have a handy reformulation for (2.10):

$$\mathcal{W}_1(\mu, \nu) = \sup_{\varphi\ \text{1-Lipschitz}} \int \varphi\, d\mu - \int \varphi\, d\nu. \tag{2.11}$$

2.1.5 Probabilistic interpretation

To conclude this section, and also to motivate what we will investigate in the next section, let us briefly examine an interpretation of Optimal Transport and Wasserstein distances, adapted from [39]. We fix again two Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and consider two probability

measures μ and ν on $\mathcal{X}$ and $\mathcal{Y}$ respectively. In this setup, we can interpret μ and ν as the laws of two random variables, X and Y. The Kantorovich problem (KP) then is equivalent to

$$\inf\{\mathbb{E}[\, c(X,Y)\,] : X \sim \mu, Y \sim \nu\},$$

Here, the infimum is taken over the joint law $(X,Y) \sim \gamma$.

The special case of $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $c(x,y) = |x-y|^p$ for some $p \geq 1$, is then just

$$\inf\{(\mathbb{E}[\, |X-Y|^p\,]) : X \sim \mu, Y \sim \nu\},$$

or to rephrase it in measure-theoretic language

$$\inf\{(||X-Y||_{L_p}^p : X \sim \mu, Y \sim \nu\}.$$

In this case, the p-Wasserstein distance $\mathcal{W}_p$ is thus equivalent to the infimum of the L_p distance of all random variables with marginal distributions μ and ν.

Finally, let us illustrate this interpretation with an example, motivated by [4].

Example 2.14. Let $\mathcal{X} = \mathbb{R}^2$ and consider for $\varepsilon > 0$ the random variable with law $\mathbb{P}_\varepsilon = \frac{1}{2}\delta_{(\varepsilon,1)} + \frac{1}{2}\delta_{(-\varepsilon,-1)}$. Then, $\mathbb{P}_\varepsilon$ converges weakly to $\mathbb{P}_0 = \frac{1}{2}\delta_{(0,1)} + \frac{1}{2}\delta_{(0,-1)}$. To show this, we take $f \in \mathcal{C}_b(\mathbb{R}^2)$ arbitrary, and observe

$$\int f(x)\,\mathbb{P}_\varepsilon(dx) - \int f(y)\,\mathbb{P}_0(dy) = \frac{1}{2}\big(f(\varepsilon,1) + f(-\varepsilon,-1)\big) - \frac{1}{2}\big(f(0,1) + f(0,-1)\big) \to 0,$$

as $\varepsilon \to 0$, as f is continuous, so $\mathbb{P}_\varepsilon \to \mathbb{P}_0$ weakly. The same is true for convergence in $\mathcal{W}_1$ distance. We consider the measure $\mathbb{P}^* = \frac{1}{2}\delta_{\big((\varepsilon,1),(0,1)\big)} + \frac{1}{2}\delta_{\big((-\varepsilon,-1),(0,-1)\big)}$, which is associated to the transport map T_ε mapping $(\varepsilon,1)$ to $(0,1)$ and $(-\varepsilon,-1)$ to $(0,-1)$. As $\mathbb{P}^* = (id, T_\varepsilon)_\# \mu$, $\mathbb{P}^* \in \Pi(\mu,\nu)$. Then

$$\mathcal{W}_1(\mathbb{P}_\varepsilon, \mathbb{P}_0) \leq \int |x-y|\,\mathbb{P}^*(dx,dy) = \frac{1}{2}|(\varepsilon,1)-(0,1)| + \frac{1}{2}|(-\varepsilon,-1)-(0,-1)| = \varepsilon,$$

so as $\varepsilon \to 0$, also $\mathcal{W}_1(\mathbb{P}_\varepsilon, \mathbb{P}_0) \to 0$. In fact, $\mathbb{P}^*$ is already the optimal transport plan, so the above inequality is indeed an equality.

2.2 Causal Optimal Transport

In the following section, we will investigate what happens when we view Optimal Transport with an added time component. To pick up the probabilistic interpretation at the end of the last section, instead of modelling the behaviour of the law of random variables, we inspect now the same questions for laws of stochastic processes in N discrete time steps. We will see that in this case, it makes sense to add further constraints on the future behaviour of these processes given the past, which is known as **causality**. This section relies mostly on two papers, [5] and [4].

2.2.1 Motivating example

We will pick up where we left in the last section and continue to investigate a variant of Example 2.14, taken from [4].

Example 2.15. We again consider, for $\varepsilon > 0$ the measure $\mathbb{P}_\varepsilon = \frac{1}{2}\delta_{(\varepsilon,1)} + \frac{1}{2}\delta_{(-\varepsilon,-1)}$. Now, instead of thinking of $\mathbb{P}_\varepsilon$ as the law of a random variable X on $\mathbb{R}^2$, let us instead interpret it as the law of a real-valued stochastic process in two time-steps, $(\{X_t^\varepsilon\}_{t=1}^2) \sim \mathbb{P}_\varepsilon$. Similarly, we consider a process $(\{X_t^0\}_{t=1}^2) \sim \mathbb{P}_0$. Then, for any transport plan $\mathbb{P}$,

$$\begin{aligned}\mathbb{P}(X_1^\varepsilon = \pm\varepsilon) &= \frac{1}{2}\\ \mathbb{P}(X_2^\varepsilon = 1 \mid X_1^\varepsilon = \varepsilon) &= 1\\ \mathbb{P}(X_2^\varepsilon = -1 \mid X_1^\varepsilon = -\varepsilon) &= 1,\end{aligned}$$

so X_2^ε is actually deterministic given X_1^ε. On the other hand, as $\mathbb{P}(X_1^0 = 0) = 1$,

$$\begin{aligned}\mathbb{P}(X_2^0 = 1 \mid X_1^0 = 0) = \mathbb{P}(X_2^0 = 1) &= \frac{1}{2}\\ \mathbb{P}(X_2^0 = -1 \mid X_1^0 = 0) = \mathbb{P}(X_2^0 = -1) &= \frac{1}{2},\end{aligned}$$

so X_2^0 given X_1^0 is not deterministic. Conceptually, the two processes X^ε and X^0 could not be more different - one is random at time 1 and deterministic at time 2, the other just the other way around. However, their Wasserstein distance is very small. Interpreting the map $T_\varepsilon = (T_\varepsilon^1, T_\varepsilon^2)$ with $T_\varepsilon^1(\varepsilon) = T_\varepsilon^1(-\varepsilon) = 0$, and $T_\varepsilon^2 = \mathrm{id}$, it gives rise to the measure $\mathbb{P}^*$ as in Example 2.14, so arguing the same way, we note that $\mathcal{W}_1(\mathbb{P}_\varepsilon, \mathbb{P}_0) = \varepsilon$. Intuitively, that two such different processes should be very similar makes little sense, and gives a hint that the Wasserstein distance is not well-adapted when comparing stochastic processes.

2.2.2 Adapted Monge problem

A key problem in why the Wasserstein distance in the motivating example gives such intuitively bad results is that it - and by extension, the Optimal Transport problem (KP) in general - ignores the structure of time inherent to stochastic processes. We model this structure using the following concept.

Definition 2.16 (Filtration)**.** We call a family of $\sigma-$algebras $\mathcal{F} = (\{\mathcal{F}_t\}_{t=1}^N)$ on $\mathcal{X}$ a **filtration**, if for every $t \leq N$, we have $\mathcal{F}_t \subset \mathcal{F}_{t+1}$.

We view $\mathcal{F}_t$ as the total information known to us up until time t, so the property that the subsets increase in time models that we gain more information as time progresses, while not losing any information we already have.

With this in mind, we will now define an adapted Optimal Transport problem for discrete stochastic processes in N time-steps, following closely [5]. While many of the following results also hold true for a general Polish space set-up as in the Optimal Transport case, we fix $\mathcal{X}$ and $\mathcal{Y}$ as closed subsets of $\mathbb{R}^N$.

We first investigate how a version of (MP) might look like that takes this structure of time into account. We define the following set of transport maps.

Definition 2.17. We say a transport map $T : \mathbb{R}^N \to \mathbb{R}^N$ with $T \in \mathcal{B}(\mathbb{R}^N)$ is **adapted** if

$$T(x_1, \dots, x_N) = \big(T_1(x_1), T_2(x_1, x_2), \dots, T_N(x_1, \dots, x_N)\big),$$

with $T_t : \mathbb{R}^t \to \mathbb{R}$, and $T_t \in \mathcal{B}(\mathbb{R}^t; \mathbb{R})$, for all $t \leq N$.

We note, that the t-th coordinate of $y = (y_1, \dots, y_N) := T(x_1, \dots, x_N)$ is then a function of the first t coordinates of $x = (x_1, \dots, x_N)$. Viewing the coordinates of x and y as time steps, as in our motivating example, y_t depends on all the information up until time t, but no future times. We can thus use this structure to define an adapted version of (MP):

$$\inf_{\substack{T_{\#}\mu=\nu \\ T \text{ adapted}}} \int c(x, T(x))\, \mu(dx). \tag{CMP}$$

Remember from the optimal transport case that the problem as above depending non-linearly on T was very hard or in many cases, impossible to solve. We gained much by moving to the relaxed version (KP). Let us do something similar here.

2.2.3 Adapted Kantorovich problem

In order to define an adapted version of (KP), we first equip our spaces $\mathcal{X}$ and $\mathcal{Y}$ with a particular filtration.

Definition 2.18. For a space $\mathcal{X}$ as above, we define $\mathcal{F}^{\mathcal{X}}$ to be the filtration such that for all $t \leq N$, $\mathcal{F}^{\mathcal{X}}_t$ is the smallest $\sigma-$algebra such that the map $(x_1, \dots, x_N) = x \in \mathcal{X} \mapsto (x_1, \dots, x_t) \in \mathbb{R}^t$ is $\mathcal{F}^{\mathcal{X}}_t-$measurable, that is,

$$\mathcal{F}^{\mathcal{X}}_t = \{B \times \mathbb{R}^{N-t} : B \in \mathcal{B}_{\mathbb{R}^t}\},$$

where $\mathcal{B}_{\mathbb{R}^t}$ is the Borel-σ-algebra on $\mathbb{R}^t$. We say $\mathcal{F}^{\mathcal{X}}$ is the filtration generated by the coordinate processes.

As before, we fix two probability measures μ and ν on $\mathcal{X}$ and $\mathcal{Y}$, respectively. We go on to define:

Definition 2.19. For a measure γ on $\mathcal{X} \times \mathcal{Y}$ and $x \in \mathcal{X}$, we write γ^x for the **conditional distribution** of γ given x. If $X \sim \mu$ and $Y \sim \nu$ are two stochastic processes with joint distribution $(X, Y) \sim \gamma$, then γ^x describes the law of Y given the event $\{X = x\}$.

Similarly, for the marginal distributions we write:

Definition 2.20. For a measure μ on $\mathcal{X} \subset \mathbb{R}^N$, $t \leq N$, and $x_1 \ldots, x_t \in \mathcal{X} \cap \mathbb{R}^t$, we write $\mu^{x_1,\ldots,x_t}$ for the **conditional distribution** of μ given $x_1, \ldots, x_t$. If $(\{X_t\}_{t=0}^N) \sim \mu$ is a stochastic process, then $\mu^{x_1,\ldots,x_t}$ describes the law of $(X_{t+1}, \ldots, X_n)$ given the event $\{X_1 = x_1, \ldots, X_t = x_t\}$.

We can now define a constrained version of the set of couplings.

Definition 2.21. We say a transport plan $\gamma \in \Pi(\mu, \nu)$ is **causal** between μ and ν if for any $t \leq N$ and $B \in \mathcal{F}_t^Y$, the mapping

$$\mathcal{X} \ni x \mapsto \gamma^x(B)$$

is $\mathcal{F}_t^{\mathcal{X}}$-measurable. We denote the set of all causal transport plans by

$$\Pi_c(\mu, \nu).$$

We note that Definition 2.21 means that for any set B depending on the first t coordinates of $\mathcal{Y}$, the mapping $x \in \mathcal{X} \mapsto \gamma^x(B)$, which in general depends on $(x_1, \ldots, x_N)$, in fact only depends on $(x_1, \ldots, x_t)$. Coming back to our probabilistic interpretation, with processes $X \sim \mu$ and $Y \sim \nu$, this means that for any $t \leq N$ and $B = B_1 \times \ldots \times B_t$ with $B_j \in \mathcal{B}_{\mathbb{R}},\ j = 1, \ldots, t$

$$\mathbb{P}(Y_1 \in B_1, \ldots, Y_t \in B_t | X_1, \ldots, X_N) = \mathbb{P}(Y_1 \in B_1, \ldots, Y_t \in B_t | X_1, \ldots, X_t),$$

or equivalently, Y_t is independent of $X_{t+1}, \ldots, X_N$ given $X_1, \ldots, X_t$, for which we write

$$Y_t \perp_{X_1,\ldots,X_t} (X_{t+1}, \ldots, X_N).$$

With this set of causal couplings, we can finally define a constrained version of (KP). Given a cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$, we define

$$\inf_{\gamma} \int_{\mathcal{X}\times\mathcal{Y}} c \, d\gamma, \text{ such that } \gamma \in \Pi_c(\mu, \nu). \tag{CP}$$

We note that in defining (CP) as above, we have preserved one key property of (KP), namely, that the mapping $\gamma \mapsto \int c \, d\gamma$ is linear. We further note that as, upon defining the product coupling $\gamma := \mu \otimes \nu$, for every $x \in \mathcal{X}$, $\gamma^x = \nu$. For every $\mathcal{B} \in \mathcal{F}_t^{\mathcal{Y}}$ and every $t \leq N$, the map $x \mapsto \gamma^x(B) = \nu(B)$ is thus constant in x, so it is clearly $\mathcal{F}_t^{\mathcal{X}}$-measurable. This means $\mu \otimes \nu \in \Pi_c(\mu, \nu)$, so there exists always at least one causal coupling, the product coupling.

We also have a similar relationship between (CMP) and (CP) as in the Optimal Transport case. Similarly, we can embed the set of adapted maps T into the set of causal couplings by defining a coupling

$$\gamma := (\mathrm{id}, T)_{\#}\mu.$$

Then clearly, $\gamma \in \Pi(\mu, \nu)$. But we also note that for any $B \in \mathcal{F}_t^{\mathcal{Y}}$, we have for a fixed $x \in \mathcal{X}$,

$$\gamma^x(B) = \delta_{T(x)}(B) = \delta_{(T_1(x_1),\dots,T_t(x_1,\dots,x_T))}(B),$$

as B is $\mathcal{F}_t^{\mathcal{Y}}$-measurable, so $\gamma^x(B)$ depends only on the first t coordinates of x. But so, by definition of $\mathcal{F}_t^{\mathcal{X}}$, for every $B \in \mathcal{F}_t^{\mathcal{Y}}$, $x \mapsto \gamma^x(B)$ is $\mathcal{F}_t^{\mathcal{X}}$-measurable, which is the causality of γ.

2.2.4 Bi-causality and adapted Wasserstein distances

In the definition of (CP), we note that we defined causality in an asymmetric way, as a one-directional property from μ to ν. We also observe, that a transport plan γ that is causal from μ to ν does not necessarily need to be causal from ν to μ. We refer to [5] for an exact specification of requirements on μ and ν for that to be the case. For us will be enough to define a more symmetric class of adapted couplings.

Definition 2.22. We say a coupling $\gamma \in \Pi(\mu, \nu)$ is **bi-causal** if it is causal from μ to ν, and causal from ν to μ. Denoting by $e(x, y) := (y, x)$, we write for the set of bi-casual couplings

$$\Pi_{bc} = \{\gamma \in \Pi_c(\mu, \nu) \text{ such that } e_{\#}\gamma \in \Pi_c(\nu, \mu)\}.$$

On these set of couplings, we can define a symmetric version of (CP), the bi-causal Optimal Transport problem,

$$\inf_{\gamma} \int_{\mathcal{X}\times\mathcal{Y}} c\, d\gamma, \text{ such that } \gamma \in \Pi_{bc}(\mu, \nu). \tag{BcP}$$

Finally, this allows us to introduce a version of the Wasserstein-distance better adapted to the discussion of stochastic processes or, in general, measures with an inherent time structure.

Definition 2.23 (Adapted Wasserstein distance)**.** For a Polish space $\mathcal{X}$, a distance d on $\mathcal{X}$, $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, we define

$$\mathcal{AW}_p(\mu, \nu) := \left(\inf_{\gamma \in \Pi_{bc}(\mu,\nu)} \int d^p(x, y)\, \gamma(dx, dy) \right)^{\frac{1}{p}}. \tag{2.12}$$

2.2.5 Duality

Similarly to the Optimal Transport case, we can also define a dual problem to (CP). To motivate this dual problem, we will again first consider an equivalent formulation to the causality constraint of γ. This will further be very helpful in later chapters when we are interested in finding a numerical solution to (CP). The results and proofs from this subsection are all taken from [5].

Proposition 2.24. *A transport plan γ is causal from μ to ν iff $\gamma \in \Pi(\mu,\nu)$ and we have, for all $t < N$ and all $f_t \in \mathcal{C}_b(\mathbb{R}^t), g_t \in \mathcal{C}_b(\mathbb{R}^N)$:*

$$\int f_t(y_1,\dots y_t)\big\{g_t(x_1,\dots,x_N)- \int g_t(x_1,\dots,x_t,\bar{x}_{t+1},\dots,\bar{x}_N)\,\mu^{x_1,\dots,x_t}(d\bar{x}_{t+1},\dots,d\bar{x}_N)\big\}\,\gamma(dx,dy)=0. \tag{2.13}$$

Proof. We define $f_t^x(x_1,\dots,x_N) := \int f_t(y_1,\dots,y_t)\,\gamma^{x_1,\dots,x_N}d(y_1,\dots,y_t)$, for $f_t \in \mathcal{C}_b(\mathbb{R}^t)$. Then, we note that by definition, $\gamma \in \Pi_c(\mu,\nu)$ if and only if,

$$f_t^x(x_1,\dots,x_N) = \int f_t^x(x_1,\dots,x_t,\bar{x}_{t+1},\dots,\bar{x}_N)\,\mu^{x_1,\dots,x_t}(d\bar{x}_{t+1},\dots,d\bar{x}_N)$$

for all f_t^x with $f_t \in \mathcal{C}_b(\mathbb{R}^t)$, that is, if the conditional law occuring in the definition of f_t^x depends only on the first t coordinates. As $\mathcal{X}$ is a Polish space, this is equivalent to

$$\int g_t(x_1,\dots,x_N)\big\{f_x(x_1,\dots,x_N)- \int f_t^x(x_1,\dots,x_t,\bar{x}_{t+1},\dots,\bar{x}_N)\,\mu^{x_1,\dots,x_t}(d\bar{x}_{t+1},\dots,d\bar{x}_N)\big\}\,d\mu=0,$$

for all $t < N$ and all $g_t \in \mathcal{C}_b(\mathbb{R}^N)$. By the tower property of conditional expectation, the above is equivalent to

$$\int f_t^x(x_1,\dots,x_N)\big\{g_t(x_1,\dots,x_N)- \int g_t(x_1,\dots,x_t,\bar{x}_{t+1},\dots,\bar{x}_N)\mu^{x_1,\dots,x_t}(d\bar{x}_{t+1},\dots,d\bar{x}_N)\big\}\,d\mu=0.$$

Finally, observing that, again by the tower property,

$$\int f_t^x(x_1,\dots,x_N)\,d\gamma = \int f_t(y_1,\dots,y_N)\,d\gamma,$$

we get that the above is equivalent to

$$\int f_t(y_1,\dots,y_N)\big\{g_t(x_1,\dots,x_N)- \int g_t(x_1,\dots,x_t,\bar{x}_{t+1},\dots,\bar{x}_N)\,\mu^{x_1,\dots,x_t}(d\bar{x}_{t+1},\dots,d\bar{x}_N)\big\}\,d\gamma=0,$$

as desired. □

Note that Proposition 2.24 tells us that the Causal Optimal Transport problem (CP) is actually a direct analogue of the Kantorovich problem in Optimal Transport (KP) under some additional linear constraints - the constraints that for every time step, an integral of the above kind on some test functions evaluates to zero. It seems thus natural to proceed in a similar fashion in order to reach a dual problem to (CP). Through Proposition 2.24, we have found a way to check causality using a sum of functions. In light of this, it is thus well motivated to define the following set of test functions:

$$\mathbb{F} := \left\{ \begin{array}{l} F : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R} \text{ s.t. } F(x_1, \dots, x_N, y_1, \dots, y_N) = \\ \sum_{t<N} f_t(y_1, \dots y_t)\big\{g_t(x_1, \dots, x_N) - \\ \int g_t(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N)\big\}, \\ \text{ with } f_t \in \mathcal{C}_b(\mathbb{R}^t), g_t \in \mathcal{C}_b(\mathbb{R}^N) \text{ for all } t < N \end{array} \right\}$$

Different to the Kantorovich problem (KP), in order to get a similar result for duality, we further need to restrict our transport plans to those where the first marginal μ fulfils the following regularity assumption.

Property 2.25 (Continuity of disintegration)**.** For all $t < N$, the mapping

$$\mathcal{X} \to \mathcal{P}(\mathcal{X}) : (x_1, \dots, x_t) \mapsto \mu^{x_1,\dots,x_t}$$

is continuous with regard to the weak topology in $\mathcal{P}(\mathcal{X})$ and with regard to the relative topology in $\mathcal{X}$.

With this restriction on the measures, there exists a statement about duality and existence of solutions for the Causal Optimal Transport problem (CP).

Theorem 2.26 (Dual formulation for (CP))**.** *Suppose the cost function $c : \mathcal{X} \times Y \to [0, \infty]$ is lower semi-continuous and bounded by below. Further assume the measure μ fulfills Property 2.25. Then,*

$$\inf_{\gamma \in \Pi_c(\mu,\nu)} \int c\, d\gamma = \sup_{\substack{\varphi,\psi \in \mathcal{C}_b(\mathbb{R}^N), F \in \mathbb{F} \\ \varphi(x)+\psi(y) \le c(x,y)+F(x,y)}} \int \varphi\, d\mu + \int \psi\, d\nu. \tag{2.14}$$

Furthermore, the infimum on the left-hand side is attained, i.e. there exists a solution to (CP).

The proof of the above is found in [5].

2.2.6 Motivating example, revisited

Before we move on to the computational point of view, let us briefly revisit the motivating Example 2.15 of the section on Causal Optimal Transport.

Example 2.15 (Continured)**.** We saw that the optimal transport plan between the measures $\mathbb{P}_\varepsilon = \frac{1}{2}\delta_{(\varepsilon,1)} + \frac{1}{2}\delta_{(-\varepsilon,-1)}$ to the measure $\mathbb{P}_0 = \frac{1}{2}\delta_{(0,1)} + \frac{1}{2}\delta_{(0,-1)}$ was the measure

$$\mathbb{P}^* = \frac{1}{2}\delta_{\big((\varepsilon,1),(0,1)\big)} + \frac{1}{2}\delta_{\big((-\varepsilon,-1),(0,-1)\big)}$$

However, we also saw that the Wasserstein distance, based on this measure, gave unintuitive results when the marginal measures $\mathbb{P}_\varepsilon$ and $\mathbb{P}_0$ were interpreted as laws of stochastic processes $X = (\{X_t\}_{t=1}^2)$ and $Y = (\{Y_t\}_{t=1}^2)$, respectively. The definitions of adaptivity and causality introduced in this section can help remedy this apparent imbalance.

We noted, staying in our language of stochastic processes and adapting the definition to the current example, that any transport plan $\mathbb{P}$ is causal from $\mathbb{P}_\varepsilon$ to $\mathbb{P}_0$, if Y_1 is independent of X_2 given X_1. Seeing as in this example, $\mathbb{P}(Y_1 = 0) = 1$, for any transport plan $\mathbb{P}$, this condition is fulfilled automatically, so $\mathbb{P}^*$ is also the solution to

$$\inf_{\gamma\in\Pi_c(\mathbb{P}_\varepsilon,\mathbb{P}_0)} \int |x-y|\ \gamma(dx,dy).$$

However, let us also inspect the problem from the other side, that is, let us investigate if $\mathbb{P}^*$ is also causal from $\mathbb{P}_0$ to $\mathbb{P}_\varepsilon$. Concretely, we need to check that X_1 is independent of Y_2 given Y_1, so that

$$\mathbb{P}^*(X_1 = x \mid Y_1, Y_2 = 1) = \mathbb{P}^*(X_1 = x \mid Y_1, Y_2 = -1), \tag{2.15}$$

for $x \in \{-\varepsilon, \varepsilon\}$. But this is not the case, as

$$\mathbb{P}^*(X_1 = \varepsilon \mid Y_1, Y_2 = 1) = \frac{1}{2}$$

but

$$\mathbb{P}^*(X_1 = \varepsilon \mid Y_1, Y_2 = -1) = 0.$$

However, (2.15) gives us an extra condition to find a transport plan that is causal from $\mathbb{P}_0$ to $\mathbb{P}_\varepsilon$, leading us to find

$$\hat{\mathbb{P}} = \frac{1}{4}\delta_{\big((\varepsilon,1),(0,1)\big)} + \frac{1}{4}\delta_{\big((-\varepsilon,-1),(0,1)\big)} + \frac{1}{4}\delta_{\big((\varepsilon,1),(0,-1)\big)} + \frac{1}{4}\delta_{\big((-\varepsilon,-1),(0,-1)\big)}.$$

In fact, $\hat{\mathbb{P}}$ is the unique, and thus optimal, causal transport plan from $\mathbb{P}_0$ to $\mathbb{P}_\varepsilon$, and thus the solution to

$$\inf_{\gamma\in\Pi_c(\mathbb{P}_\varepsilon,\mathbb{P}_0)} \int |x-y|\ \gamma(dx,dy).$$

As any transport plan is causal from $\mathbb{P}_\varepsilon$ to $\mathbb{P}_0$, it is then also the optimal bi-causal plan. It is thus easy to compute

$$\mathcal{AW}_1(\mathbb{P}_\varepsilon,\mathbb{P}_0) = \int |x-y|\ \hat{\mathbb{P}}(dx,dy) = 1+\varepsilon.$$

Different to the regular Wasserstein distance, as $\varepsilon \to 0$, $\mathcal{AW}_1(\mathbb{P}_\varepsilon,\mathbb{P}_0) \geq 1$, so the discrepancy between these two very different stochastic processes is better preserved.

Chapter 3

Reformulation of the problem

In this chapter, we will attempt to reformulate the two problems discussed in the previous chapter - the Optimal Transport problem (KP) and the Causal Optimal Transport problem (CP) - in a way that facilitates a numerical approximation. The idea behind this comes from a paper from De Gennaro Aquino and Eckstein, [15], where the authors formulated the same ideas and considerations for a class of problems which include in particular the Optimal Transport problem (KP). In this chapter, we will use these ideas as a basis and extend them to the Causal Optimal Transport problem in N time steps, (CP).

3.1 Problem setting

Let $\mathcal{P}(\mathbb{R}^{2\times N})$ be the set of probability measures on $\mathbb{R}^N \times \mathbb{R}^N = \mathbb{R}^{2\times N}$ - modelling for instance the law of a discrete stochastic process $Z = (\{Z_t\}_{t=1}^N)$ with values in $\mathbb{R}^2$ in $N \in \mathbb{N}$ time steps - and $\mathcal{C}_b(\mathbb{R}^{2\times N})$ the space of continuous and bounded functions mapping from $\mathbb{R}^{2\times N}$ to $\mathbb{R}$. We consider $\pi \in \mathcal{P}(\mathbb{R}^{2\times N})$ a probability measure with marginals μ and ν, two probability measures on $\mathbb{R}^N$, a bounded continuous cost function $c \in \mathcal{C}_b(\mathbb{R}^{2\times N})$ and a set of test functions $\mathcal{H} \subset \mathcal{C}_b(\mathbb{R}^{2\times N})$. The two problems (KP) and (CP) can then be summarised, for different sets of test functions, in the following way.

$$\inf_{\gamma\in\mathcal{Q}} \int c\, d\gamma, \tag{P}$$
$$\text{where } \mathcal{Q} := \left\{\gamma \in \mathcal{P}(\mathbb{R}^{2\times N}) : \int h\, d\gamma = \int h\, d\pi \text{ for all } h \in \mathcal{H}.\right\}$$

3.1.1 Optimal Transport

Firstly, we consider the Optimal Transport problem (KP),

$$\inf_{\gamma\in\Pi(\mu,\nu)} \int c\, d\gamma, \tag{KP}$$

where we consider only one time-step and thus $\pi \in \mathcal{P}(\mathbb{R}^2)$ with marginals μ and ν, for example $\pi = \mu \otimes \nu$. We noted in 2.3 that $\gamma \in \Pi(\mu, \nu)$ if and only if

$$\int (h_1(x) + h_2(y))\, \gamma(dx, dy) = \int h_1(x)\, \mu(dx) + \int h_2(y)\, \nu(dy),$$

for all $h_1, h_2 \in \mathcal{C}_b(\mathbb{R})$, so clearly (KP) is equivalent to (P) taking as test functions the set

$$\mathcal{H}_{KP} := \big\{ h \in \mathcal{C}_b(\mathbb{R}^2) : h(x, y) = h_1(x) + h_2(y), h_1, h_2 \in \mathcal{C}_b(\mathbb{R}) \big\}.$$

3.1.2 Causal Optimal Transport

For the Causal Optimal Transport problem in N time-steps (CP),

$$\inf_{\gamma \in \Pi_c(\mu,\nu)} \int c\, d\gamma \tag{CP}$$

we observed in Proposition 2.13 that a transport plan γ is causal from μ to ν iff $\gamma \in \Pi(\mu, \nu)$ and we have, for all $t < N$ and all $f_t \in \mathcal{C}_n(\mathbb{R}^t), g_t \in \mathcal{C}_b(\mathbb{R}^N)$:

$$\int f_t(y_1, \dots y_t) \big\{ g_t(x_1, \dots, x_N) - \int g_t(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N) \big\}\, \gamma(dx, dy) = 0. \tag{3.1}$$

If we then define the set of test functions

$$\begin{aligned} \mathcal{H}_{CP} := \Big\{ & h \in \mathcal{C}_b(\mathbb{R}^{2\times N}) : h(x, y) = h_1(x) + h_2(y) + \sum_{t<N} f_t(y_1, \dots y_t) \\ & \big\{ g_t(x) - \int g_t(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N) \big\}, \\ & h_1, h_2, g_t \in \mathcal{C}_b(\mathbb{R}^N), f_t \in \mathcal{C}_b(\mathbb{R}^t)\ \forall t < N \Big\}, \end{aligned} \tag{3.2}$$

we note that for all $h \in \mathcal{H}_{CP}$, by (3.1)

$$\int h\, d\gamma = \int h\, d\pi$$

if and only if $\pi, \gamma \in \Pi_c(\mu, \nu)$. So (3.2) is our desired set of test functions such that (P) is equal to the Caual Optimal Transport problem (CP). Here too, $\pi = \mu \otimes \nu$ is a possible choice.

3.2 Problem reformulation

The problem (P) as stated above is a minimization problem on measures with additional constraints dependent on a set of test functions. From a numerical point of view, both of these properties are not easy to handle. A main idea in the paper [15] was to reformulate these two facets of the problem (P) in a way that is easier to handle numerically.

Firstly, we would like to remove the constraints dependent on the test functions, that is, we want to state an **unconstrained** minimization problem which is equivalent to (P). For this, we notice that, similar to when we tried to motivate the dual problem for the Optimal Transport problem (KP),

$$\sup_{h\in\mathcal{H}} \int h d\pi - \int h d\gamma = \begin{cases} 0 & \text{if } \gamma \in \mathcal{Q} \\ \infty & \text{else} \end{cases}.$$

We thus observe,

$$(\mathrm{P}) = \inf_{\gamma\in\mathcal{Q}} \int c d\gamma = \inf_{\gamma\in\mathcal{P}(\mathbb{R}^{2\times N})} \sup_{h\in\mathcal{H}} \int c d\gamma - \int h d\gamma + \int h d\pi.$$

We have thus replaced our constrained minimization problem with an unconstrained **Min-Max** problem, that is, an optimization problem which tries to simultaneously minimize and maximize two quantities.

However, in this formulation, while we try to maximize a function out of a given set of functions, we still try to minimize a measure, which is computationally hard. Another key insight from [15] was to rewrite this constraint using push-forward measures.

We consider $D \in \mathbb{N}$ and a probability measure $\theta \in \mathcal{P}(\mathbb{R}^{D\times N})$. Provided for every $\gamma \in \mathcal{P}(\mathbb{R}^{2\times N})$ we can find a map T such that $\gamma = T_{\#}\theta$, then instead of finding an optimizing measure, the task then becomes to find an optimizing function T. This is guaranteed if θ can be reduced to the uniform distribution on the unit interval $(0, 1)$, following an argument from [15].

Indeed, let $\mathcal{U}$ be the uniform distribution on $(0, 1)$ and $S : (0, 1) \mapsto \mathbb{R}^{D\times N}$ a measurable map such that $S_{\#}\mathcal{U} = \theta$. Then, by [16], there exists a bimeasurable bijection $\varphi : (0, 1) \to \mathbb{R}^{2\times N}$ such that we can write any measure $\gamma \in \mathcal{P}(\mathbb{R}^{2\times N})$ as a pushforward of the unit measure, by defining $Q_{\tilde{\gamma}}$ as the quantile function of $\tilde{\gamma} := (\varphi^{-1})_{\#}\gamma$ and writing

$$\gamma = (\varphi \circ Q_{\tilde{\gamma}})_{\#}\mathcal{U}.$$

But then, we can write γ as the pushforward of θ, writing

$$\gamma = (\varphi \circ Q_{\tilde{\gamma}} \circ S^{-1})_{\#}\theta.$$

We thus fix $D \in \mathbb{N}$ and a measure $\theta \in \mathcal{P}(\mathbb{R}^{D\times N})$ with the above property, calling θ a **latent measure**. We further denote by $\mathcal{B}_{D,2}$ the set of Borel-measurable functions mapping from $\mathbb{R}^{D\times N}$ to $\mathbb{R}^{2\times N}$. By the above considerations, we can thus reformulate the

problem (P) to read

$$\begin{aligned}
(\mathrm{P}) &= \inf_{\gamma\in\mathcal{Q}} \int c\, d\gamma \\
&= \inf_{\gamma\in\mathcal{P}(\mathbb{R}^{2\times N})} \sup_{h\in\mathcal{H}} \int c\, d\gamma - \int h\, d\gamma + \int h\, d\pi \\
&= \inf_{T\in\mathcal{B}_{D,2}} \sup_{h\in\mathcal{H}} \int c\, dT_{\#}\theta - \int h\, dT_{\#}\theta + \int h\, d\pi \\
&= \inf_{T\in\mathcal{B}_{D,2}} \sup_{h\in\mathcal{H}} \int c\circ T\, d\theta - \int h\circ T\, d\theta + \int h\, d\pi \\
&= \inf_{T\in\mathcal{B}_{D,2}} \sup_{h\in\mathcal{H}} \Phi(T,h)
\end{aligned}$$

where we define the objective function

$$\Phi(T,h) := \int c\circ T\, d\theta - \int h\circ T\, d\theta + \int h\, d\pi.$$

Starting with a constraint minimization problem on measures, we have obtained an unconstrained MinMax problem on two function spaces. From a computational point of view, we are then interested in how to effectively approximate those two function spaces - on the one hand, the set $\mathcal{H}$, on the other, the set $\mathcal{B}_{D,2}$ of Borel-measurable functions mapping from $\mathbb{R}^{D\times N}$ to $\mathbb{R}^{2\times N}$ - by a space of functions that can be explicitly computed. Our next task is thus to find spaces $\mathcal{B}^m_{D,2}$ and $\mathcal{H}^m$, solve an approximate problem (P^m) to (P) on those spaces and obtain optimizers T^m, h^m such that, in some sense of convergence,

$$\inf_{T^m}\sup_{h^m} \Phi(T^m,h^m) \to \inf_{T}\sup_{h} \Phi(T,h).$$

Chapter 4

Neural networks

In this chapter, we will give a brief overview on the theory of neural networks, which in the following will be used to construct the approximation (P^m) to the reformulated problem (P) as discussed in the previous chapter. We will inspect their general structure and investigate their approximation capabilities.

4.1 Introduction to neural networks

As discussed at the end of the previous chapter, we are interested in finding a space of functions that can both be computed relatively easily while at the same time approximate a large class of functions relatively well. One candidate for this is the set of neural networks, which will be introduced in the following.

We are going to inspect the notion of artificial feed-forward neural networks - also known as multilayer perceptrons - which we will in brief only call **neural networks**. Originally proposed by Rosenblatt in 1958, [35], taking inspiration from the structure of neurons in the human brain, we will use a definition from a more mathematical point of view, taken from [10] and [13]. While neural networks can be defined in a more general setting, for our purposes we will restrict ourselves to the following definition.

Definition 4.1. Let $L, d_1, d_2 \in \mathbb{N}$. A **neural network** F with L **layers**, input dimension d_1 and output dimension d_2 is the composition of L functions $f_l : \mathbb{R}^{n_l} \to \mathbb{R}^{n_{l+1}}$, with $l = 1, \dots, L$,

$$F(x) = (f_L \circ \dots \circ f_1)(x),$$

where for every l, $n_l \in \mathbb{N}$ and $n_1 = d_1, n_{L+1} = d_2$. For every layer $l = 1, \dots, L$ we can explicitly represent the layer function f_l with a matrix of **weights** $W_l \in \mathbb{R}^{n_{l+1} \times n_l}$, a vector of **biases** $b_l \in \mathbb{R}^{n_{l+1}}$ and an activation function $\Psi_l : \mathbb{R}^{n_{l+1}} \to \mathbb{R}^{n_{l+1}}$, as

$$f_l(x_l) = \Psi_l(W_l x_l + b_l),$$

for $x_l \in \mathbb{R}^{n_l}$. We call the dimensions n_l the number of *hidden neurons* of each layer of the neural network and define

$$m := \max_l n_l$$

to be the **hidden dimension** of the network.

In this definition, we understand the activation functions Ψ_l to be applied element-wise. This means, there exists a real-valued function $\sigma_l : \mathbb{R} \to \mathbb{R}$ such that for every $x \in \mathbb{R}^{n_{l+1}}$,

$$\Psi_l(x) = \begin{pmatrix} \sigma_l(x_1) \\ \dots \\ \sigma_l(x_{n_{l+1}}) \end{pmatrix}.$$

With a little abuse of notation, we will often refer to these σ_l as the activation function of the l-th layer.

For simplicity, we will introduce a short notation for the set of neural networks with a fixed number of layers $L \geq 2$ as defined above. As we will see in the following section, the exact number of layers is not important for the theoretical approximation capabilities of the neural network. The choice of number of layers has practical implications, for instance on the rate of convergence for certain types of functions, [18], or on the ability to exactly represent greater classes of functions, [41]. In this and the following theoretical discussions, for a more general outlook, we will however only demand the number of layers to be a fixed number $L \geq 2$, not going into further detail.

Definition 4.2. We denote by $N^m_{d_1,d_2}$ the set of neural networks with input dimension d_1, output dimension d_2, hidden dimension m, a fixed number of layers $L \geq 2$, and where for every layer, the activation function Ψ_l is not a polynomial.

We note that the restriction that Ψ_l is not polynomial gives rise to a huge variety in the set of neural networks. Common, easily computable choices for activation functions that will later be used in the numerical experiments are the following

1. The **ReLU** (rectified linear unit) function

$$\sigma(x) := \max\{0, x\} = (x)_+$$

2. The hyperbolic tangent

$$\sigma(x) := \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

3. The sigmoid function

$$\sigma(x) := \frac{1}{1 + e^{-x}}$$

Specifying the activation functions to be one of the above or similar, we can view a neural network as a series of concatenations of matrix-vector multiplications, vector additions, and evaluations of simple, non-polynomial functions. Computationally, none of these operations are very expensive. A big reason we are using neural networks for our numerical solutions comes from this fact. Much more importantly though, neural networks exhibit great approximation properties, which we will inspect in the following section.

4.2 Approximation properties

Let us begin with specifying the notion of convergence we will use in the following, adapting this definition from [33].

Definition 4.3. We say a sequence of functions $f_n : \mathcal{X} \to \mathbb{R}$ mapping from a metric space $\mathcal{X}$ to $\mathbb{R}$ converges **uniformly on compact sets** to some function f if for every compact set $K \subset \mathcal{X}$, $f_n|_K \to f|_K$ uniformly as $n \to \infty$. Equipping $K \subset \mathcal{X}$ with the supremum norm

$$||f||_{\infty,K} := \sup_{x \in K} |f(x)|,$$

this means that for every $\varepsilon > 0$ we have

$$\lim_{n \to \infty} ||f_n - f||_{\infty,K} < \varepsilon,$$

for every compact subset $K \subset \mathcal{X}$.

In this notion of convergence, there exists a powerful theorem on the approximation capabilities of neural networks, stating that uniformly on compact sets, neural networks can approximate continuous bounded functions arbitrarily well.

Theorem 4.4 (Universal Approximation Theorem)**.** *Let $d, n \in \mathbb{N}$ and write $N_{d,n} = \cup_m N^m_{d,n}$ for the set of neural networks with input dimension d, output dimension n and arbitrary hidden dimension and number of layers $L \geq 2$. Let the activation functions Ψ_l be continuous, non-constant and bounded. Then, for every bounded and continuous $f \in \mathcal{C}_b(\mathbb{R}^d; \mathbb{R}^n)$ there exists a sequence $f_m \in N^m_{d,n}$ such that f_m converges uniformly on compact sets to f.*

There exist multiple different proofs with slightly different assumptions and statements for this result, most important of which by Cybenko, [14], and Hornik, Stinchcombe and White, [26]. We also refer to [32] for an extensive study of different proofs and related statements.

Useful for us later will also be the following corollary to Theorem 4.4.

Corollary 4.5. *Let $d, n \in \mathbb{N}$ and write $N_{d,n} = \cup_m N_{d,n}^m$ for the set of neural networks with input dimension d, output dimension n and arbitrary hidden dimension and number of layers $L \geq 2$. Let the activation functions Ψ_l be continuous, non-constant and bounded. Then, for every $f \in \mathcal{C}(\mathbb{R}^d; \mathbb{R}^n)$ there exists a sequence $f_m \in N_{d,n}^m$ such that for all compact sets $K \subset \mathbb{R}^d$, f_m converges uniformly to f on K. In particular, f_m converges pointwise on $\mathbb{R}^d$ to f.*

Proof. By Theorem 4.4, for every g in $C_b(\mathbb{R}^d; \mathbb{R}^n)$, there exists a sequence of g_m such that $g_m \to g$ converges uniformly on compact sets. To show the same statement for merely continuous functions, let us use a diagonal argument. Let $f \in \mathcal{C}(\mathbb{R}^d; \mathbb{R}^n)$ be a continuous function mapping from $\mathbb{R}^d$ to $\mathbb{R}^n$. We denote, for $k \in \mathbb{N}$, by

$$B_k := \{x \in \mathbb{R}^d : ||x|| \leq k\},$$

the closed balls around 0 with radius k. Note that for any $k \in \mathbb{N}$, $f|_{B_k}$ is bounded and continuous. For every $k \in \mathbb{N}$, by Theorem 4.4 there thus exists a sequence $f_m^{(k)}$ that converges uniformly to $f|_{B_k}$ on all compact sets. Let us then set

$$\tilde{f}_m := f_m^{(m)},$$

for $m \in \mathbb{N}$. Then, for every $k \in \mathbb{N}$, $(\tilde{f}_m)_{m \geq k}$ is a subsequence of $(f_m^k)_{m \in \mathbb{N}}$. But as for every k, f_m^k converges uniformly on compact sets to $f|_{B_k}$ on B_k, so does $\tilde{f}_m$ to f on $\cup_{k \in \mathbb{N}} B_k = \mathbb{R}^d$. Secondly, as for every $x \in \mathbb{R}^d$, by the above $\tilde{f}_m$ converges uniformly to f on $\{x\}$, in general, $\tilde{f}_m$ converges pointwise on $\mathbb{R}^d$ to f. □

A similar result exists further also for the space of L^p-functions. We denote the norm on L^p-spaces for a measure μ and $1 \leq p < \infty$ by

$$||f||_{L^p(\mu)} = \left(\int |f(x)|^p \mu(dx) \right)^{1/p}$$

and say $f \in L^p(\mu)$ if $||f||_{L^p(\mu)} < \infty$.

Theorem 4.6 (Universal Approximation Theorem for L^p-functions)**.** *Let $d, n \in \mathbb{N}$, $1 \leq p < \infty$, and μ be a finite measure on $\mathbb{R}^d$. Then, if the activation functions Ψ_l are unbounded, and non-constant, the set $N_{d,n}$ is dense in $L^p(\mu)$ for every finite measure μ, that is, for every $f \in L^p(\mu)$, and every $\varepsilon > 0$ there is a function $g \in N_{d,n}$ such that*

$$||f - g||_{L^p(\mu)} < \varepsilon.$$

For a proof of the above statement, we refer to [25].

Chapter 5

Approximation of the problem

In this chapter, we will more closely investigate the possibility to approximate the functions and function spaces appearing in the reformulated problem (P) with a set of neural network based functions, as introduced in the previous chapter. First, we will show that it is possible to find sets $\mathcal{B}^m_{D,2}$ and $\mathcal{H}^m$, of neural network functions such that the objective function $\Phi(T,h)$ can be approximated for every T and h by neural network functions $T^m \in \mathcal{B}^m_{D,2}$ and $h^m \in \mathcal{H}^m$ such that $\Phi(T^m,h^m) \to \Phi(T,h)$.

5.1 Approximation of the objective function

We will investigate this approximation in the following always for the more specific Causal Optimal Transport case, that is (P) with the set of test functions $\mathcal{H}_{CP}$. A similar approximation for the unconstrained Optimal Transport problem - (P) with test functions $\mathcal{H}_{KP}$ - is introduced and proven to converge in [15]. Furthermore, as shown in Chapter 2, the problem (CP) in fact is based on the problem (KP) under some additional, linear constraints. Proving that the more restricted problem can be approximated then immediately also implies the approximation of the more general problem, by the fact that the additional constraints are automatically fulfilled for the more general problem.

First, we need to define the spaces on which we want to find our approximating functions. We note that the outer infimum is taken over all Borel-measurable functions mapping from $\mathbb{R}^{D\times N}$ to $\mathbb{R}^{2\times N}$. By Luzin's theorem, all Borel-measurable functions are continuous almost everywhere. Thus, with the help of Corollary 4.5 to the Universal Approximation Theorem for continuous functions, we can hope to approximate all such $T \in \mathcal{B}_{D,2}$ by neural networks with sufficiently large hidden dimension m, input dimension $\mathbb{R}^{D\times N}$ and output dimension $\mathbb{R}^{2\times N}$. For ease of notation, we will denote this set of functions by $N^m_{D,2}$.

The inner supremum on the other hand is taken over a more special set, the set of test functions

$$\mathcal{H}_{CP} := \Big\{ h \in \mathcal{C}_b(\mathbb{R}^{2\times N}) : h(x,y) = h_1(x) + h_2(y) + \sum_{t<N} f_t(y_1, \dots y_t) \\ \big\{ g_t(x) - \int g_t(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N) \big\}, \\ h_1, h_2, g_t \in \mathcal{C}_b(\mathbb{R}^N), f_t \in \mathcal{C}_b(\mathbb{R}^t)\ \forall t < N \Big\}. \tag{5.1}$$

Note that every function in $\mathcal{H}_{CP}$ is composed of different continuous and bounded functions. By Theorem 4.4, we know that uniformly on compact sets, we can approximate bounded continuous functions by neural networks. It is perhaps reasonable to assume that by building a similar set as $\mathcal{H}_{CP}$ composed of neural network functions, we could achieve similar approximation results. We thus consider for $m \in \mathbb{N}$:

$$\mathcal{H}^m := \Big\{ h_1(x) + h_2(y) + \sum_{t<N} f_t(y_1, \dots y_t) \big\{ g_t(x) - \\ \int g_t(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N) \big\}, \\ h_1, h_2, g_t \in N^m_{N,1}, f_t \in N^m_{t,1}\ \forall t < N \Big\}.$$

Lemma 5.1. *For every $h \in \mathcal{H}_{CP}$, there exists a sequence $h^m \in \mathcal{H}^m$ such that for all such that for all compact sets $K \subset \mathbb{R}$, f_m converges uniformly to f on $K^{2\times N}$.*

Proof. Consider a function $h \in \mathcal{H}_{CP}$, and write using the above decomposition

$$h = h_1(x) + h_2(y) + \sum_{t<N} f_t(y_1, \dots y_t) \big\{ g_t(x) - \\ \int g_t(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N) \big\},$$

for some h_1, h_2, f_t, g_t. Now, by the Universal Approximation Theorem, there exists for every $\varepsilon > 0$ and functions $h_1^m, h_2^m, f_t^m, g_t^m \in N^m_{d,1}$ such that for all compact subsets $K \subset \mathbb{R}$,

$$\begin{aligned} ||h_1 - h_1^m||_{\infty,K^N} &< \varepsilon, \\ ||h_2 - h_2^m||_{\infty,K^N} &< \varepsilon, \\ ||f_t - f_t^m||_{\infty,K^t} &< \varepsilon, \\ ||g_t - g_t^m||_{\infty,K^N} &< \varepsilon, \end{aligned} \tag{5.2}$$

for all $t < N$ and all m large enough. We will use these functions to define our candidate approximating function $h^m \in \mathcal{H}^m$, defining

$$h^m(x,y) := h_1^m(x) + h_2^m(y) + \sum_{t<N} f_t^m(y_1, \dots y_t) \\ \big\{ g_t^m(x) - \int g_t^m(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N) \big\}.$$

Fix $K \subset \mathbb{R}$, and choose m large enough. Note then, for g_t, g_t^m as above,

$$\begin{aligned}
&|| \int g_t(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N) - \\
&\int g_t^m(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N)||_{\infty,K^t} \\
&\leq \int ||g_t - g_t^m||_{\infty,K^N}\, d\mu^{x_1,\dots,x_t} \\
&\leq ||g_t - g_t^m||_{\infty,K^N} \int d\mu^{x_1,\dots,x_t} < \varepsilon, \qquad (5.3)
\end{aligned}$$

as $d\mu^{x_1,\dots,x_t}$ is a probability measure and thus the integral $\int d\mu^{x_1,\dots,x_t} = 1$. We observe, using abbreviated notation, that

$$\begin{aligned}
||h - h^m||_{\infty,K^{2\times N}} &= ||\big(h_1 + h_2 + \sum_{t<N} f_t\big\{g_t - \int g_t\, d\mu^{x_1,\dots,x_t}\big\}\big) - \\
&\qquad \big(h_1^m + h_2^m + \sum_{t<N} f_t^m\big\{g_t^m - \int g_t^m\, d\mu^{x_1,\dots,x_t}\big\}\big)||_{\infty,K^{2\times N}} \\
&\leq ||\big(h_1 - h_1^m\big)||_{\infty,K^N} + ||\big(h_2 - h_2^m\big)||_{\infty,K^N} + \\
&\qquad \sum_{t<N} ||f_t\big\{g_t - \int g_t\, d\mu^{x_1,\dots,x_t}\big\} - f_t^m\big\{g_t^m - \int g_t^m\, d\mu^{x_1,\dots,x_t}\big\}||_{\infty,K^N}. \qquad (5.4)
\end{aligned}$$

Adding and subtracting a mixed term $f_t\{g_t^m - \int g_t^m\, d\mu^{x_1,\dots,x_t}\}$ and noting that by (5.2), the first two summands are smaller than ε, we have

$$\begin{aligned}
||h - h^m||_{\infty,K^{2\times N}} \leq 2\varepsilon + \sum_{t<N} ||f_t||_{\infty,K^t} \Big(&||g_t - g_t^m||_{\infty,K^N} + \\
&|| \int g_t\, d\mu^{x_1,\dots,x_t} - \int g_t^m\, d\mu^{x_1,\dots,x_t}||_{\infty,K^t}\Big) + \\
&||f_t - f_t^m||_{\infty,K^t} ||g_t^m - \int g_t^m\, d\mu^{x_1,\dots,x_t}||_{\infty,K^N}. \qquad (5.5)
\end{aligned}$$

Now, for all $t < N$, f_t and g_t are bounded functions, so in particular for some constant $C_1 < \infty$, $||f_t||_{\infty,K^t} < C_1$ and $||g_t||_{\infty,K^N} < C_1$ and so by (5.2), also $||g_t^m||_{\infty,K^N} < C_2$, for some $C_2 < \infty$. As the integral of a bounded function on a compact set against a probability distribution is also bounded, we also have

$$||\big(g_t^m - \int g_t^m\, d\mu^{x_1,\dots,x_t}\big)||_{\infty,K^N} < C_3,$$

for some $C_3 < \infty$, so putting everything together and using (5.2) and (5.3) yields

$$||h - h^m||_{\infty,K^{2\times N}} \leq 2\varepsilon + N(2C_1\varepsilon + \varepsilon C_3) \leq \tilde{C}\varepsilon,$$

for some constant $\tilde{C} < \infty$. As $h \in \mathcal{H}_{CP}$ was arbitrary, this gives us the statement. □

We want to use this fact to show that we can approximate the objective function Φ using neural network functions. However, we will run into difficulties if we allow approximation of $h \in \mathcal{H}_{CP}$ by arbitrary sequences $h^m \in \mathcal{H}^m$. While for every m, these functions h^m are bounded as they are finite linear combinations of bounded activation functions, they are not necessarily uniformly bounded. We thus first need to consider a further restricted set of candidate approximation functions. For $m \in \mathbb{N}$, we define

$$\tilde{\mathcal{H}}^m := \Big\{(-c) \vee h^m \wedge (c) : \ h^m \in \mathcal{H}^m, c \in \mathbb{R}\Big\}.$$

We note in Corollary 5.2 that this restricted set is still dense in $\mathcal{H}_{CP}$ in the compact topology. More important, it also allows for approximation by uniformly bounded sequences.

Corollary 5.2. *For every $h \in \mathcal{H}_{CP}$, there exists a sequence $h^m \in \tilde{\mathcal{H}}^m$ such that for all such that for all compact sets $K \subset \mathbb{R}$, f_m converges uniformly to f on $K^{2\times N}$, and for every m, $||h^m||_\infty \leq 2||h||_\infty$.*

Proof. We first note that by Lemma 5.1, for every $h \in \mathcal{H}_{CP}$ there exists a sequence $h^m \in \mathcal{H}^m$ such that h^m converges to h uniformly on compact sets. We define a new sequence

$$\tilde{h}^m := (-2||h||_\infty) \vee h^m \wedge (2||h||_\infty)$$

Then, clearly, for every $m \in \mathbb{N}$, $\tilde{h}^m \in \tilde{\mathcal{H}}^m$ and $||\tilde{h}^m||_\infty \leq 2||h||_\infty$. Take an arbitrary compact subset $K \subset \mathbb{R}^{2\times N}$ and $\varepsilon > 0$. As h^m converges uniformly on compact sets to h, for m large enough we have that

$$(-2||h||_\infty) \leq h^m|_K \leq (2||h||_\infty).$$

But then, $\tilde{h}^m = h^m$ on K, so by Lemma 5.1,

$$||\tilde{h}^m - h||_{\infty,K} < \varepsilon.$$

As K and ε were arbitrary, the statement follows. □

Remark 5.3. We note that restricted on compact sets, the sets $\mathcal{H}^m$ and $\tilde{\mathcal{H}}^m$ are equivalent. Indeed, take $h^m \in \mathcal{H}^m$, and an arbitrary compact set K. Then, $h^m|_K = (-||h||_{\infty,K}) \vee h^m \wedge (||h||_{\infty,K})$, so $h^m|_K \in \tilde{\mathcal{H}}^m$. On the other hand, for every arbitrary $\tilde{h}^m \in \tilde{\mathcal{H}}^m$ we have $\tilde{h}^m \in \mathcal{H}^m$ in general. This will be important later, when we show some results only restricted on compact sets. Further, this is important from the computational point of view, as for this, we are often only interested on results restricted on compact sets.

Secondly, to show a convergence for the objective function, we will need to use Luzin's theorem, a proof of which can be found in [37].

Theorem 5.4 (Luzin's theorem)**.** *Suppose f is a Borel-measurable function mapping from $\mathbb{R}^{D\times N}$ to $\mathbb{R}^{2\times N}$, μ a finite measure on $\mathbb{R}^{D\times N}$ and $\varepsilon > 0$. Then, there exists a compact set $K \subset \mathbb{R}^{D\times N}$ such that*

$$f|_K \in \mathcal{C}(K) \quad \text{and } \mu(K) > 1 - \varepsilon.$$

We are now ready to state the following result.

Proposition 5.5. *For any $T \in \mathcal{B}_{D,2}$ and any $h \in \mathcal{H}_{CP}$ there exist $T^m \in N^m_{D,2}$ and $h^m \in \tilde{\mathcal{H}}^m$ such that $\Phi(T^m, h^m) \to \Phi(T, h)$, as $m \to \infty$.*

Proof. We consider arbitrary functions $T \in \mathcal{B}_{D,2}$ and $h \in \mathcal{H}_{CP}$ as above. As $T \in \mathcal{B}_{D,2}$, by Theorem 5.4, for all $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset \mathbb{R}^{D\times N}$ such that

$$T|_{K_\varepsilon} \in \mathcal{C}(K_\varepsilon) \text{ and } \theta(K_\varepsilon) > 1 - \varepsilon.$$

But as $T|_{K_\varepsilon}$ is continuous, by Corollary 4.5 to the Universal Approximation Theorem for continuous functions, there exists a sequence $T^m \in N^m_{D,2}$ such that for all $x \in K_\varepsilon$,

$$T^m(x) \to T(x),$$

as $m \to \infty$, that is, that the T^m converge pointwise on K_ε to T. We then observe,

$$\begin{aligned}|\Phi(T,h) - \Phi(T^m,h^m)| &= |\int c\circ T - c\circ T^m\, d\theta - \int h\circ T - h^m\circ T^m\, d\theta + \int h - h^m\, d\pi| \\ &\leq |\int c\circ T - c\circ T^m\, d\theta| + |\int h\circ T - h^m\circ T^m\, d\theta| + |\int h - h^m\, d\pi|.\end{aligned}$$

As c is continuous and as $T^m \to T$ pointwise on K_ε, also $c\circ T_m \to c\circ T$ pointwise on K_ε. Now,

$$|\int c\circ T - c\circ T^m\, d\theta| \leq 2||c||_\infty \theta(\mathbb{R}^{D\times N}\setminus K_\varepsilon) + |\int_{K_\varepsilon} c\circ T - c\circ T^m\, d\theta| \leq 2C\varepsilon + \varepsilon,$$

for m large enough, as c is bounded and thus by the Dominated Convergence Theorem, the integrals converge.
For the second and third term, we note that by Corollary 5.2, for every $\varepsilon > 0$ there exists a function $h^m \in \tilde{\mathcal{H}}^m$ such that for every compact subset $K \subset \mathbb{R}$,

$$||h - h^m||_{\infty, K^{2\times N}} < \varepsilon, \tag{5.6}$$

for all m large enough. We consider now a compact set $K \subset \mathbb{R}$ such that $\pi(\mathbb{R}^{2\times N}\setminus K^{2\times N}) < \varepsilon$ and $\theta(\mathbb{R}^{D\times N}\setminus K^{D\times N}) < \varepsilon$. Then,

$$\begin{aligned}|\int h\circ T - h^m\circ T^m\, d\theta| &\leq |\int_{\mathbb{R}^{D\times N}\setminus K^{D\times N}} h\circ T - h^m\circ T^m\, d\theta| + |\int_{K^{D\times N}} h\circ T - h^m\circ T^m\, d\theta| \\ &\leq (||h||_\infty + ||h^m||_\infty)\theta(\mathbb{R}^{D\times N}\setminus K^{D\times N}) + \\ &\quad |\int_{K^{D\times N}} h\circ T - h\circ T^m\, d\theta| + |\int_{K^{D\times N}} h\circ T^m - h^m\circ T^m\, d\theta|.\end{aligned}$$

We note that h and h^m are bounded, by $||h||_\infty < \infty$ and $2||h||_\infty < \infty$, respectively. By by Corollary 4.5 to the Universal Approximation Theorem, T^m converges to T uniformly on compact subsets, so also on $K^{D\times N}$. But as h is bounded, by the Dominated Convergence Theorem, the first integral thus becomes arbitrarily small for all m large enough. Finally, for the third term, we note that

$$|\int_{K^{D\times N}} h\circ T^m - h^m \circ T^m \, d\theta| \leq ||h-h^m||_{\infty, K^{T^m(K^{D\times N})}} \theta(K^{D\times N}).$$

As $T^m \to T$ uniformly on compact sets, we note that for all m large enough,

$$T^m(K^{D\times N}) \subset \mathcal{B}_{2||T(K^{D\times N})||_\infty},$$

that is, the closed ball around the origin with radius $2||T(K^{D\times N})||_\infty$, which is a compact subset of $\mathbb{R}^{2\times N}$. But as h^m converges to h uniformly on compact sets, we finally observe that for m large enough,

$$\begin{aligned} |\int h\circ T - h^m \circ T^m \, d\theta| &\leq 3||\infty||_{\infty,K}\varepsilon + \varepsilon + \varepsilon\theta(K^{D\times N}) \\ &\leq C\varepsilon, \end{aligned}$$

where C is a finite constant.
Similarly,

$$\begin{aligned} |\int h - h^m \, d\pi| &\leq |\int_{\mathbb{R}^{2\times N}\setminus 2^{D\times N}} h - h^m \, d\pi| + |\int_{K^{2\times N}} h - h^m \, d\pi| \\ &\leq (||h||_\infty + ||h^m||_\infty)\pi(\mathbb{R}^{2\times N} \setminus K^{2\times N}) + \pi(K^{2\times N})||h-h^m||_{\infty,K^{2\times N}} \\ &\leq C\varepsilon, \end{aligned}$$

by (5.6), choice of K and boundedness of h and h^m. Putting all this together, and noting that $\varepsilon > 0$ was chosen arbitrarily, this gives us

$$|\Phi(T,h) - \Phi(T^m, h^m)| \to 0,$$

as $m \to \infty$, as desired. □

With this result in mind, we can thus define a similar approximating problem

$$\begin{aligned} (P^m) &:= \inf_{T\in N^m_{D,2}} \sup_{h\in\tilde{\mathcal{H}}^m} \int c\circ T \, d\theta - \int h\circ T \, d\theta + \int h \, d\pi \\ &= \inf_{T\in N^m_{D,2}} \sup_{h\in\tilde{\mathcal{H}}^m} \Phi(T,h). \end{aligned} \tag{5.7}$$

We have proven an approximation result for the objective function, leading us to hope that in a similar fashion, we could show $(P^m) \to (\mathrm{P})$. However, due to the fact that we evaluate an outer infimum and and inner supremum at the same time, this is in general not true. We will illustrate this using the following example, taken from [15].

Example 5.6. Let $\pi \in \mathcal{P}(\mathbb{R}^2)$ be a measure with marginal distributions given by the Lebesgue density $\frac{d\pi_1}{d\lambda}(x) := \mathbb{1}_{(0,1)}(x)\kappa x|\sin(1/x)|$ for some $\kappa > 0$ and $\pi_2 = \pi_1$. Take $\theta = \mathcal{U}_{(0,1)^2}$, the uniform distribution on $(0,1)^2$ and for all layers of the neural network functions, set the activation function σ to be the ReLU. Then, we cannot exactly represent π as a pushforward of θ under some neural network function T^m, $\pi \neq \theta_{T^m}$ for all $T^m \in N^m_{2,2}$, as although the density above can be approximated up to arbitrary degree, it cannot be directly represented by neural network functions in our setting. Next, we note that the function

$$\mathbb{R} \ni x \mapsto a\sigma(x-b) = a(x-b)_+ \in \mathcal{H}^m,$$

for all $a, b \in \mathbb{R}$. The key observation now is that since $\pi \neq \theta_{T^m}$ for all T^m, we can find some $b \in \mathbb{R}$ such that

$$\int (x-b)_+ \, \pi(dx) \neq \int (x-b)_+ \, \theta_{T^m}(dx),$$

which is shown in [8]. But then, we get that for all T^m

$$\sup_{h^m \in \mathcal{H}^m} \int h^m \, d\pi - \int h^m \, d\theta_{T^m} \geq a \left(\int (x-b)_+ \, \pi(dx) - \int (x-b)_+ \, \theta_{T^m}(dx) \right),$$

where we choose b such that the integrals are not equal. But as a is arbitrary, we have that

$$\sup_{h^m \in \mathcal{H}^m} \int h^m \, d\pi - \int h^m \, d\theta_{T^m} = \infty$$

for all T^m. By choosing our cost function $c \equiv 0$, we note that $(\mathrm{P}) = 0$ but by the above considerations, $(P^m) = \infty$.

5.2 Relaxation of constraints

In order to obtain convergence not only of the objective function Φ but of the neural network approximation of the problem (P) in general, we have to move to a relaxation of the problem, similar as we did when moving from the Monge problem (MP) to Kantorovich's formulation (KP). We do so by replacing the set of test functions by a different set, involving only Lipschitz-continuous functions. We denote for some $L > 0$ by $\mathrm{Lip}_L(\mathbb{R}^d)$ the set of L-Lipschitz continuous functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(0) = 0$.

To give a motivation for this, we take note of an idea from [44] for the Optimal Transport case. The idea presented in this paper is to reformulate the constraint $\gamma \in \Pi(\mu, \nu)$ in terms of the Wasserstein-distance. Instead of demanding that $\gamma_1 = \mu$ and $\gamma_2 = \nu$ holds true for both marginals of γ, we specify the condition $\mathcal{W}_1(\mu, \gamma_1) = 0$ and $\mathcal{W}_1(\nu, \gamma_2) = 0$. Relaxing this constraint by adding a Lagrange-multiplier $L > 0$ then results in the problem

$$\inf_{\gamma \in \mathcal{P}(R^2)} \int c \, d\gamma + L \left(\mathcal{W}_1(\gamma_1, \mu) + \mathcal{W}_1(\gamma_2, \nu) \right).$$

By equation (2.12), the Kantorovich-Rubinstein-formula, this is equal to

$$\inf_{T\in\mathcal{B}(\mathbb{R}^2)} \sup_{\substack{h_1\in \mathrm{Lip}_1(\mathbb{R})\\ h_2\in\mathrm{Lip}_1(\mathbb{R})}} \int c\, d\theta_T + L\Big[\Big(\int h_1\, d(\theta_T)_1 - \int h_1\, d\mu\Big) + \Big(\int h_2\, d(\theta_T)_2 - \int h_2 d\,\nu\Big)\Big],$$

or equivalently

$$\inf_{T\in\mathcal{B}(\mathbb{R}^2)} \sup_{\substack{h_1\in \mathrm{Lip}_L(\mathbb{R})\\ h_2\in\mathrm{Lip}_L(\mathbb{R})}} \int c\, d\theta_T - \int (h_1+h_2)\, d\theta_T + \int (h_1+h_2)\, d\pi.$$

The above idea was used to relax the constraints for the problem (P) in the cases studied in [15]. As before, we will try to extend it also to the Causal Optimal Transport case. For this, we define a new set of test functions

$$\begin{aligned}\mathcal{H}_L := \Big\{ h(x,y) =& h_1(x) + h_2(y) + \sum_{t<N} f_t(y_1,\dots y_t)\big\{g_t(x) - \\ & \int g_t(x_1,\dots,x_t,\bar{x}_{t+1},\dots,\bar{x}_N)\,\mu^{x_1,\dots,x_t}(d\bar{x}_{t+1},\dots,d\bar{x}_N)\big\}, \\ & h_1,h_2,g_t\in \mathrm{Lip}_L(\mathbb{R}^N), f_t\in \mathrm{Lip}_L(\mathbb{R}^t)\ \forall t<N\Big\}, \end{aligned} \tag{5.8}$$

for some $L>0$. The only difference to the set $\mathcal{H}_{CP}$ is that instead of bounded continuous functions we now demand the test functions to be composed of L-Lipschitz functions.

Accordingly, we must restrict ourselves in the same fashion for our neural network approximation. Setting

$$\mathrm{Lip}_L^m(\mathbb{R}^d) := N_{d,1}^m \cap \mathrm{Lip}_L(\mathbb{R}^d)$$

leads us to a set of candidate approximating test functions

$$\begin{aligned}\mathcal{H}_L^m := \Big\{ & h_1(x) + h_2(y) + \sum_{t<N} f_t(y_1,\dots y_t)\big\{g_t(x) - \\ & \int g_t(x_1,\dots,x_t,\bar{x_{t+1}},\dots,\bar{x_N})\,\mu^{x_1,\dots,x_t}(d\bar{x_{t+1}},\dots,d\bar{x_N})\big\}, \\ & h_1,h_2,g_t\in \mathrm{Lip}_L^m(\mathbb{R}^N), f_t\in \mathrm{Lip}_L^m(\mathbb{R}^t)\ \forall t<N\Big\}, \end{aligned} \tag{5.9}$$

By [17], the set $\mathrm{Lip}_L^m(\mathbb{R}^d)$ satisfies universal approximation properties for L-Lipschitz functions in $\mathrm{Lip}_L(\mathbb{R}^d)$. We note that we have defined this new set $\mathcal{H}_L^m$ without the boundedness restriction we introduced for $\mathcal{H}^m$ that we needed for arguing that the objective function can be approximated. We do not need this restriction here, as we will in the following show all results only restricted to compact subsets, and we remember from Remark 5.3 that the original set $\mathcal{H}^m$ and the restricted set $\tilde{\mathcal{H}}^m$ were equivalent on compact sets.

We can thus define the relaxed problem

$$(P_L) := \inf_{T\in\mathcal{B}_{K,2}} \sup_{h\in\mathcal{H}_L} \int c\circ T\, d\theta - \int h\circ T\, d\theta + \int h\, d\pi, \tag{5.10}$$

and an approximative problem

$$(P_L^m) := \inf_{T\in N_{K,2}^m} \sup_{h\in\mathcal{H}_L^m} \int c\circ T\, d\theta - \int h\circ T\, d\theta + \int h\, d\pi. \tag{5.11}$$

5.2.1 Approximation of (P_L)

We will now show that the problem (P_L^m) converges to (P_L) as $m \to \infty$, in the case that the given measure π and its first marginal μ are compactly supported and fulfil the following property, taken from [4].

Property 5.7 (Lipschitz-regularity of conditional distributions)**.** Take K a compact subset of $\mathbb{R}$. We say μ has Lipschitz-regular conditional distributions on K^N if for all $t < N$, the mapping

$$\mathbb{R}^t \ni K^t \to \mathcal{P}(K^{N-t}) : (x_1, \dots, x_t) \mapsto \mu^{x_1,\dots,x_t}$$

is Lipschitz continuous, with the space $\mathcal{P}(K^{N-t})$ endowed with the $\mathcal{W}_1$-distance. The above thus says that there exists a constant L such that

$$\mathcal{W}_1(\mu^{x_1,\dots,x_t}, \mu^{y_1,\dots,y_t}) \le L|(x_1, \dots, x_t) - (y_1, \dots, y_t)|, \tag{5.12}$$

for all $x, y \in K^t$.

The above property is fulfilled by a number of distributions. We inspect the following examples from [4], which are listed here without proof.

Example 5.8. 1. Assume μ is the law of a stochastic process $(X_t)_{t=1,\dots,N}$ with the following property

$$X_{t+1} = F_{t+1}(X_1, \dots, X_t, \varepsilon_{t+1}),$$

for all $t < N$, where ε_{t+1} is a real-valued random variable and F_{t+1} an $\mathbb{R}^{t+1}-$valued function such that $(x_1, \dots, x_t) \mapsto F_{t+1}(x_1, \dots, x_t, z)$ is Lipschitz for all z. Then μ fulfils Property 5.7.

2. Assume μ has compact support and a density f with regard to the Lebesgue measure on $\mathbb{R}^N$, f is Lipschitz with constant L and there exists a $\delta > 0$ such that $f \ge \delta$ on the compact support of μ. Then, μ fulfils Property 5.7 with Lipschitz-constant $\frac{2L}{\sqrt{\delta}}$.

3. Any measure μ supported on finitely many points fulfils Property 5.7.

Given a marginal μ fulfilling Property 5.7, we can then state an approximation statement for the problem (P_L).

Proposition 5.9. *Assume μ and ν are supported on K^N for a compact subset $K \subseteq \mathbb{R}$ and μ fulfills Property 5.7 for some Lipschitz-constant $\tilde{L}$. Further, assume all possible maps T are projected to have range $K^{2\times N}$. Then*

$$(P_L^m) \to (P_L) \text{ as } m \to \infty.$$

Proof. The proof follows with some adaptions specific to the Causal Optimal Transport Problem the proof of Theorem 1 from [15]. We are going to divide the proof into two parts and show that for all $\varepsilon > 0$ there exists an $m \in \mathbb{N}$ such that

$$(P_L) \leq (P_L^m) + \varepsilon \tag{a}$$
$$(P_L) \geq (P_L^m) - \varepsilon \tag{b}$$

We further also note that the restriction that all possible maps T have range $K^{2\times N}$ comes from the fact that μ and ν are each supported on K^N, and so any optimal T should not map any mass into $\mathbb{R}^{2\times N} \setminus K^{2\times N}$.

(a) We choose $\tilde{\varepsilon} > 0$ arbitrary and note that, by Theorem 1 from [17], for all compact sets $K \subset \mathbb{R}$ there exists an $m \in \mathbb{N}$ such that any function $f \in \mathrm{Lip}_L(\mathbb{R}^N)$ can be approximated by functions from the set $\mathrm{Lip}_L^m(\mathbb{R}^N)$ up to accuracy $\tilde{\varepsilon}$, uniformly on K^N. Then, we have for all $K \subset \mathbb{R}$, for all T and $i = 1, 2$:

$$\begin{aligned}
&\Big| \sup_{h_i \in \mathrm{Lip}_L} \int h_i \, d\theta_T - \int h_i \, d\pi - \sup_{h_i^m \in \mathrm{Lip}_L^m} \int h_i^m \, d\theta_T - \int h_i^m \, d\pi \Big| \\
&\leq \sup_{\substack{h_i^m \in \mathrm{Lip}_L^m \\ h_i \in \mathrm{Lip}_L}} \Big(||h_i - h_i^m||_{\infty, K^N} \int d\theta_T + ||h_i - h_i^m||_{\infty, K^N} \int d\pi \Big) < 2\tilde{\varepsilon}.
\end{aligned}$$

We further note, for all $K \subset \mathbb{R}$, for all T and for any $t < N$,

$$\begin{aligned}
&\Big| \sup_{f_t, g_t \in \mathrm{Lip}_L} \int f_t \big\{ g_t - \int g_t \, d\mu^{x_1,\dots,x_t} \big\} \, d\theta_T - \int f_t \big\{ g_t - \int g_t \, d\mu^{x_1,\dots,x_t} \big\} \, d\pi \\
&- \sup_{f_t^m, g_t^m \in \mathrm{Lip}_L^m} \int f_t^m \big\{ g_t^m - \int g_t^m \, d\mu^{x_1,\dots,x_t} \big\} \, d\theta_T - \int f_t^m \big\{ g_t^m - \int g_t^m \, d\mu^{x_1,\dots,x_t} \big\} \, d\pi \Big| \\
&\leq \sup_{\substack{f_t^m, g_t^m \in \mathrm{Lip}_L^m \\ f_t, g_t \in \mathrm{Lip}_L}} \Big| \int \big(f_t \big\{ g_t - \int g_t \, d\mu^{x_1,\dots,x_t} \big\} \big) - \big(f_t^m \big\{ g_t^m - \int g_t^m \, d\mu^{x_1,\dots,x_t} \big\} \big) \, d\theta_T \Big| \\
&\qquad + \Big| \int \big(f_t \big\{ g_t - \int g_t \, d\mu^{x_1,\dots,x_t} \big\} \big) - \big(f_t^m \big\{ g_t^m - \int g_t^m \, d\mu^{x_1,\dots,x_t} \big\} \big) \, d\pi \Big| \\
&= \sup_{\substack{f_t^m, g_t^m \in \mathrm{Lip}_L^m \\ f_t, g_t \in \mathrm{Lip}_L}} \Big| \int f_t \big\{ g_t - g_t^m - \big(\int g_t \, d\mu^{x_1,\dots,x_t} - \int g_t^m \, d\mu^{x_1,\dots,x_t} \big) \big\} \\
&\qquad + \big(f_t - f_t^m \big) \big\{ g_t^m - \int g_t^m \, d\mu^{x_1,\dots,x_t} \big\} \, d\theta_T \Big| \\
&\qquad + \Big| \int f_t \big\{ g_t - g_t^m - \big(\int g_t \, d\mu^{x_1,\dots,x_t} - \int g_t^m \, d\mu^{x_1,\dots,x_t} \big) \big\} \\
&\qquad + \big(f_t - f_t^m \big) \big\{ g_t^m - \int g_t^m \, d\mu^{x_1,\dots,x_t} \big\} \, d\pi \Big|,
\end{aligned}$$

where in the last line we have added and subtracted a mixed term $f_t\{g_t^m - \int g_t^m \, d\mu^{x_1,\dots,x_t}\}$. Now, $||f_t||_{\infty,K^t}$ and $||\{g_t^m - \int g_t^m \, d\mu^{x_1,\dots,x_t}\}||_{\infty,K^N}$ are bounded by some constants C_1 and C_2, respectively, where we argue the same way as in the proof of Lemma 5.1. As K and L are fixed, we further note that the constants do not depend on m. Choosing m large enough, we then get that the above is less or equal than

$$\begin{aligned}
\sup_{\substack{f_t^m, g_t^m \in \mathrm{Lip}_L^m \\ f_t, g_t \in \mathrm{Lip}_L}} \Big(& ||f_t||_{\infty,K^t} \big\{ ||g_t - g_t^m||_{\infty,K^N} + ||g_t - g_t^m||_{\infty,K^N} \int d\mu^{x_1,\dots,x_t} \big\} \\
& + ||f_t - f_t^m||_{\infty,K^t} C_2 \Big) \int d\theta_T \\
+ \Big(& ||f_t||_{\infty,K^t} \big\{ ||g_t - g_t^m||_{\infty,K^N} + ||g_t - g_t^m||_{\infty,K^N} \int d\mu^{x_1,\dots,x_t} \big\} \\
& + ||f_t - f_t^m||_{\infty,K^t} C_2 \Big) \int d\pi \\
\leq 2\big(& C_1\{\tilde{\varepsilon} + \tilde{\varepsilon}\} + \tilde{\varepsilon} C_2 \big) \leq C\tilde{\varepsilon},
\end{aligned}$$

for some constant C. Putting all this together gives us for all T :

$$\sup_{h \in \mathcal{H}_L} \int c \circ T \, d\theta - \int h \circ T \, d\theta + \int h \, d\pi \geq \sup_{h \in \mathcal{H}_L^m} \int c \circ T \, d\theta - \int h \circ T \, d\theta + \int h \, d\pi + C\tilde{\varepsilon}.$$

Setting $\varepsilon := \frac{\tilde{\varepsilon}}{C}$ then implies, as $N_{D,2}^m \subset \mathcal{B}_{D,2}$,

$$(P_L) \leq \inf_{T \in N_{D,2}^m} \sup_{h \in \mathcal{H}_L} \Phi(T,h) \leq \inf_{T \in N_{D,2}^m} \sup_{h \in \mathcal{H}_L^m} \Phi(T,h) + \varepsilon = (P_L^m) + \varepsilon,$$

which shows (a).

(b) We choose an optimizer $\hat{\gamma} = \hat{T}_{\#}\theta$ for (P_L). As $\hat{T}$ is restricted to take values only in the compact set $K^{2\times N}$, $\hat{T} \in \mathcal{L}^1(\theta)$, so by the Universal Approximation Theorem for $\mathcal{L}^1$-functions, Theorem 4.6, we can choose a sequence $T^m \in N_{K,2}^m$ with $T^m \to \hat{T}$ in $\mathcal{L}_1(\theta)$ for $m \to \infty$. It follows by [15] that $\gamma^m := T^m_{\#}\theta$ converges weakly to $\hat{\gamma}$. As the measures are supported on the compact set K, this means that

$$\mathcal{W}_1(\gamma^m, \gamma) \to 0 \tag{5.13}$$

We now note, for some $\tilde{\varepsilon} > 0$ and m large enough:

$$\begin{aligned}
& \big| \sup_{h \in \mathcal{H}_L} \Phi(\hat{T}, h) - \sup_{h \in \mathcal{H}_L} \Phi(T^m, h) \big| \leq \big| \int c \, d\gamma - \int c \, d\gamma^m \big| + \sup_{h \in \mathcal{H}_L} \big| \int h \, d\gamma - \int h \, d\gamma^m \big| \\
& \leq \tilde{\varepsilon} + \sup_{h_1, h_2, f_1, g_1, \dots, f_{N-1}, g_{N-1}} \big| \int h_1 \, d\gamma - \int h_1 \, d\gamma^m \big| + \big| \int h_2 \, d\gamma - \int h_2 \, d\gamma^m \big| \\
& + \sum_{t<N} \big| \int f_t \big\{ g_t - \int g_t \, d\mu^{x_1,\dots,x_t} \big\} \, d\gamma - \int f_t \big\{ g_t - \int g_t \, d\mu^{x_1,\dots,x_t} \big\} \, d\gamma^m \big|,
\end{aligned}$$

where we have used the fact that c is bounded and continuous and thus by weak convergence of the measures the difference between the first integrals becomes arbitrarily small. For notational simplicity, let us denote by

$$\Omega(F) := \int F \, d\gamma^m - \int F \, d\gamma,$$

where F takes the role of any of the above functions. We note, that if F is Lipschitz-continuous for some constant L, we have by the Kantorovich-Rubinstein formula, Proposition 2.12, and equation (5.13) that

$$|\Omega(F)| \leq L\mathcal{W}_1(\gamma^m, \gamma) < \tilde{\varepsilon},$$

for m large enough. Now, h_1, h_2, f_t, g_t are all $L-$Lipschitz by assumption. Further, we assume that Property 5.7 holds for all t, so in particular $x \mapsto \int g_t(x) \, d\mu^{x_1,\dots,x_t}(x)$ is Lipschitz, with a Lipschitz constant smaller or equal the product of the Lipschitz constants of g and $x \mapsto \int d\mu^{x_1,\dots,x_t}(x)$. As all functions are Lipschitz and thus bounded on K, in particular also their products are Lipschitz. We thus note that

$$\begin{aligned} \big| \sup_{h \in \mathcal{H}_L} \Phi(\hat{T}, h) - \sup_{h \in \mathcal{H}_L} \Phi(T^m, h) \big| &\leq \tilde{\varepsilon} + (N+2)C\tilde{\varepsilon} \\ &\leq \tilde{C}\tilde{\varepsilon} \end{aligned}$$

for m large enough, and hence, setting $\varepsilon := \frac{\tilde{\varepsilon}}{\tilde{C}}$, and noting that $\mathrm{Lip}_L^m(\mathbb{R}^d) \subset \mathrm{Lip}_L(\mathbb{R}^d)$,

$$(P_L) = \sup_{h \in \mathcal{H}_L} \Phi(\hat{T}, h) \geq \sup_{h \in \mathcal{H}_L} \Phi(T^m, h) - \varepsilon \geq \sup_{h \in \mathcal{H}_L^m} \Phi(T^m, h) - \varepsilon \geq (P_L^m) - \varepsilon,$$

yielding the claim. □

5.2.2 Solutions of (P_L) are solutions of (P)

By moving to reformulated constraints on the set of test functions, we have thus constructed a Min-Max problem (P_L) that can in fact be approximated by the neural network function based problem (P_L^m) and does not exhibit the troubles in approximating supremum and infimum at the same time, as we have seen for the original problem (P) in Example 5.6. However, we still need to argue why, at least for the case where we have shown the approximation capabilities - that is, when all measures are compactly supported and the first marginal μ possesses Lipschitz-regular kernels - solving the relaxed problem (P_L) gives us the solution of (P). We will first show that we can impose a necessary condition in the form of linear constraints for a transport plan to be causal using Lipschitz-functions, same as we did before using bounded continuous functions.

Lemma 5.10. *Assume μ and ν are supported on K^N for a compact subset $K \subseteq \mathbb{R}$. Then, if a measure γ is causal from μ to ν, we have $\gamma \in \Pi(\mu, \nu)$ and, for all $t < N$ and*

all $f_t \in Lip_L(\mathbb{R}^t), g_t \in Lip_L(\mathbb{R}^N)$

$$\int f_t(y_1, \dots y_t)\big\{g_t(x_1, \dots, x_N) - \int g_t(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots, d\bar{x}_N)\big\}\, \gamma(dx, dy) = 0. \tag{5.14}$$

for some $L > 0$.

Proof. According to our definition of causality and Proposition 2.24, we know that any measure γ is causal from μ to ν iff $\gamma \in \Pi(\mu, \nu)$ and for all $t < N$, $f_t \in \mathcal{C}_b(\mathbb{R}^t), g_t \in \mathcal{C}_b(\mathbb{R}^N)$ we have

$$\int f_t\big\{g_t - \int g_t\, d\mu^{x_1,\dots,x_t}\big\}\, d\gamma = 0 \;\text{ for all } f_t \in \mathcal{C}_b(\mathbb{R}^t), g_t \in \mathcal{C}_b(\mathbb{R}^N) \tag{5.15}$$

We note, that due to the fact that γ and μ are supported on K^N, (5.15) is equivalent to the same statement for all $t < N$, $f_t \in \mathcal{C}_b(K^t), g_t \in \mathcal{C}_b(K^N)$, where we restrict the domain of the test functions to the respective cartesian products of the compactum K,

$$\int f_t\big\{g_t - \int g_t\, d\mu^{x_1,\dots,x_t}\big\}\, d\gamma = 0 \;\text{ for all } f_t \in \mathcal{C}_b(K^t), g_t \in \mathcal{C}_b(K^N) \tag{5.16}$$

We now fix $L > 0$ and note that for any $\varphi \in \mathrm{Lip}_L(\mathbb{R}^d)$, $d \leq N$ and any $x \in K^d$ we have

$$|\varphi(x)| \leq L||x||_{K^d} \leq L \sup_{x \in K^d} ||x||_{K^d} < \infty,$$

as K^d is compact, so in fact $f|_{K^d} \in \mathcal{C}_b(K^d)$. But then, (5.16) implies

$$\int f_t\big\{g_t - \int g_t\, d\mu^{x_1,\dots,x_t}\big\}\, d\gamma = 0, \tag{5.17}$$

for all $t < N$, $f_t \in \mathrm{Lip}_L(\mathbb{R}^t), g_t \in \mathrm{Lip}_L(\mathbb{R}^N)$, so the statement holds. □

In order to show that the solutions to (P_L) and (P) are equivalent, we further need to impose a sufficient condition for causality. We need to argue why by checking these constraints, we already optimize over all possible causal transport plans. We start by first showing that on compact sets, we can approximate continuous and bounded functions by Lipschitz functions with an arbitrary Lipschitz constant.

Lemma 5.11. *Let* $K \subseteq \mathbb{R}^N$ *be a compact subset and let* $f \in \mathcal{C}_b(K)$. *Then, for all* $\varepsilon > 0$, *there exists an* $L > 0$ *and an* $L-$*Lipschitz function* $g \in Lip_L(K)$ *such that*

$$||f - g||_{\infty,K} < \varepsilon.$$

Proof. We begin to note that the functions

$$g_L(x) := \inf_{y \in K} \{f(y) + L|x - y|\} \tag{5.18}$$

are $L-$Lipschitz for every $L > 0$, as for all $x, z \in K$

$$\begin{aligned}\big|g_L(x) - g_L(z)\big| &= \big|\inf_{y\in K}\{f(y) + L|x-y|\} - \inf_{y\in K}\{f(y) + L|z-y|\}\big| \\ &\leq \sup_{y\in K}\big|\{L|x-y| - L|z-y|\}\big| \leq L|x-z|,\end{aligned}$$

by the triangle inequality. As f is continuous on a compact set, it is in particular uniformly continuous, so we have that for all $a > 0$ there exists a $b > 0$ such that

$$|f(x) - f(y)| \leq a + b|x-y|,$$

for all $x, y \in K$. Hence if $f(y) + L|x-y| \leq f(x)$, we must have $(L-b)|x-y| \leq a$. If L is large, then $|x-y| \leq \frac{a}{L-b}$, whenever $f(y) + L|x-y| \leq f(x)$. As $g_L(x) \leq f(x)$, for all L, the infimum in $g_L(x)$ must thus be obtained over those y such that $|x-y| \leq \frac{a}{L-b}$. Now, again as f is uniformly continuous, for any $\varepsilon > 0$, and for all $x, y \in K$, $|f(x) - f(y)| < \varepsilon$, if $|x-y| \leq \min\{\frac{\varepsilon - a}{b}, \frac{a}{L-b}\}$. We thus observe, for all $x \in K$,

$$\begin{aligned}\big|f(x) - g_L(x)\big| = \big|f(x) - \inf_{y\in K}\{f(y) + L|x-y|\}\big| &\leq \big|f(x) - \inf_{|x-y|\leq \frac{a}{L-b}}\{f(y) + L|x-y|\}\big| \\ &\leq \varepsilon + L\Big(\frac{a}{L-b}\Big) \leq \varepsilon + 2a,\end{aligned}$$

for large enough L. But so, as ε and a are arbitrary, we get the desired result. □

This gives us the possibility to reformulate the causality constraint we have encountered in Proposition 2.24 using Lipschitz-functions with a fixed constant $L > 0$.

Proposition 5.12. *Assume μ and ν are supported on K^N for a compact subset $K \subseteq \mathbb{R}$. Assume $\gamma \in \Pi(\mu, \nu)$ and for some $L > 0$ it holds that*

$$\int f_t(y_1, \dots y_t)\big\{g_t(x_1, \dots, x_N) - \int g_t(x_1, \dots, x_t, \bar{x}_{t+1}, \dots, \bar{x}_N)\ \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, \dots,\ d\bar{x}_N)\big\}\gamma(dx, dy) = 0,$$

for all $t < N$ and all $f_t \in Lip_L(\mathbb{R}^t), g_t \in Lip_L(\mathbb{R}^N)$. Then, γ is causal from μ to ν.

Proof. As shown in Proposition 2.24, we know that if $\gamma \in \Pi(\mu, \nu)$ and for all $t < N, f_t \in \mathcal{C}_b(\mathbb{R}^t), g_t \in \mathcal{C}_b(\mathbb{R}^N)$ we have

$$\int f_t\big\{g_t - \int g_t\, d\mu^{x_1,\dots,x_t}\big\}\, d\gamma = 0 \ \text{ for all } f_t \in \mathcal{C}_b(\mathbb{R}^t), g_t \in \mathcal{C}_b(\mathbb{R}^N), \tag{5.19}$$

then γ is causal from μ to ν. In Lemma 5.11, we showed that we can approximate bounded continuous functions on compact subsets arbitrarily well by Lipschitz functions. Let us thus fix $\varepsilon > 0$ and for $t < N$, functions $f_t \in \mathcal{C}_b(\mathbb{R}^t), g_t \in \mathcal{C}_b(\mathbb{R}^N)$. Then, by Lemma 5.11,

we can find a constant $\hat{L} > 0$ and functions $\hat{f}_t \in \mathrm{Lip}_{\hat{L}}(\mathbb{R}^t), \hat{g}_t \in \mathrm{Lip}_{\hat{L}}(\mathbb{R}^N)$ such that for all $t < N$,

$$||f_t - \hat{f}_t||_{\infty,K^t} < \varepsilon \tag{5.20}$$

$$||g_t - \hat{g}_t||_{\infty,K^N} < \varepsilon. \tag{5.21}$$

We note, for all $t < N$, by (5.21),

$$||\int g_t \, d\mu^{x_1,\dots,x_t} - \int \hat{g}_t \, d\mu^{x_1,\dots,x_t}||_{\infty,K^N} \leq ||g_t - \hat{g}_t||_{\infty,K^N} \int d\mu^{x_1,\dots,x_t} < \varepsilon,$$

as $\mu^{x_1,\dots,x_t}$ is a probability measure. But then, following similar arguments as in the proof of Proposition 5.9, we find

$$\begin{aligned}
&||\int f_t\big\{g_t - \int g_t \, d\mu^{x_1,\dots,x_t}\big\} \, d\gamma - \int \hat{f}_t\big\{\hat{g}_t - \int \hat{g}_t \, d\mu^{x_1,\dots,x_t}\big\} \, d\gamma||_{\infty,K^N\times K^t} \\
\leq &||f_t\big\{g_t - \int g_t \, d\mu^{x_1,\dots,x_t}\big\} - \hat{f}_t\big\{\hat{g}_t - \int \hat{g}_t \, d\mu^{x_1,\dots,x_t}\big\}||_{\infty,K^N\times K^t} \int d\gamma \\
\leq &||f_t g_t - \hat{f}_t\hat{g}_t||_{\infty,K^N\times K^t} + ||f_t \int g_t \, d\mu^{x_1,\dots,x_t} - \hat{f}_t \int \hat{g}_t \, d\mu^{x_1,\dots,x_t}||_{\infty,K^N\times K^t},
\end{aligned}$$

as γ is a probability measure. Adding and subtracting mixed terms for both summands allows us to note

$$\begin{aligned}
&||\int f_t\big\{g_t - \int g_t \, d\mu^{x_1,\dots,x_t}\big\} \, d\gamma - \int \hat{f}_t\big\{\hat{g}_t - \int \hat{g}_t \, d\mu^{x_1,\dots,x_t}\big\} \, d\gamma||_{\infty,K^N\times K^t} \\
\leq &||f_t||_{\infty,K^t}||g_t - \hat{g}_t||_{\infty,K^N} + ||f_t - \hat{f}_t||_{\infty,K^t}||\hat{g}_t||_{\infty,K^N} + \\
&||f_t||_{\infty,K^t}||\int g_t \, d\mu^{x_1,\dots,x_t} - \int \hat{g}_t \, d\mu^{x_1,\dots,x_t}||_{\infty,K^N} + ||f_t - \hat{f}_t||_{\infty,K^t}||\int \hat{g}_t \, d\mu^{x_1,\dots,x_t}||_{\infty,K^t} \\
\leq &C\varepsilon,
\end{aligned} \tag{5.22}$$

where we note that f_t is bounded on K by assumption, $\hat{g}_t$ is bounded on the compact set K as it is $\hat{L}-$Lipschitz and $\int \hat{g}_t \, d\mu^{x_1,\dots,x_t}$ is bounded as the integral of a bounded function over a compact set, and the norms of the differences between the functions are arbitrarily small by (5.20) and (5.21).
Now, by assumption, for some $L > 0$, we have

$$\int \tilde{f}_t\big\{\tilde{g}_t - \int \tilde{g}_t \, d\mu^{x_1,\dots,x_t}\big\} \, d\gamma = 0 \;\text{ for all } \tilde{f}_t \in \mathrm{Lip}_L(\mathbb{R}^t), \tilde{g}_t \in \mathrm{Lip}_L(\mathbb{R}^N). \tag{5.23}$$

We note, that for any $\tilde{g}_t \in \mathrm{Lip}_L(\mathbb{R}^N)$,

$$\big|\frac{\hat{L}}{L}\tilde{g}_t(x) - \frac{\hat{L}}{L}\tilde{g}_t(y)\big| = \frac{\hat{L}}{L}|\tilde{g}_t(x) - \tilde{g}_t(y)| \leq \frac{\hat{L}}{L}L|x - y| = \hat{L}|x - y|,$$

for all $x, y \in \mathbb{R}^N$ as $\tilde{g}_t$ is L-Lipschitz, so in particular, $\frac{\hat{L}}{L}\tilde{g}_t$ is $\hat{L}-$Lipschitz. The same argument works for all $\tilde{f}_t \in \mathrm{Lip}_L(\mathbb{R}^t)$.

But if

$$\int \tilde{f}_t\big\{\tilde{g}_t - \int \tilde{g}_t \, d\mu^{x_1,\dots,x_t}\big\}\, d\gamma = 0 \;\text{ for all } \tilde{f}_t \in \mathrm{Lip}_L(\mathbb{R}^t), \tilde{g}_t \in \mathrm{Lip}_L(\mathbb{R}^N),$$

it follows that

$$\int \frac{\hat{L}}{L}\tilde{f}_t\big\{\frac{\hat{L}}{L}\tilde{g}_t - \int \frac{\hat{L}}{L}\tilde{g}_t \, d\mu^{x_1,\dots,x_t}\big\}\, d\gamma = 0 \;\text{ for all } \tilde{f}_t \in \mathrm{Lip}_L(\mathbb{R}^t), \tilde{g}_t \in \mathrm{Lip}_L(\mathbb{R}^N),$$

so in fact, that (5.23) holds true for all $\hat{L}-$Lipschitz functions, is equivalent to

$$\int \hat{f}_t\big\{\hat{g}_t - \int \hat{g}_t \, d\mu^{x_1,\dots,x_t}\big\}\, d\gamma = 0 \;\text{ for all } \hat{f}_t \in \mathrm{Lip}_{\hat{L}}(\mathbb{R}^t), \tilde{g}_t \in \mathrm{Lip}_{\hat{L}}(\mathbb{R}^N). \tag{5.24}$$

By (5.22), this gives us (5.19), and thus causality of γ, as desired. □

In the framework we find ourselves in, we have thus shown in Lemma 5.10 and Proposition 5.12 that

$\gamma \in \Pi_c(\mu,\nu)$ iff $\gamma \in \Pi(\mu,\nu)$ and for some (and then all) L:
$\int f_t\big\{g_t - \int g_t \, d\mu^{x_1,\dots,x_t}\big\}\, d\gamma = 0,$ for all $t < N$ and $f_t \in \mathrm{Lip}_L(\mathbb{R}^t), g_t \in \mathrm{Lip}_L(\mathbb{R}^N)$.

Assuming μ and ν are supported on K^N for a compact subset $K \subseteq \mathbb{R}$, solving (P_L) then gives us the same solution as solving the original problem (P).

Chapter 6

Description of the algorithm

In this chapter, we will describe in detail how the approximations investigated and discussed in the last chapter can be used to develop a numerical algorithm that solves the Causal Optimal Transport problem (CP). The idea behind the algorithm again is again based on the method developed in [15] for a group of problems including the Optimal Transport problem (KP). Same as in the last chapters, we will now inspect an adaptation of this method to the Causal Optimal Transport problem in N time steps, which was developed in the course of this thesis.

6.1 Strategy

In Chapter 3, we established that both the Optimal Transport problem (KP) and the Causal Optimal Transport problem (CP) can be reformulated as a problem of the following form,

$$(\mathrm{P}) = \inf_{T \in \mathcal{B}_{D,2}} \sup_{h \in \mathcal{H}} \int c \circ T \, d\theta - \int h \circ T \, d\theta + \int h \, d\pi,$$

where the set of test functions $\mathcal{H}$ depends on the problem at hand. Here, π is a measure on $\mathbb{R}^{2\times N}$ with marginals μ and ν, c is a cost function and θ is a given latent measure on $\mathbb{R}^{D\times N}$. In the same chapter, we gave an explicit expression of the set $\mathcal{H}$ for the two problems.

We then saw in Chapter 4 that the space of neural networks has great approximation properties, leading us to define an approximative problem (P^m) in Chapter 5. Finally, in the same chapter, we investigated a relaxation of problem (P) involving Lipschitz constraints, (P_L), for which we could prove under certain conditions, convergence of solutions of an approximative problem involving neural networks, (P_L^m).

In light of these results, the strategy to numerically solve problem (P) looks as follows.

1. Given a measure π on $\mathbb{R}^{2\times N}$ with marginals μ and ν on $\mathbb{R}^N$ and a cost function c for which we want to solve problem (P), we first choose a latent measure θ on $\mathbb{R}^{D\times N}$ such that θ can be bijectively mapped onto the uniform measure on $(0,1)$.

2. Next, we choose a Lipschitz constant $L > 0$ and a dimension $m \in \mathbb{N}$ large enough and move to the approximative relaxed problem

$$(P_L^m) = \inf_{T\in N_{D,2}^m} \sup_{h\in\mathcal{H}_L^m} \int c\circ T\, d\theta - \int h\circ T\, d\theta + \int h\, d\pi$$

   We remember that for the Causal Optimal Transport problem (CP) the set $\mathcal{H}_L^m$ was defined as

$$\begin{aligned}\mathcal{H}_L^m := \Big\{ & h_1(x) + h_2(y) + \sum_{t<N} f_t(y_1,\dots y_t)\big\{g_t(x) - \\ & \int g_t(x_1,\dots,x_t,\bar{x_{t+1}},\dots,\bar{x_N})\,\mu^{x_1,\dots,x_t}(d\bar{x_{t+1}},\dots,d\bar{x_N})\big\}, \\ & h_1,h_2,g_t \in \mathrm{Lip}_L^m(\mathbb{R}^N), f_t \in \mathrm{Lip}_L^m(\mathbb{R}^t)\ \forall t<N\Big\}\end{aligned}$$

3. Seeing as the problem (P_L^m) optimizes over spaces of neural network functions, we build the functions by specifying sets of weights w_T for the function $T \in N_{D,2}^m$ and $w_h = (w_{h_1}, w_{h_2}, w_{f_1}, w_{g_1}, \dots, w_{f_{N-1}}, w_{g_{N-1}})$ for $h \in \mathcal{H}_L^m$, where the entries of w_h are the weights of the respective neural networks h decomposes into. We can then write the objective function over which to optimize, explicitly specifying the inputs and dependencies, as

$$\begin{aligned}\Phi(w_T, w_h) := & \int c(T(z;w_T))\,\theta(dz) \\ & - \int h(T(z;w_T);w_h)\,\theta(dz) + \int h((x,y);w_h)\,\pi(dx,dy)\end{aligned}$$

4. We numerically approximate this objective function and the occurring integrals by sampling S points $\{(x^i,y^i)\}_{i=1}^S \sim \pi$ and $\{z^i\}_{i=1}^S \sim \theta$ in order to approximate the integrals using a Monte Carlo approach,

$$\hat{\Phi}(w_T,w_h) = \frac{1}{S}\sum_{i=1}^S c(T(z^i;w_T)) - h(T(z^i;w_T);w_h) + h((x^i,y^i);w_h)$$

5. We note that in order to evaluate the conditional integrals appearing in all functions $h \in \mathcal{H}^m$ for the Causal Optimal Transport problem, we sample, for all $t < N$, $\tilde{S}$ points $\{\xi_t^i\}_{i=1}^{\tilde{S}} \sim \mu^{x_1,\dots,x_t}$ and similarly approximate the integrals within h using a Monte Carlo approach

6. We choose some optimization algorithm to maximize the objective over $\mathcal{H}_L^m$ while simultaneously minimizing it over $T \in N_{D,2}^m$. Given the structure of the problem,

this means that over n steps we successively update the weights w_h and w_T. In the first step, random initial weights w_h^0 and w_T^0 are chosen. Then, successively, we first update w_h leaving w_T fixed,

$$w_h^{n+1} = \max_{w_h} \hat{\Phi}(w_T^n, w_h),$$

and then do the same for w_T, leaving the updated value for w_h^{n+1} fixed,

$$w_T^{n+1} = \min_{w_T} \hat{\Phi}(w_T, w_h^{n+1}).$$

We call this process **training** the network. While all weights are allowed in the minimization step, additional care has to be taken to ensure that the weights put out by the maximization step fulfill the Lipschitz criterion such that the functions $h \in \mathcal{H}_L^m$

7. We iterate this process either until some fixed maximum number of training rounds, or until both simultaneous optimizations converge to the same objective value up to some small tolerance

A key aspect of the problem is that it involves the calculation of an inner supremum and an outer infimum at the same time. In some way, this can be interpreted as optimizing two competing networks, a **generator** generating a candidate supremum, which is then tested to satisfy the linear constraint by an **adversary** via evaluation of the inner infimum. As hinted by the usage of these terms, a network architecture of this sort is known as a **Generative Adversarial Network** - in short, GAN - which was first proposed by Goodfellow et al. in [21].

6.2 Optimization

As outlined in the last section, one key feature of the algorithm will be to choose an optimization algorithm. The algorithm chosen in [15] and in this thesis is the optimizer **Adam**, introduced by Kingma and Ba in their paper [28]. To motivate the use of Adam, we first introduce a basic optimization algorithm it is based on.

In general, an optimization problem such as the one at hand can be summarized in the following way. Given a functional F defined on some space $\mathcal{W}$, we are interested in finding local minima, that is to find an $w^* \in \mathcal{W}$ such that $F(w^*) \leq F(w)$ for all w in some neighborhood of w^*. If F is differentiable, a first-order condition for the above is $\nabla F(w^*) = 0$. If F is a convex functional then every local minimum is in fact already a global minimum, so we want to find $w^* \in \mathcal{W}$ such that $F(w^*) \leq F(w)$ for all $w \in \mathcal{W}$.

A classical approach to solve such optimization problems for differentiable functions is the **gradient descent** method, as described in [38]. Choosing an initial value $w^0 \in \mathcal{W}$, we define a sequence

$$w^{n+1} = w^n - \tau \nabla F(w^n) \tag{6.1}$$

for some $\tau > 0$, where the constant τ is termed the step-size, and $\nabla F(w^n) = \partial_w F(w^n)$ is the gradient of F with respect to the coordinate w. Given some conditions on F, the gradient descent method guarantees convergence to some local optimum. We present the following theorem, the proof of which can be found in [38].

Theorem 6.1. *Assume F is continuously differentiable with respect to w, bounded from below, and its gradient with respect to w is Lipschitz continuous for some constant L. Then, if $\tau < \frac{1}{L}$, for every w_0 the sequence $\{w_n\}$ satisfies*

$$\lim_{n\to\infty} \nabla F(w_n) = 0.$$

We note that if the objective function F is convex, all local minima are in particular global minima, so the gradient descent algorithm converges to a global minimum.

Let us inspect the gradient descent method more closely in our setting. Seeing as we defined the approximative objective function $\hat{\Phi}$ as a sum of functions evaluated at each sample (z^i) and (x^i, y^i), for $i = 1, \ldots, S$ an estimator for the computation of the gradient is given by

$$\nabla \hat{\Phi} = \frac{1}{S} \sum_{i=1}^{S} \nabla \hat{\Phi}_i(x^i, y^i, z^i),$$

or the sum of S gradients evaluated at each of the S samples. Not only does the complexity of this calculation grow linearly with the sample size, often the gradients evaluated at samples close to each other are themselves very similar.

An algorithm trying to remedy this fact is (on-line) **stochastic gradient descent** (SGD), as described in a survey on gradient-based optimization methods, [36]. Instead of in every step calculating the gradient at every sampled point, we only need to calculate the gradient for one randomly chosen point,

$$w^{n+1} = w^n - \tau \nabla F(w^n; \mathbf{x}^i), \tag{6.2}$$

where $\mathbf{x}^i$ is a randomly chosen point from a sample of points. While the convergence behaviour of SGD is more erratic than for standard gradient descent, if the step-size τ is successively decreased, the algorithm converges to local minima, or global minima, if F is convex, see [12]. A further adaption of stochastic gradient descent is to calculate the gradient in every step on a batch of $s \ll S$ points, which can increase stability in convergence, see in [36].

As hinted upon by the behaviour of SGD, adaptive step-sizes might be beneficial for quicker or more robust convergence. Similarly, many of the methods described in the survey [36], use as update direction at step n a weighted average of the recent update directions. Compared with using purely the gradient - evaluated at either one, a subset of, or all samples - this can also increase speed and robustness of convergence. The algorithm chosen for this thesis, **Adam**, combines these two features:

$$\begin{aligned} m_{n+1} &= \left(\beta_1 m_n + (1-\beta_1)\nabla F(w_n)\right)\frac{1}{1-\beta_1^n} \\ v_{n+1} &= \left(\beta_2 v_n + (1-\beta_2)|\nabla F(w_n)|^2\right)\frac{1}{1-\beta_2^n} \\ w_{n+1} &= w_n - \frac{\tau}{\sqrt{v_{n+1}}+\epsilon} m_{n+1}, \end{aligned} \tag{6.3}$$

where $\beta_1, \beta_2, \epsilon$ are regularization constants, and initially we choose $m_0, v_0 = 0$. The gradient at the n-th step is evaluated at a batch of s samples. The fixed step-size τ is in every step divided by an exponential average of the gradients squared, v_n, while the update is performed in direction of the exponential average of the gradients, m_n. The update rule 6.3 converges for convex functionals F to a global optimum, see [28].

6.3 Modelling the Lipschitz constraint

For the algorithm, we need to find a method to enforce Lipschitz continuity for the individual functions that the functions $h \in \mathcal{H}_L^m$ are composed of. By default, the weights w_h^n put out by the optimizer in the n-th training step will not produce Lipschitz continuous functions h_1^n, h_2^n and f_t^n, g_t^n, for all $t < N$. To remedy this, a strategy proposed by [23] and used in the algorithm solving the Optimal Transport problem in [15] is to subtract a penalty term from the functions,

$$\varphi^* = \varphi - \lambda\left((|| \bigtriangledown \varphi|| - L)^+\right)^2,$$

for some $\lambda > 0$, and $\varphi \in \{h_1^n, h_2^n, f_1^n, g_1^n, \ldots, f_{N-1}^n, g_{N-1}^n\}$. We remind ourselves that a function φ is $L-$Lipschitz if and only if the norm of its gradient is at most L. By subtracting the above penalty term, which is 0 if any only if $|| \bigtriangledown \varphi|| \leq L$ - in other words, if φ is L-Lipschitz - and positive otherwise, we ensure that the supremum over all functions in $\mathcal{H}^m$ is actually taken over these h such that all individual functions comprising h are close to being Lipschitz-continuous. The output of the maximization steps will thus, as $n \to \infty$, converge to the maximally possible function in $\mathcal{H}_L^m$.

6.4 Sampling from the conditional distribution

As noted in section 6.1, building the function $h \in \mathcal{H}_L^m$ requires at every step the computation of a Monte Carlo integration evaluated at points $\{\xi_t^j\}_{j=1}^{\tilde{S}} \sim \mu^{x_1,\dots,x_t}$ for all $t < N$. While for some examples, the conditional distribution $\mu^{x_1,\dots,x_t}$ can be computed explicitly and thus sampled from directly, in general we have to make use of a sampling algorithm. The algorithm used in this thesis is based on kernel estimation, taken from [40], and functions as follows. For every time-step t and given sampled values $\{(x_1^i, \dots, x_t^i)\}_{i=1}^S$, we first fix a bandwidth η following a rule of thumb specified in [40],

$$\eta = 1.06\hat{\sigma}s^{-1/5},$$

where $\hat{\sigma}$ is the standard deviation of the sample $\{x_1^i, \dots, x_t^i\}_{i=1}^S$. We then sample more points $\{\hat{x}\}_{j=1}^{4S}$ from the joint distribution μ of all N time steps - in numerical experiments, $4S$ points for every step has shown good results - and consider the set of points $(\hat{x}_1^j, \dots, \hat{x}_N^j)$ such that

$$||(\hat{x}_1^j, \dots, \hat{x}_t^j) - (x_1^j, \dots, x_t^j)|| < \eta \tag{6.4}$$

We then sample $\tilde{S}$ points $\{\xi_t^j\}_{j=1}^{\tilde{S}}$ from the set of all $\{\hat{x}\}_{j \text{ fulfills (6.4)}}$. At these points, we perform the Monte Carlo integration to approximate the integrals evaluated at the $i-$th sampled point x^i,

$$\int g_t(x_1^i, \dots, x_t^i, \bar{x}_{t+1}, \dots, \bar{x}_N)\, \mu^{x_1,\dots,x_t}(d\bar{x}_{t+1}, d\bar{x}_N) \approx \frac{1}{\tilde{S}} \sum_{j=1}^{\tilde{S}} g_t(x_1^i, \dots, x_t^i, \xi_t^j).$$

6.5 Algorithm

Finally, we present in pseudo-code the algorithm used for the numerical experiments in this thesis. The algorithm is based on an algorithm proposed in [15] for the Optimal Optimal Transport problem and adapted to solve the additional constraints poised for the Causal Optimal Transport problem. For the algorithm, we first need to specify the parameters related to the problem, that is, the marginal distributions, the number of time steps and the cost function. Further, parameters related to the algorithm, such as the latent measure θ, a batch size and conditional sample size, s and $\tilde{s}$, respectively, and a maximum number of iterations n_{max}, need to be chosen. Further, we specify a number of steps n_{mean}, over which we will average the return value of the algorithm. This is done to increase stability and to flatten out some approximation errors. Finally, we can institute an early stopping criterion. We want the algorithm to stop training before reaching n_{max} rounds, if the average change over the last n_{mean} rounds was less or equal then some tolerance level. If this never occurs during training, the final value after n_{max} rounds will be accepted as output value. If it does occur, the algorithm will put out the value at which the tolerance level is reached.

Algorithm 1 Causal Optimal Transport

Inputs: cost function c, marginal distributions μ and ν, time steps N, latent measure θ, batch size s, conditional batch sitze $\tilde{s}$, Lipschitz constant λ, maximum number of iterations n_{max}, number of steps over which to average return value n_{mean}

Require: random initialization of neural network weights w_T and $w_h = (w_{h_1}, w_{h_2}, w_{f_1}, w_{g_1}, \dots, w_{f_{N-1}}, w_{g_{N-1}})$

for $\tau \leq n_{max}$ **do**

sample $\{(x^i, y^i)\}_{i=1}^s \sim \pi$

sample $\{z^i\}_{i=1}^s \sim \theta$

sample $\{\xi_t^j\}_{j=1}^{\tilde{s}} \sim \mu^{x_1^i,\dots,x_t^i}$ for all $t < N$ and all $i = 1, \dots, s$

build $h^i = h_1(x^i) + h_2(y^i) + \sum_{t<N} f_t(y_1^i, \dots, y_t^i)\{g_t(x^i) - \frac{1}{\tilde{s}} \sum_{j=1}^{\tilde{s}} g_t(x_1^i, \dots, x_t^i, \xi_t^j)\}$ for all $i = 1, \dots, s$

add Lipschitz-constraint $\varphi = \varphi - \lambda \left((|| \bigtriangledown h_1^i|| - L)^+\right)$ for all $\varphi \in \{h_1^n, h_2^n, f_1^n, g_1^n, \dots, f_{N-1}^n, g_{N-1}^n\}$

evaluate $\hat{\Phi}(w_T, w_h) = \frac{1}{s} \sum_{i=1}^s c(T(z^i; w_T)) - h(T(z^i; w_T); w_h) + h((x^i, y^i); w_h)$

$w_h \leftarrow \text{Adam}(\text{-}\hat{\Phi}(w_T, w_h))$

$w_T \leftarrow \text{Adam}(\hat{\Phi}(w_T, w_h))$

end for

Return: $\frac{1}{n_{mean}} \sum_{k=n_{max}-n_{mean}+1}^{n_{max}} \hat{\Phi}_k(w_T, w_h)$

We will use this algorithm for the numerical experiments in the following chapter.

Chapter 7

Numerical experiments

In the final chapter, we will present the results of some numerical experiments solving Optimal Transport and Causal Optimal Transport problems. We will investigate three different examples, inspect some details of the algorithm as described in the previous chapter and discuss some difficulties and interesting features.

7.1 List of examples

We will apply the algorithm to three related examples. For all of these, we will investigate what theoretical solution to expect, and we will argue why the marginal distributions fulfil the criteria that we have outlined in Chapter 5, and thus why we can expect the algorithm to converge.

Similar experiments were also undertaken for a different class of examples, computing Wasserstein distances between discrete distributions, supported on finite sets. However, the algorithm had difficulties finding the theoretical values, eventually always converging to different values. This happened both when starting with discrete and continuous latent measures, and the difficulty seemed to lie in the outer infimum problem, with the algorithm unable to find the optimal map T. Due to time constraints, we were not able to resolve these problems and so, the results of numerical experiments on this class of examples were finally not added into the thesis. An interesting direction for further work could be to apply the algorithm to experiments of this class. A possible remedy could be to approximate the discrete distributions by convolutions with Gaussian distributions, thereby hopefully alleviating the difficulties in finding the correct solutions.

7.1.1 Experiments

We inspect both the Optimal Transport problem (KP) and the Causal Optimal Transport problem in N time steps (CP) in the case that the marginals μ and ν, and the marginals

at each of the $t \leq N$ time steps μ_t and ν_t, respectively, are two Gaussian distributions. Numerical experiments are undertaken for the following cases.

1. The first experiment is to solve the standard Optimal transport problem (KP) for two-dimensional Gaussians $\mu^{(1)}, \nu^{(1)}$, with

$$\mu^{(1)} \sim \mathcal{N}\left(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)\right), \quad \nu^{(1)} \sim \mathcal{N}\left(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}4&0\\0&4\end{smallmatrix}\right)\right).$$

   We use the cost function $c(x,y) = ||x-y||_2^2$, with $||\cdot||_2$ the Euclidean norm, so (KP) in this case then is

$$\inf_{\gamma \in \Pi(\mu^{(1)}, \nu^{(1)})} \int ||x-y||_2^2 \, \gamma(dx, dy). \tag{EXP1}$$

   Solving (EXP1) then gives us the value of $\mathcal{W}_2^2(\mu^{(1)}, \nu^{(1)})$. By [20], the squared Wasserstein distance between two Gaussians μ and ν with means m_1, m_2 and covariance matrices Σ_1, Σ_2, respectively, is given by

$$\mathcal{W}_2^2(\mu, \nu) = ||m_1^2 + m_2^2||_2^2 + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}),$$

   where Tr denotes the euclidean norm and Tr the trace of a matrix. Using this formula, we can easily compute that the value of (EXP1) is given by $\mathcal{W}_2^2(\mu^{(1)}, \nu^{(1)}) = 2(\sqrt{4} - \sqrt{1})^2 = 2$. We can even directly specify the optimal transport plan for (EXP1). Take the transport plan induced by the map $T : x \mapsto 2x$, that is the pushforward under T, $\gamma := (\mathrm{id}, T)_{\#}\mu^{(1)}$. Then, $\gamma_1 = \mu^{(1)}$, and $\gamma_2 = \nu^{(1)}$, so $\gamma \in \Pi(\mu, \nu)$, and furthermore,

$$\int ||x-y||_2^2 \, \gamma(dx, dy) = \int ||x - 2x||_2^2 \, \mu^{(1)}(dx) = \int (x_1^2 + x_2^2) \, \mu^{(1)}(dx) = 2,$$

   as $\mu_1^{(1)} \sim \mathcal{N}(0,1)$ and $\mu_2^{(1)} \sim \mathcal{N}(0,1)$, so γ is indeed optimal. Finally, seeing as the first marginal $\mu^{(1)}$ has a density f whose derivative is bounded, f is in particular Lipschitz and further, for all compact $K \subset \mathbb{R}^2$ there exists a $\delta > 0$ such that $f|_K \geq \delta$. Restricted on to a sufficiently large compact set K, $\mu^{(1)}$ thus fulfills Property 5.7. By Proposition 5.9 and Proposition 5.12, and noting that for numerical purposes, convergence restricted on to a large enough compact set is enough, solving (P_L^m) for m large enough gives us a good approximation to the theoretical value.

2. In the second experiment, we want to investigate a Causal Optimal Transport problem in $N = 2$ time steps. For this, we will use the same marginal distributions, but instead, now view them as the law of a stochastic process in two time steps. The marginal distributions $\mu^{(2)}$ and $\nu^{(2)}$ are here then given by

$$\mu^{(2)} = (\mu_1^{(2)}, \mu_2^{(2)}), \text{ with } \mu_1^{(2)} \sim \mathcal{N}(0,1), \mu_2^{(2)} \sim \mathcal{N}(0,1), \mu_1^{(2)} \perp \mu_2^{(2)},$$

   and

$$\nu^{(2)} = (\nu_1^{(2)}, \nu_2^{(2)}), \text{ with } \nu_1^{(2)} \sim \mathcal{N}(0,4), \nu_2^{(2)} \sim \mathcal{N}(0,4), \nu_1^{(2)} \perp \nu_2^{(2)}.$$

We will again use the same cost function, $c(x, y) = ||x - y||_2^2$, so that the problem presents as

$$\inf_{\gamma \in \Pi_c(\mu^{(2)}, \nu^{(2)})} \int ||x - y||_2^2 \, \gamma(dx, dy). \tag{EXP2}$$

We note now that in fact, the values of (EXP2) and (EXP1) coincide. In specifying the marginal distributions, we required $\mu_1^{(2)} \perp \mu_2^{(2)}$ and $\nu_1^{(2)} \perp \nu_2^{(2)}$, that is, the two time steps to be independent. But this means, that the joint distributions of the marginals in both time steps, $\mu^{(2)}$ and $\nu^{(2)}$, are Gaussian themselves, and in fact, coincide with $\mu^{(1)}$ and $\nu^{(1)}$. We then already know the optimal transport plan from $\mu^{(2)}$ to $\nu^{(2)}$, namely $\gamma := (\mathrm{id}, T)_{\#}\mu^{(2)}$, for $T : x \mapsto 2x$. But γ is also casual. Indeed, $T(x_1, x_2) = 2(x_1, x_2)$, is adapted in the sense of Definition 2.17. But then, the transport plan induced by T is causal. Seeing as γ is optimal and causal, it is then optimal for the Causal Optimal Transport problem (EXP2). The values of (EXP1) and (EXP2) thus coincide at $\mathcal{W}_2^2(\mu^{(2)}, \nu^{(2)}) = 2$. By a similar reasoning as above, the marginal $\mu^{(2)}$ also fulfills the needed criteria for the algorithm to converge.

3. Thirdly, we adapt the second experiment in some way. Instead of requiring the distributions of the two time steps to be independent, we now view the problem with them being equivalent. Conceptually, this means that the marginals now model the law of a stochastic process $(\{X_t\}_{t=1}^2)$ with $X_2 = X_1$, and $(\{Y_t\}_{t=1}^2)$ with $Y_2 = Y_1$, respectively. The marginal distributions are then given as

$$\mu^{(3)}(dx_1, dx_2) = \mu_1^{(3)}(dx_1)\delta_{x_1}(dx_2), \text{ with } \mu_1^{(3)} \sim \mathcal{N}(0, 1)$$

and

$$\nu^{(3)}(dy_1, dy_2) = \nu_1^{(3)}(dy_1)\delta_{y_1}(dy_2), \text{ with } \nu_1^{(3)} \sim \mathcal{N}(0, 4).$$

Again, we use the same cost function $c(x, y) = ||x - y||_2^2$, to arrive at the problem

$$\inf_{\gamma \in \Pi_c(\mu^{(3)}, \nu^{(3)})} \int ||x - y||_2^2 \, \gamma(dx, dy). \tag{EXP3}$$

However, the problem differs in a way that now, the conditional distributions of the marginals at the second time step given the first are now given by $(\mu^{(3)})^{x_1} = \delta_{x_1}$ and $(\nu^{(3)})^{y_1} = \delta_{y_2}$, respectively. But in this case, any transport plan is already causal. Indeed, take a transport plan $\gamma \in \Pi(\mu^{(3)}, \nu^{(3)})$, and an arbitrary Borel-measurable set B_1. Take two processes X and Y as above with laws $\mu^{(3)}$ and $\nu^{(3)}$. Then, $\mathbb{P}(Y_1 \in B_1 \mid X_1, X_2) = \mathbb{P}(Y_1 \in B_1 \mid X_1)$, as $X_2 = X_1$ and hence, the conditional distribution of Y_1 given X_1 and X_2 is already determined by the value of X_1. The optimal transport is thus already also the optimal causal transport. In fact, the same transport plan as for the first two examples, $\gamma := (\mathrm{id}, T)_{\#}\mu^{(3)}$, with

$T : x \mapsto 2x$, is optimal also for (EXP3). We note that

$$\int ||x-y||_2^2 \, \gamma(dx, dy) = \int ||x - 2x||_2^2 \, \mu^{(3)}(dx) = \int (x_1^2 + x_2^2) \, \mu_1^{(3)} d(x_1) \delta_{x_1} d(x_2)$$
$$= 2 \int x_1^2 \, \mu_1^{(3)} d(x_1) = 2,$$

as $\mu_1^{(3)} \sim \mathcal{N}(0, 1)$. Finally, we need to argue why our algorithm converges for this experiment. This is the case as $\mu^{(3)}$ can be viewed as the law a stochastic process $(X_t)_{t=1}^2$ with $X_1 \sim \mathcal{N}(0, 1)$, and $X_2 = X_1$, so $X_2 = F(X_1)$ for $F : x \mapsto x$. But as F is clearly Lipschitz, $\mu^{(3)}$ thus fulfills Property 5.7, so the algorithm converges.

7.2 Presentation of results

For each of the examples introduced above, we will now investigate the behaviour of training and convergence to the final value. In implementing the experiments, we followed the strategy outlined in Chapter 6, in particular Algorithm 1.

For all three experiments, the networks were trained over a maximum of 10000 rounds. Concurrently, if the mean of change over the last 500 rounds at some point became smaller than a tolerance value of 0.01, the network training was stopped. In this case, the mean of the last 500 rounds was accepted as final value of convergence. In case of no interruption, similarly, the mean value over the last 500 rounds was output as final value.

In all experiments, the network architecture was composed of four layers. While a larger number of layers should theoretically improve approximation capabilities, during trials we found that in relation, increasing the sample size and network size seemingly gave a better improved convergence when compared to how much increasing that parameter would influence computing time. We thus followed [15] in choosing four layers. Further, we followed their example in specifying the hyperbolic tangent as activation function for the generator maximizing network and the ReLU for the adversarial minimizing network. Further, we always chose a Lipschitz constant $\lambda = 1$. This can be reasoned by the fact that the cost functions in all experiments are 1-Lipschitz, and a higher Lipschitz constraint led to less stable convergence in trials.

For all trials, we chose as latent measure $\theta = \mathcal{U}_{[0,1]^{2\times 2}}$, the uniform measure on the unit cube in dimension 2×2. Further, we always chose a large batch size of $s = 512$ at which to perform the update direction calculation of the optimizer in every step. While this increased computation time significantly, in most trials, at lower sample sizes the convergence behaviour was erratic or very slow otherwise. All of the following computations were performed in Python 3.7 using the library Tensorflow 1.15.0., [1], for

building the neural networks, initiating the network weights, and training the network using the Adam optimizer. The code for the following experiments can be found at `https://github.com/FabianPfurtscheller/causal_optimal_transport`.

7.2.1 Wasserstein distance between gaussians

(EXP1) In the first experiment, we investigated convergence over a mean of five different runs at a fixed sample size s but different network sizes m. Below, compare the convergence at a fixed sample size of $s = 1024$ and network sizes $m = 64, 128, 256$. Plotted is in all three cases the averaged value over all five different runs at each training step, further averaged over the last 500 values, and plotted against the theoretical value, a dashed line in the background. We note that for increasing network sizes, convergence happens both faster and in a much more stable way. Inspecting the three graphs more closely, we see that also the smaller network sizes produce objective values around the theoretical value quite quickly. However, while the largest network changes very little as soon as the theoretical value is reached, the smaller networks still move around and never reach the tolerance value that was put in in order to stop training before the final round. The biggest network, at a network size of $m = 256$ did reach this threshold, however, in order to compare better with the behaviour of the smaller networks, we showcase here the results over a full round of training.

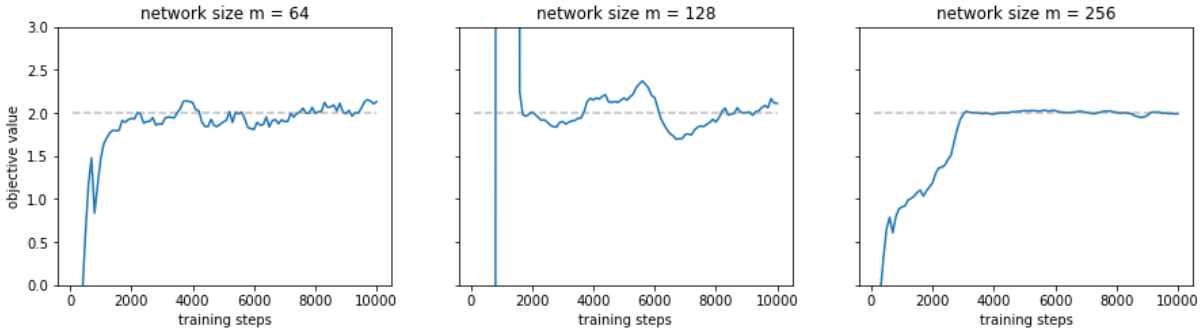

Figure 7.1: Convergence of (EXP1) at different network sizes

However, this faster and more stable convergence comes at the cost of significantly increased computing time. In Table 7.2.1, we note the average runtime of a full run of training for 10000 steps for each of the three networks, at sample size $s = 1024$.

network size m	average runtime (seconds)
64	530
128	1834
256	7876

Table 7.1: Average computation times per run for (EXP1) at different network sizes

(EXP2) Noting the long computation times and convergence behaviour already for the computationally simpler classical Optimal Transport problem, we produced the following trials only with one run, instead of taking the mean of five training runs. The following figure then shows the convergence for the algorithm approximating (EXP2) with a sample size $s = 512$ and network size $m = 512$, at a runtime of 31283 seconds. We observe that while the convergence behaviour is relatively stable, it takes many rounds for the algorithm to come close to the theoretical value. Further, in the trials, the threshold to stop training prematurely before the maximum 10000 steps was never reached.

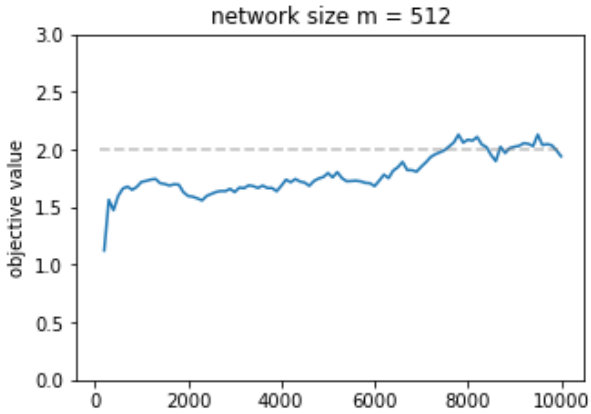

Figure 7.2: Convergence of (EXP2) at network size $m = 512$

(EXP3) The seemingly similar problem (EXP3) was investigated with the same specifications, a sample size $s = 512$ and a network size $m = 512$. Interestingly, the runtime for this trial, while still very high, was lower compared to (EXP2) at 18564 seconds, which is due to the sampling of the conditional distribution $\mu^{x_1,\dots,x_t} = \delta_{x_1}$ being easier in this case when compared to (EXP2). However, convergence behaviour was much more unstable, with the algorithm moving erratically around the desired theoretical value. Below again we inspect the graph of convergence.

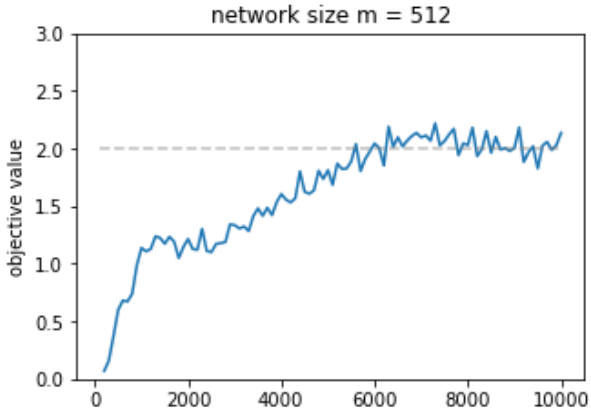

Figure 7.3: Convergence of (EXP3) at network size $m = 512$

Chapter 8

Conclusion

In this thesis, an algorithm developed for the Optimal Transport problem was successfully adapted for a class of transport problems with linear constraints, the Causal Optimal Transport problem in N discrete time steps. We first introduced the problem, illustrated it with some examples, and stated some important results from the theory of transport problems. Next, we reformulated the problem based on the scheme for the Optimal Transport problem from [15]. We then showed the theoretical approximation capabilities of an approximative problem based on neural networks, and investigated its convergence to the solution of the Causal Optimal Transport problem under some further conditions. Finally, we described the proposed algorithm, and undertook some numerical experiments for commonly occurring examples. We demonstrated that the approach works and produces good results for all presented examples, even if the time needed for convergence was quite high for some examples.

Here lie already the first possible directions for further investigations. Either by using different frameworks, or by adapting the existing implementations, it could certainly be possible to increase the convergence speed, making the algorithm more practicable to use. Secondly, as already alluded to in the last chapter, not all examples for which trials were undertaken were successfully solved in the course of the thesis. It could be interesting to inspect more closely certain classes of examples if or how the algorithm converges to the theoretical values there. Thirdly, in the literature on Generative Adversarial Networks, many different algorithms and strategies beyond simple gradient-descent-based methods as used in this algorithm have been proposed. Studying in detail their applicability in this setting, and adapting the algorithm accordingly, might lead to faster or more stable convergence. Fourthly, while we have proven statements about theoretical convergence, we have not looked into how to quantify the rate of convergence in terms of network sizes.

Furthermore, at least for the discrete setting, it is feasible to numerically solve the Causal Optimal Transport problem also by using methods of linear programming. It could

be interesting to compare those two methods in terms of convergence speed and stability. Finally, other than the Causal Optimal Transport problem, there exists a number of different, linearly and non-linearly, constrained transport problems with a large variety of applications, such as Martingale Optimal Transport, as described in [8], or Weak Optimal Transport, studied for example in [22] or [6]. It could be interesting to study if the scheme and algorithms proposed in [15] and adapted in this thesis could be similarly adapted for these settings, especially in the case of Weak Optimal Transport, where the linear structure of constraints used to construct the algorithm for the problems investigated in this thesis is lost.

References

[1] M. Abadi et al. "Tensorflow: A system for large-scale machine learning". In: *12th USENIX Symposium on Operating Systems Designand Implementation*. 2016. URL: `www.tensorflow.org`.

[2] B. Acciaio, J. Backhoff-Veraguas, and A. Zalashko. "Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization". In: *Stochastic Processes and their Applications* 130.5 (2020).

[3] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. "Adapted Wasserstein distances and stability in mathematical finance". In: *Finance and Stochastics* 24 (2020).

[4] J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and J. Wiesel. "Estimating Processes in Adapted Wasserstein Distance". In: *Annals of Applied Probability* 32.1 (2022).

[5] J. Backhoff-Veraguas, M. Beiglböck, Y. Lin, and A. Zalashko. "Causal Transport in Discrete Time and Applications". In: *SIAM Journal on Optimization* 27.4 (2017).

[6] J. Backhoff-Veraguas, M. Beiglböck, and G. Pammer. "Existence, duality, and cyclical monotonicity for weak transport costs". In: *Calculus of Variations* 203.58 (2019).

[7] J. Backhoff-Veraguas and M. Huesman. *Stochastic Mass Transfer - Lecture Notes*. 2021.

[8] M. Beiglböck, P. Henry-Labordère, and F. Penkner. "Model-Independent Bounds for Option Prices: A Mass Transport Approach". In: *Finance and Stochastics* 17.3 (2013).

[9] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics. Wiley & Sons, 1999.

[10] C. Bishop. *Pattern Recognition and Machine Learning*. Information Sciences and Statistics. Springer, 2006.

[11] V. I. Bogachev and A. V. Kolesnikov. "The Monge-Kantorovich problem: achievements, connections, and perspectives". In: *Russian Mathematical Survey* 67 (2012).

[12] L. Bottou. "Online Algorithms and Stochastic Approximations". In: *Online Learning and Neural Networks*. Cambridge University Press, 1998.

[13] A. Canterini and D. E. Chang. *Deep Neural Networks in a Mathematical Framework*. Springer Briefs in Computer Science. Springer, 2018.

[14] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals and Systems* 2 (1989).

[15] L. De Gennaro Aquino and S. Eckstein. "MinMax Methods for Optimal Transport and Beyond: Regularization, Approximation and Numerics". In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. 2020.

[16] R. Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[17] S. Eckstein. "Lipschitz neural networks are dense in the set of all Lipschitz functions". In: *arXiv preprint, arXiv:2009.13881* (2020). Arxiv preprint.

[18] C. L. Frenzen, T. Satao, and J. T. Butler. "On the number of segments needed in a piecewise linear approximation". In: *Journal of Computational and Applied mathematics* 234.2 (2010).

[19] A. Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2017.

[20] C. R. Givens and R. M. A. Shortt. "A class of Wasserstein metrics for probability distributions". In: *Michigan Mathematical Journal* 31.2 (1984).

[21] I. Goodfellow et al. "Generative Adversarial Networks". In: *Neural Information Processing Systems Conference*. 2015.

[22] N. Gozlan, C. Roberto, P.-M. Samson, and P. Tetali. "Kantorovich duality for general transport costs and applications". In: *Journal of Functional Analysis* 273.11 (2017).

[23] I. Gulrajani et al. "Improved Training of Wasserstein GANs". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017.

[24] B. Han. "Distributionally robust risk evaluation with causality constraintand structural information". In: *arXiv preprint: arXiv 2203.10571v2* (2022).

[25] K. Hornik. "Approximation Capabilities of Multilayer Feedforward Networks". In: *Neural Networks* 4 (1991).

[26] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2 (1989).

[27] L. Kantorovich. "On the transformation of masses". In: *Doklady Akademii Nauk SSSR* 37 (1942).

[28] D. P. Kingma and J. L. Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations*. 2015.

[29] R. Lasalle. "Causal transport plans and their Monge–Kantorovich problems". In: *Stochastic Analysis and Applications* 36.3 (2018).

[30] G. Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences de Paris* (1781).

[31] G. Peyré and M. Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019).

[32] A. Pinkus. "Approximation theory of the MLP model in neural networks". In: *Acta Numerica* (1999).

[33] R. Remmert. *Theory of complex functions*. Springer, 1991.

[34] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[35] F. Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain". In: *Psychological review* 65.6 (1958).

[36] S. Ruder. "An overview of gradient descent optimization algorithms". In: *arXiv preprint, arXiv:1609.04747* (2016).

[37] W. Rudin. *Real and Complex Analysis*. 3rd ed. McGraw-Hill Book Company, 1987.

[38] A. Ruszczynski. *Nonlinear Optimization*. Princeton University Press, 2006.

[39] F. Santambrogio. *Optimal Transport for Applied Mathematicians - Calculus of Variation, PDEs and Modeling*. Birkhäuser/Springer, 2015.

[40] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1992.

[41] M. Telgarsky. "Benefits of depth in neural networks". In: *Journal of Machine Learning Research: Workshop and Conference Proceedings*. Vol. 49. 1. 2016.

[42] C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

[43] C. Villani. *Topics in Optimal Transportation*. Vol. 58. Graduate Studies in Mathematics. American Mathematical Society, 2003.

[44] Y. Xie et al. "On scalable and efficient computation of large scale Optimal Transport". In: *International Conference on Machine Learning*. 2019.

[45] T. Xu, L. K. Wenliang, M. Munn, and B. Acciaio. "COT-GAN: Generating Sequential Datavia Causal Optimal Transport". In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. 2020.