Constrained Approximate Optimal Transport Maps

Eloi Tanguy[1], Agnès Desolneux[2], and Julie Delon[1]

[1]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France
[2]Centre Borelli, CNRS and ENS Paris-Saclay, F-91190 Gif-sur-Yvette, France

12th March 2025

Abstract

We investigate finding a map g within a function class G that minimises an Optimal Transport (OT) cost between a target measure ν and the image by g of a source measure μ. This is relevant when an OT map from μ to ν does not exist or does not satisfy the desired constraints of G. We address existence and uniqueness for generic subclasses of L-Lipschitz functions, including gradients of (strongly) convex functions and typical Neural Networks. We explore a variant that approaches a transport plan, showing equivalence to a map problem in some cases. For the squared Euclidean cost, we propose alternating minimisation over a transport plan π and map g, with the optimisation over g being the L^2 projection on G of the barycentric mapping $\overline{\pi}$. In dimension one, this global problem equates the L^2 projection of $\overline{\pi^*}$ onto G for an OT plan π^* between μ and ν, but this does not extend to higher dimensions. We introduce a simple kernel method to find g within a Reproducing Kernel Hilbert Space in the discrete case. We present numerical methods for L-Lipschitz gradients of ℓ-strongly convex potentials, and study the convergence of Stochastic Gradient Descent methods for Neural Networks. We finish with an illustration on colour transfer, applying learned maps on new images, and showcasing outlier robustness.

Table of Contents

1 Introduction

Let μ and ν denote two probability distributions on two (potentially different) measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. Many problems in applied fields can be written under the form

$$\inf_{g \in G} \mathcal{D}(g\#\mu, \nu), \tag{1}$$

where $\#$ denotes the *push-forward* operation[1], $\mathcal{D}$ is a non-negative discrepancy (such as a distance metric or a ϕ-divergence) measuring the similarity between $g\#\mu$ and ν, and G is a set of acceptable functions from $\mathcal{X}$ to $\mathcal{Y}$. Under appropriate assumptions on $\mathcal{D}$, this problem can be interpreted as a projection of ν on the set $G\#\mu := \{g\#\mu, g \in G\}$ for the discrepancy $\mathcal{D}$. In this paper, we focus on cases where ν cannot be written as $g\#\mu$ for $g \in G$.[2]

In the highly popular field of generative modelling, the target distribution is usually an empirical distribution composed of m samples, $\nu = \frac{1}{m}\sum_{i=1}^{m} \delta_{x_i}$, μ is an easy-to-sample latent distribution (for instance a Gaussian distribution), and the set $G = \{g_\theta,\ \theta \in \Theta\}$ generally denotes functions represented by a specific neural network architecture. The goal is to find the parameter θ such that $\mu_\theta := g_\theta\#\mu$ fits ν as well as possible. Models taking this form are often called push-forward generative models [45], and include Variational Auto-Encoders (VAEs) [31], Generative Adversarial Networks (GANs) [28], Normalising flows [43] and even Diffusion Models [49], which can be reinterpreted as indirect push-forward generative models [45]. In these works, the discrepancy $\mathcal{D}$ is often chosen as the Kullback-Leibler divergence, as it is the case for traditional GANs and VAEs, or as the Wasserstein distance, like in Wasserstein-GANs [5]. The discrepancy $\mathcal{D}(g_\theta\#\mu, \nu)$ is minimised in θ, for instance by using sophisticated versions of stochastic gradient descent. In such problems, it is clear that $g_\theta\#\mu$ does not target exactly ν, since it would mean that the model has only learned to reproduce existing samples, and not to create new ones. This is possible because the expressivity of neural networks is limited, but also because the training steps usually impose regularity properties on g_θ and constrain its Lipschitz constant in order to increase its robustness [55, 25] or stabilise its training [36]. It is therefore natural to wonder to what extent the optimisation of such discrepancies with regularity constraints on the set of functions G is well-posed, depending on the choice of $\mathcal{D}$, and what this means in practice.

In an Euclidean setting, another example of Eq. (1) appears when we need to compare two distributions μ and ν potentially living in spaces of different dimensions, or when invariance to geometric transformations is required (for problems such as shape matching or word embedding). In such cases, it is usual to choose G as a well chosen set of linear or affine embeddings (such as matrices in the Stiefel manifold if the space dimension is different between $\mathcal{X}$ and $\mathcal{Y}$). For instance, this idea underpins several sets of works introducing global invariances in optimal transport [2, 46].

In both of the previous examples, G is parametrised by a set Θ of parameters which is potentially extremely large (for neural networks) but of finite dimension. Alternatively, the set of functions G can be much more complex and characterised by regularity or convexity assumptions, the problem becoming non-parametric. This is typically the case in the field of optimal transport [54, 47]. Given μ and ν probability measures on respective Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, Monge's Optimal Transport consists in finding a Transport map T such that $T\#\mu = \nu$ and which minimises a given displacement

[1] The image measure $g\#\mu$ is defined as the law of $g(X)$ for X a random variable of law μ, or more abstractly by $g\#\mu(B) = \mu(g^{-1}(B))$ for any Borel set $B \subset \mathcal{Y}$.

[2] Obviously, if ν belongs to $G\#\mu$, the problem is trivial and the infimum in Eq. (3) is 0.

cost. From a theoretical standpoint, the existence of (unconstrained) Monge maps has been widely studied [11, 27, 42], under regularity assumption for μ. When there is no map T such that $T\#\mu = \nu$ (for example, if μ is discrete and if ν is not), or when the map solution does not meet the regularity requirements for some given practical application, it makes sense instead to solve problems of the form of Eq. (1), with $\mathcal{D}$ a Wasserstein distance and G a set of functions with acceptable regularity. For instance, as studied in [37], G can be composed of functions $g = \nabla\phi$ with ϕ ℓ-strongly-convex with an L-Lipschitz gradient. For cases where μ is discrete, this formulation also overcomes a classic shortcoming of numerical optimal transport approaches, which usually compute solutions which are only defined on the support of μ. If a machine learning algorithm requires the computation of the transport of new inputs, the map must be either recomputed, or an approximation of the previous map must be defined outside of the support of μ. Several solutions have been proposed in the literature to solve this problem [16, 7, 32, 37, 41, 48, 38], and some of them [32, 37] consists in solving Eq. (1) with an appropriate set of functions G. Consistency and asymptotic properties of such estimators are also the subject of several of these works [16, 32, 29].

For the sake of legibility and to avoid excessive technicality, we focus on the case where the target space is $\mathbb{R}^d$, however it is possible to extend our considerations to a target space $\mathcal{Y}$ which is a Polish space verifying the Heine-Borel property (i.e. that any bounded and closed set is a compact set), which in particular allows the case where $\mathcal{Y}$ is a connected and complete Riemannian manifold (in which case the Heine-Borel property follows from the Hopf-Rinow Theorem, see [19] Theorem 2.8). Similarly, the problem naturally extends to the case where the codomain of the maps g and the target measure ν are different spaces $\mathcal{Y}, \mathcal{Y}'$.

OT discrepancies. In this paper, we focus on problems of the form Eq. (1) when $\mathcal{D}$ is chosen as an optimal transport discrepancy for a general ground cost c. We recall that if $\mathcal{X}$ and $\mathcal{Y}$ are two Polish spaces, the Optimal Transport cost between two measures $\nu_1 \in \mathcal{P}(\mathcal{X})$ and $\nu_2 \in \mathcal{P}(\mathcal{Y})$ for a ground cost function $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ is defined by the following optimisation problem

$$\mathcal{T}_c(\nu_1, \nu_2) = \min_{\pi \in \Pi(\nu_1, \nu_2)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y), \tag{2}$$

where $\Pi(\nu_1, \nu_2)$ is the set of probability measures on $\mathcal{X} \times \mathcal{Y}$ whose first marginal is ν_1 and second marginal is ν_2[3]. Given this method of quantifying the discrepancy between $g\#\mu$ and ν, Eq. (1) becomes

$$\inf_{g \in G} \mathcal{T}_c(g\#\mu, \nu). \tag{3}$$

In the case where the source measure is discrete and the target measure is absolutely continuous, the Optimal Transport problem in Eq. (3) is said to be semi-discrete, and has a slightly more explicit expression (see [34] for a course on the matter). If we suppose in addition that $c(x, y) = \|x - y\|_2^2$ and that the source measure weights are uniform ($a_i = 1/n$), then Eq. (3) is a constrained version the Optimal Uniform Quantization problem studied thoroughly in [33].

Existence of minimisers. An important question regarding this optimisation problem concerns the existence of minimisers, depending on the ground cost c and the set of functions G. While numerous works in the literature have focused on the convergence of optimisation algorithms (such as stochastic gradient descent) to critical points for this kind of problem [24], the existence of minimisers has surprisingly been little studied. We derive in Theorems 2.3 and 2.4 generic conditions to ensure existence of such minimisers in G, and show counter-examples when these conditions are not met. We also show that these conditions are satisfied for two classes of functions, namely classes of L-Lipschitz functions which can be written as gradient of l-strongly convex functions (recovering a result shown in [37] as a particular case of Theorem 2.4), and classes of neural networks with Lipschitz activation

[3] The fact that the minimum is attained is a consequence of the direct method of calculus of variations (see [47], Theorem 1.7). The value of $\mathcal{T}_c(\nu_1, \nu_2)$ may be $+\infty$, but a sufficient condition for $\mathcal{T}_c(\nu_1, \nu_2) < +\infty$ ([54], Remark 5.14) is that

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\nu_1(x) \mathrm{d}\nu_2(y) < +\infty.$$

functions. We also discuss uniqueness of the solutions, which is usually not satisfied, and remains a difficult question without strong assumptions on the set of functions G.

Approximating a coupling. In the field of optimal transport, a particular setting where Eq. (3) is interesting is when we have access to a non deterministic coupling π solution of a regularised version of a optimal transport between two probability measures μ and ν. For instance, the entropic optimal transport [40], or the mixture Wasserstein formulation [18] both yield optimal plans π which cannot be trivially written as optimal maps between μ and ν. For some applications, it can be interesting to approximate π by another transport plan supported by the graph of a function with possible additional regularity assumptions. This can be done by approximating π by $(I, g)\#\mu$, with specific regularity properties on g, which is a particular case of Eq. (3), replacing ν by π and G by the set $H := \{(I, g), g \in G\}$. In this specific setting, we show in Section 2.7 under which conditions on the ground cost c the solutions of this problem between plans are equivalent to solutions of the original Eq. (3) when $\pi \in \Pi(\mu, \nu)$. Numerical approaches seeking maps that approach barycentric projections have been studied in [48, 38].

Alternate minimisation. Under appropriate assumptions, Eq. (3) can be rewritten as a minimisation problem over $\pi \in \Pi(\mu, \nu)$ and $g \in G$:

$$\min_{g \in G} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(g(x), y) \mathrm{d}\pi(x, y). \tag{4}$$

This naturally leads to consider Eq. (3) as an alternate minimisation problem, that we study in Section 3 in the Euclidean case when $c(x, y) = \|x-y\|_2^2$. More precisely, we show that Eq. (3) is strongly linked to the barycentric projection problem: when π is fixed, the solution g minimising Eq. (4) can be reinterpreted as the L^2-projection of the barycentric projection of π on the set G. In the one-dimensional case, when G is a subclass of increasing functions, this yields an explicit solution to the problem (as it was shown in [37] in a more specific case), and we show that this explicit solution does not hold in dimension larger than 1 by presenting a counter-example.

Outline of the paper. In this work, we address problem Eq. (3) for large classes of functions G. In Section 2, we define the problem and establish general conditions for the existence of solutions, exploring examples involving gradients of convex functions and neural networks. Section 3 examines the link between Eq. (3) and a constrained barycentric projection problem, demonstrating an explicit solution in one dimension and providing a counterexample in higher dimension. Section 4 focuses on practical numerical methods to solve the optimisation problems for Lipschitz gradients of strongly convex potentials, kernel methods and Neural Networks. We conclude with an illustration on colour transfer.

2 A Constrained Approximate Transport Map Problem

2.1 Problem Definition

We consider $(\mathcal{X}, d_{\mathcal{X}})$ a locally compact Polish space, and $\mu \in \mathcal{P}(\mathcal{X})$ a probability measure on $\mathcal{X}$. Our objective is to find a map $g : \mathcal{X} \longrightarrow \mathbb{R}^d$ verifying the constraint $g \in G$ for some class of functions $G \subset (\mathbb{R}^d)^{\mathcal{X}}$, such that the image measure $g\#\mu$ is "close" to a fixed probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$, in the sense of Eq. (3).

Applying the definition of $\mathcal{T}_c$ directly (Eq. (2)) yields the following expression for Eq. (3):

$$\inf_{g \in G} \min_{\pi \in \Pi(g\#\mu,\nu)} \int_{\mathcal{X}\times\mathbb{R}^d} c(x, y) \mathrm{d}\pi(x, y). \tag{5}$$

The optimisation variable g acts on the set of constraints of the Optimal Transport problem, however thanks to a well-known "change of variables" result ([20] Lemmas 1 and 2 for a reference), we will be able to reformulate Eq. (5). In the following, we shall denote by $\Pi_c^*(\nu_1, \nu_2)$ the set of minimisers of the optimal transport problem Eq. (2) between two measures ν_1 and ν_2.

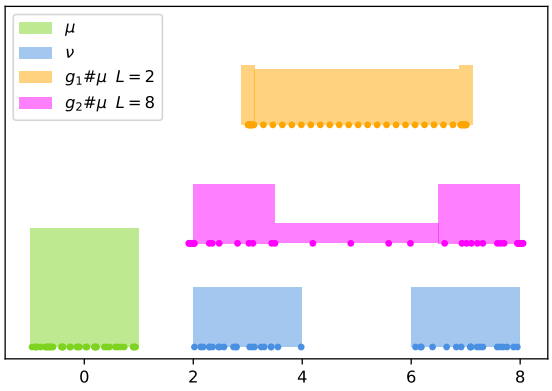
Lemma 2.1. *([20], Lemmas 2.6 and 2.7) Let $\mathcal{X}, \mathcal{Y}, \mathcal{X}', \mathcal{Y}'$ be Polish spaces. Let $g : \mathcal{X} \longrightarrow \mathcal{X}'$ and $h : \mathcal{Y} \longrightarrow \mathcal{Y}'$ two measurable maps and let $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$. Consider two costs $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ and $c' : \mathcal{X}' \times \mathcal{Y}' \longrightarrow \mathbb{R}$ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$, $c(x, y) = c'(g(x), h(y))$.*

- *For any $\gamma' \in \Pi(g\#\mu, h\#\nu)$, there exists $\gamma \in \Pi(\mu, \nu)$ such that $\gamma' = (g, h)\#\gamma$.*
- *We have $\Pi^*_{c'}(g\#\mu, h\#\nu) = (g, h)\#\Pi^*_c(\mu, \nu)$.*

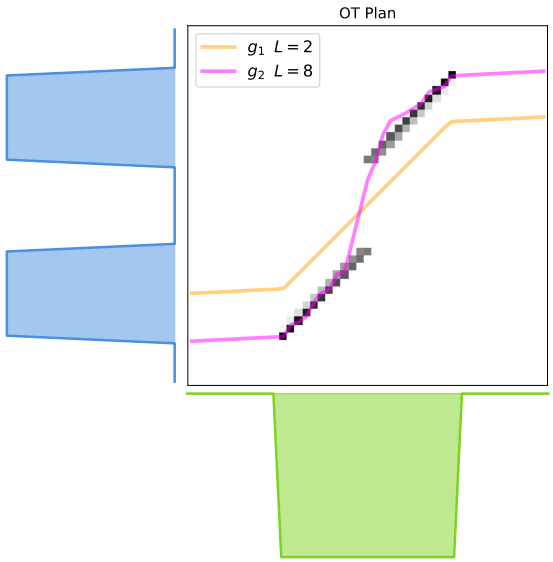
Using Lemma 2.1, the energy of the map problem Eq. (3) can be written as follows:

$$\mathcal{T}_c(g\#\mu, \gamma) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathbb{R}^d} c(g(x), y) \mathrm{d}\pi(x, y). \tag{6}$$

In our study of the map problem Eq. (3), we will consider classes G that are a subset of the L-Lipschitz functions. The first reason is that with unbounded Lipschitz constants, the problem may not have a solution, as we shall see in Section 2.5. Moreover, there are multiple practical considerations that lead to choosing functions with an upper-bounded Lipschitz constant. To begin with, numerous practical models enforce this condition, such as Wasserstein GANs [5], and diffusion models [49] (see also [45] Appendix S2), furthermore most neural networks are Lipschitz (since typical non-linearities are chosen as Lipschitz), and the control of the Lipschitz constant is often used as a regularisation method [55]. In Fig. 1, we illustrate a solution of the map problem using numerical methods introduced in Section 4.2, for two different values of L (the Lipschitz constant of the maps g).



(a) Illustration of the source, target and image measures using their samples and (approximate) densities.



(b) Illustration of map solutions and comparison with the (discontinuous) Optimal Transport coupling.

Figure 1: Illustration of solutions of maps problems (Eq. (3)) on a toy dataset with a source measure $\mu = \mathcal{U}([-1, 1])$ and a target measure $\nu = \frac{1}{2}\mathcal{U}([2, 4]) + \frac{1}{2}\mathcal{U}([6, 8])$. The two solutions are respectively $L = 2$ and $L = 8$ Lipschitz.

2.2 Existence of a Solution

To formulate an existence result, we shall apply the direct method of calculus of variations, which requires a technical condition on the stability of the class of functions G with respect to certain limits. To formulate this condition, we will introduce the notion of closedness of a class of continuous functions with respect to the compact-open topology. By [30] Chapter 7 Theorem 11, in our setting, this topology is equivalent to the topology of uniform convergence on compact sets, which allows us to formulate Definition 2.2 in terms of local uniform convergence.

Definition 2.2. *We say that a set of functions $G \subset (\mathbb{R}^d)^{\mathcal{X}}$ is **closed for the compact-open topology** if there exists a sequence $(\mathcal{K}_m)$ of compact sets of $\mathcal{X}$ verifying $\cup_m \mathcal{K}_m = \mathcal{X}$ such that: for any sequence $(g_n)_{n\in\mathbb{N}} \in G^{\mathbb{N}}$ such that for all m, $(g_n|_{\mathcal{K}_m})_{n\in\mathbb{N}}$ converges uniformly towards a function $g_{\mathcal{K}_m} : \mathcal{K}_m \longrightarrow \mathbb{R}^d$, there exists $g \in G$ such that $g|_{\mathcal{K}_m} = g_{\mathcal{K}_m}$ for all m.*

One can understand this condition as a form of "local uniform closedness" of the class G.

Theorem 2.3. *Let $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ be a lower semi-continuous cost function, $\mu \in \mathcal{P}(\mathcal{X})$ be a probability measure on a locally compact Polish space $(\mathcal{X}, d_{\mathcal{X}})$, and $\nu \in \mathcal{P}(\mathbb{R}^d)$. Assume that*

*i) **(Coercive cost)** There exists $\eta : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ non-decreasing and such that $\eta(t) \xrightarrow[t\longrightarrow+\infty]{} +\infty$, and $\alpha \in \mathbb{R}$ such that $\forall y, y' \in \mathbb{R}^d$, $c(y, y') \geq \alpha + \eta(\|y - y'\|_2)$ and $\int \eta(\|\cdot\|_2)\mathrm{d}\nu < +\infty$;*

*ii) **(Lipschitzness and Closedness of G)** G is a subset of the space of L-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}^d$, that is closed for the compact-open topology (see Definition 2.2);*

*iii) **(Problem finiteness)** There exists $g \in G$ such that $\mathcal{T}_c(g\#\mu, \nu) < +\infty$.*

Then the problem $\underset{g\in G}{\operatorname{argmin}}\ \mathcal{T}_c(g\#\mu, \nu)$ has a solution.

Proof. — *Step 1*: Defining a minimising sequence.

We introduce the notation $J(g) := \mathcal{T}_c(g\#\mu, \nu)$ for convenience, and J^* the problem value, which is finite by Assumption iii). Consider a minimising sequence $(g_n)_{n\in\mathbb{N}} \in G^{\mathbb{N}}$ such that

$$\forall n \in \mathbb{N},\ J(g_n) \leq J^* + 2^{-n}.$$

— *Step 2*: Bounding g_n.

First, we fix $n \in \mathbb{N}$, and take $a \in \mathcal{X}$ in the support of μ and $r > 0$, then set $A := B_{d_{\mathcal{X}}}(a, r)$ the ball of centre a and radius r for the distance $d_{\mathcal{X}}$, so that $\mu(A) > 0$. The transport problem has a solution since c is lower semi-continuous ([47] Theorem 1.7). We introduce $\pi_n^* \in \Pi(\mu, \nu)$ optimal for the OT cost $\mathcal{T}_c(g_n\#\mu, \nu)$. We lower-bound:

$$J(g_n) \geq \int_{A\times\mathbb{R}^d} c(g_n(x), y)\mathrm{d}\pi_n^*(x, y) \geq \int_{A\times\mathbb{R}^d} \eta(\|g_n(x) - y\|_2)\mathrm{d}\pi_n^*(x, y) + \alpha\mu(A).$$

To separate variables, we will use an elementary inequality: let $z, w \in \mathbb{R}^d$, the triangle inequality yields $\|z\|_2 \leq 2\max(\|w - z\|_2, \|w\|_2)$, applying the non-decreasing and non-negative function η provides $\eta(\|z\|_2/2) \leq \max(\eta(\|w - z\|_2), \eta(\|w\|_2)) \leq \eta(\|w - z\|_2) + \eta(\|w\|_2)$. Finally, we have

$$\forall w, z \in \mathbb{R}^d,\ \eta(\|w - z\|_2) \geq \eta(\|z\|_2/2) - \eta(\|w\|_2). \tag{7}$$

By assumption, we remind that $\int \eta(\|\cdot\|_2)\mathrm{d}\nu < +\infty$, and resume lower-bounding using Eq. (7) with $w := y$ and $z := g_n(x)$:

$$J(g_n) \geq \int_A \eta\left(\frac{\|g_n(x)\|_2}{2}\right)\mathrm{d}\mu(x) - \mu(A)\int \eta(\|\cdot\|_2)\mathrm{d}\nu + \alpha\mu(A).$$

Let $x \in A$, we apply again Eq. (7) with $w := (g_n(a) - g_n(x))/2$ and $z := g_n(a)/2$:

$$\eta(\|g_n(x)\|_2/2) \geq \eta(\|g_n(a)\|_2/4) - \eta(\|g_n(a) - g_n(x)\|_2/2) \geq \eta(\|g_n(a)\|_2/4) - \eta(Lr/2),$$

where the second inequality comes from the fact that g_n is L-Lipschitz, $d_{\mathcal{X}}(x, a) \leq r$ and that η is non-decreasing. Gathering our inequalities leads to the following lower-bound:

$$J^* + 1 \geq J(g_n) \geq \mu(A)\left(\eta(\|g_n(a)\|_2/4) - \eta(Lr/2) - \int \eta(\|\cdot\|_2)\mathrm{d}\nu + \alpha\right).$$

This implies that there exists $M > 0$ independent of n such that $\|g_n(a)\|_2 \leq M$. (Since by coercivity of η, the right-hand side of the equation above would tend to $+\infty$ if $\|g_n(a)\|_2 \xrightarrow[n\longrightarrow+\infty]{} +\infty$).

— *Step 3*: Applying Arzelà-Ascoli's Theorem.

For $n \in \mathbb{N}$, we use the upper-bound from Step 2 and the fact that each g_n is L-Lipschitz:

$$\forall x \in \mathcal{X},\ \|g_n(x)\|_2 \leq M + Ld_{\mathcal{X}}(x, a),$$

which shows that $\forall x \in \mathcal{X},\ \{g_n(x),\ n \in \mathbb{N}\}$ has compact closure in $\mathbb{R}^d$. The sequence (g_n) is equi-Lipschitz and thus equi-continuous, and is closed for the compact-open topology (Definition 2.2) by assumption. By Arzelà-Ascoli's theorem (as stated in [30], Chapter 7, Theorem 17), we can choose $\beta : \mathbb{N} \longrightarrow \mathbb{N}$ an extraction such that $g_{\beta(n)} \xrightarrow[n\longrightarrow+\infty]{} g$ locally uniformly on $\mathcal{X}$, for a certain function $g \in G$.

— *Step 4*: Showing that the limit g is optimal.

First, the sequence $(g_{\beta(n)}\#\mu)$ converges weakly towards the probability measure $g\#\mu$: take a continuous and compactly supported test function $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}$, the dominated convergence theorem shows that

$$\int_{\mathcal{X}} \phi \circ g_{\beta(n)} \mathrm{d}\mu \xrightarrow[n\longrightarrow+\infty]{} \int_{\mathcal{X}} \phi \circ g \mathrm{d}\mu,$$

where convergence of the integrands is ensured by the point-wise convergence of $(g_{\beta(n)})$, and domination by $\|\phi\|_\infty$ suffices. Since c is lower semi-continuous, the OT cost is itself lower semi-continuous for the weak convergence of measures (see [3] Theorem 2.6), we obtain the following inequality:

$$\liminf_{n\longrightarrow+\infty} J(g_{\beta(n)}) \geq J(g),$$

where J was introduced in Step 1, where we also chose g_n such as $J(g_n) \leq J^* + 2^{-n}$, thus we conclude $J^* \geq J(g)$, hence g is optimal.

□

Theorem 2.3 can be extended to the case where the regularity of the functions of G is only assumed *on a partition* of $\mathcal{X}$. Note that to avoid pathological ambiguity and unnecessary complications, we will consider partitions whose borders have no mass for μ, such that the problem objective can be split according to the partition.

Theorem 2.4. *Let $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ be a continuous cost function, a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ on a locally compact Polish space $(\mathcal{X}, d_{\mathcal{X}})$, and $\nu \in \mathcal{P}(\mathbb{R}^d)$. Consider $(E_i)_{[\![1,K]\!]}$ a partition of $\mathcal{X}$ such that for every $i \in [\![1, K]\!]$, $\mu(\partial E_i) = 0$. Under the same conditions as Theorem 2.3, and replacing assumption ii) by*

ii') The class of functions $G \subset (\mathbb{R}^d)^{\mathcal{X}}$ is of the form

$$G = \left\{ g : \mathcal{X} \longrightarrow \mathbb{R}^d \mid \forall i \in [\![1, K]\!],\ g|_{\mathring{E}_i} = g_i,\ g_i \in G_i \right\},$$

where for every $i \in [\![1, K]\!]$, the set of functions $G_i \subset (\mathbb{R}^d)^{\mathring{E}_i}$ is a subset of the space of L-Lipschitz functions from $\mathring{E}_i$ to $\mathbb{R}^d$, that is closed for the compact-open topology (see Definition 2.2),

then the problem $\underset{g\in G}{\operatorname{argmin}}\ \mathcal{T}_c(g\#\mu, \nu)$ has a solution.

Proof. We shall follow closely the proof of Theorem 2.3, and point out the technical differences. We introduce a minimising sequence exactly identically to Step 1. The computations from Step 2 can be done verbatim, choosing instead $A_i \subset \mathring{E}_i$, and concluding $\|g_n(a_i)\|_2 \leq M_i$ for a fixed $a_i \in A_i$.

Step 3 is then done separately on each $\mathring{E}_i$, yielding extractions (β_i) such that each $g_{\beta_i(n)}$ converges locally uniformly on $\mathring{E}_i$ towards a function $g_i \in G_i$. Considering the extraction $\beta := \beta_1 \circ \cdots \circ \beta_K$, we have for all $i \in [\![1, K]\!]$ the uniform convergence of $(g_{\beta(n)})$ towards $g \in G$ on all compact sets of $\mathring{E}_i$.

Finally, Step 4 is done likewise to Theorem 2.3, with the technicality that since $\mu(\partial E_i) = 0$, the pointwise convergence of $(g_{\beta(n)})$ towards g at each point of $\mathring{E}_i$ suffices to show that $g_{\beta(n)}(x) \xrightarrow[n\longrightarrow+\infty]{} g(x)$ for μ-almost-every $x \in \mathcal{X}$, which yields the convergence in law $g_{\beta(n)}\#\mu \xrightarrow[n\longrightarrow+\infty]{w} g\#\mu$. The rest follows verbatim. □

In Remarks 2.5 and 2.6, we present some natural extensions of Theorems 2.3 and 2.4, which we kept separate for legibility.

Remark 2.5. *The existence results of Theorems 2.3 and 2.4 also hold if the objective functional is changed into a regularised version*

$$J(g) = \mathcal{T}_c(g\#\mu, \nu) + R(g),$$

where $R : G \longrightarrow \mathbb{R}_+ \cup \{+\infty\}$ is lower semi-continuous with respect to uniform local convergence. One also has to assume that there still exists $g \in G$ such that the new cost J is finite. The proofs can be written almost identically: in Step 1, it suffices to lower-bound $R(g_n) \geq 0$, and in Step 4, one obtains $\liminf J(g_{\beta(n)}) \geq J(g)$ thanks to the lower semi-continuity of R.

Remark 2.6. *Condition i) on c can be generalised to the case where the target space $\mathbb{R}^d$ is instead a Polish space $\mathcal{Y}$ verifying the Heine-Borel property (i.e. all closed and bounded sets are compact), in which case Condition i) can be replaced with the condition that $c(\cdot, y_0)$ be **proper**, which is to say that its preimage by any compact set $S \subset \mathbb{R}_+$ is a compact set of $\mathcal{Y}$. This property would be used in Step 2 to show that $g_n(a) \in C$ for some compact set $C \subset \mathcal{Y}$ independent of n, then in Step 3, we would use the Lipschitz property of g_n and the triangle inequality on $d_{\mathcal{Y}}$ to show that $\forall x \in \mathcal{K},\ g_n(x) \in \overline{B}_{\mathcal{Y}}(y_0, Lr + d_{\mathcal{Y}}(y_0, C))$, for a compact set $\mathcal{K} \subset \mathcal{X}$ of diameter r and $y_0 \in \mathcal{Y}$. This would show that for each $x \in \mathcal{K}$, the set $\{g_n(x)\}_{n\in\mathbb{N}}$ is pre-compact in $\mathcal{Y}$, and allow one to apply Arzelà-Ascoli likewise.*

A natural context for Optimal Transport is the case where the ground cost is of the from $c(x, y) = \|x - y\|^p$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$ and $p \geq 1$. In Proposition 2.7, we show that such costs verify the assumptions to our existence theorems.

Proposition 2.7. *Cost functions of the form $c(x, y) := \|x - y\|^p$, where $p > 0$ and $\|\cdot\|$ is a norm on $\mathbb{R}^d$ satisfy Assumption i) of Theorems 2.3 and 2.4, as long as $\nu \in \mathcal{P}_p(\mathbb{R}^d)$.*

Proof. Take $\eta := t \longmapsto (Kt)^p$, where $K > 0$ is provided by the norm equivalence inequality $\|\cdot\| \geq K\|\cdot\|_2$. □

2.3 Function Class Example: Gradients of Convex Functions

An interesting class of functions G to optimise over is the set of L-Lipschitz functions that are gradients of (ℓ-strongly) convex functions. Indeed, this can be seen as a regularising assumption, and was studied in [37] for the cost $c(x, y) = \|x - y\|_2^2$. We shall see in Proposition 2.9 that classes of such functions on *arc-connected* partitions verify the conditions of our existence result Theorem 2.4. In particular, [37] Definition 1 (which states existence, with a simplified proof due to lack of space) is a consequence of Theorem 2.4. Before this result, we will present a technical lemma on arc-connectedness. In this paper, we will say that a set $A \subset \mathbb{R}^d$ is *arc-connected* if any pair of points of A can be joined by a Lipschitz curve contained in A.

Lemma 2.8. *Let $\mathcal{O}$ be an arc-connected open set of $\mathbb{R}^d$. There exists $(C_k)_{k\in\mathbb{N}}$ a sequence of arc-connected compact sets such that $\forall k \in \mathbb{N},\ C_k \subset C_{k+1}$ and $\bigcup_{k\in\mathbb{N}} C_k = \mathcal{O}$.*

Proof. Consider the collection $(\overline{B}(q,r_q))_{q\in\mathcal{O}\cap\mathbb{Q}^d}$ where for each $q\in\mathcal{O}\cap\mathbb{Q}^d$, we take $r_q>0$ such that $\overline{B}(q,r_q)\subset\mathcal{O}$. Using a bijection between $\mathbb{N}$ and $\mathcal{O}\cap\mathbb{Q}^d$, we can introduce sequences $(q_k)\in(\mathcal{O}\cap\mathbb{Q}^d)^{\mathbb{N}}$ and $(r_k)\in(0,+\infty)^{\mathbb{N}}$ such that the sequence of the $A_k:=\overline{B}(q_k,r_k)$ enumerates the previous collection. The sequence (A_k) is made of compact arc-connected sets and verifies $\mathcal{O}=\cup_k A_k$. We can now defined recursively the sequence (C_k) by $C_0:=A_0$ and $C_{k+1}:=C_k\cup A_{k+1}\cup w_k([0,1])$, where for $k\in\mathbb{N}$, $w_k:[0,1]\longrightarrow\mathcal{O}$ is a Lipschitz curve between q_k and q_{k+1} contained in $\mathcal{O}$ (which exists by assumption on $\mathcal{O}$). By induction, the sequence (C_k) verifies the desired properties. □

We now have the technical tools to prove that L-Lipschitz functions that are gradients of (ℓ-strongly) convex functions verifies the local convergence stability assumption of Theorem 2.4 on partitions of $\mathbb{R}^d$.

Proposition 2.9. *Consider $\mathcal{X}:=\mathbb{R}^d$, and a partition $\mathcal{E}:=(E_i)_{[\![1,K]\!]}$, where each $\mathring{E}_i$ is arc-connected. Let $0\leq\ell\leq L$, the set of functions*

$$\mathcal{F}_{\mathcal{E},L,\ell}:=\Big\{g:\mathbb{R}^d\longrightarrow\mathbb{R}^d\mid\forall i\in[\![1,K]\!],\ g|_{\mathring{E}_i}\ L-\textit{Lipschitz};\ g|_{\mathring{E}_i}=\nabla\varphi_i,\ \varphi_i\in\mathcal{C}^1(\mathring{E}_i,\mathbb{R}),\ \varphi_i\ \ell-\textit{strongly convex}\Big\}$$

verifies Assumption ii') of Theorem 2.4.

Proof. Let $i\in[\![1,K]\!]$, and define for notational convenience $\mathcal{U}:=\mathring{E}_i$. We want to show that the set of functions

$$G:=\Big\{g:\mathcal{U}\longrightarrow\mathbb{R}^d\ L-\text{Lipschitz}\mid g=\nabla\varphi,\ \varphi\in\mathcal{C}^1(\mathcal{U},\mathbb{R}),\ \varphi\ \ell-\text{strongly convex}\Big\}$$

is closed for the compact-open topology (Definition 2.2). By Lemma 2.8, since $\mathcal{U}$ is open and arc-connected, we can choose an increasing sequence of arc-connected compact sets $\mathcal{K}_m\subset\mathcal{U}$ such that $\cup_m\mathcal{K}_m=\mathcal{U}$. We fix $a\in\mathcal{K}_0$.

Take a sequence $(g_n)_{n\in\mathbb{N}}\in G^{\mathbb{N}}$ such that for each $m\in\mathbb{N}$, $g_n|_{\mathcal{K}_m}$ converges uniformly to some function $h_m\in\mathcal{C}^0(\mathcal{K}_m,\mathbb{R}^d)$. We will show that there exists $g\in G$ that coincides with h_m on each $\mathcal{K}_m$. Regarding the Lipschitz constraint, by point-wise convergence, each function h_m is L-Lipschitz.

For any $n\in\mathbb{N}$, since $g_n\in G$, we can introduce an ℓ-strongly convex function $\varphi_n\in\mathcal{C}^1(\mathcal{U},\mathbb{R})$ such that $g_n=\nabla\varphi_n$. Since φ_n can be chosen up to an additive constant, we can assume $\varphi_n(a)=0$. We study the point-wise convergence of (φ_n) on $\mathcal{K}_m$ for $m\in\mathbb{N}$ fixed, so we fix $x\in\mathcal{K}_m$. Since $\mathcal{K}_m$ is arc-connected, we can choose $w:[0,1]\longrightarrow\mathcal{K}_m$ a Lipschitz curve such that $w(0)=a$ and $w(1)=x$. Noticing that for almost-every $t\in[0,1]$, $\frac{\mathrm{d}}{\mathrm{d}t}\varphi_n(w(t))=\langle\nabla\varphi_n(w(t)),\dot{w}(t)\rangle$ and using $\varphi_n(a)=0$, we write (by absolute continuity of $\varphi_n\circ w$):

$$\varphi_n(x)=\int_0^1\langle\nabla\varphi_n(w(t)),\dot{w}(t)\rangle\mathrm{d}t\xrightarrow[n\longrightarrow+\infty]{}\int_0^1\langle h_m(w(t)),\dot{w}(t)\rangle\mathrm{d}t=:\psi_m(x),$$

where the convergence is obtained by the dominated convergence theorem.

Our objective is now to prove that ψ_m is $\mathcal{C}^1$-smooth on $\mathring{\mathcal{K}}_m$, and that $\nabla\psi_m=h_m$. Let $x\in\mathring{\mathcal{K}}_m$, $v\in\mathbb{R}^d$ and $\delta>0$ such that $\forall t\in[-\delta,\delta]$, $x+tv\in\mathring{\mathcal{K}}_m$. For $n\in\mathbb{N}$ and $t\in[-\delta,\delta]$, let $f_n(t):=\varphi_n(x+tv)$. We have shown that the sequence (f_n) converges pointwise to $f:=t\longmapsto\psi_m(x+tv)$. Furthermore, by convergence of (g_n), the derivative sequence $f_n'=t\longmapsto\langle\nabla\varphi_n(x+tv),v\rangle$ converges uniformly on $[-\delta,\delta]$ to $t\longmapsto\langle h_m(x+tv),v\rangle$. A standard calculus theorem then shows that f is differentiable on $(-\delta,\delta)$, with $f'(t)=\frac{\mathrm{d}}{\mathrm{d}t}\langle h_m(x+tv),v\rangle$. In particular, by setting $t=0$ we have shown that the directional derivative $D_v\psi_m(x)$ exists and has the value $\langle h_m(x),v\rangle$. Since h_m is continuous (we saw that it is Lipschitz), this shows that ψ_m is of class $\mathcal{C}^1$, with $\nabla\psi_m=g_m$ on $\mathring{\mathcal{K}}_m$.

For $x\in\mathcal{U}$, letting $m:=\min\{m\in\mathbb{N}:x\in\mathcal{K}_m\}$, we define $\psi(x):=\psi_m(x)$, which is well-defined since $x\in\mathcal{K}_m$. For $m<m'$, since $\mathcal{K}_m\subset\mathcal{K}_{m'}$, we have $\psi_{m'}|_{\mathcal{K}_m}=\psi_m$, as a consequence, for any $m\in\mathbb{N}$, $\psi|_{\mathcal{K}_m}=\psi_m$ without ambiguity. The previous result implies in particular that ψ is of class $\mathcal{C}^1$ on each $\mathring{\mathcal{K}}_m$, and thus everywhere on $\mathcal{U}$. We define $g:\mathcal{U}\longrightarrow\mathbb{R}^d$ similarly, with the same property

$g|_{\mathcal{K}_m} = h_m$. With this construction, on each $\mathring{\mathcal{K}}_m$, one has $g = g_m = \nabla\psi_m = \nabla\psi$. As a result, we have $g = \nabla\psi$ on all of $\mathcal{U}$. Since each g_m is L-Lipschitz, it follows that g is L-Lipschitz on $\mathcal{U}$.

To see that $g \in G$, it only remains to show that ψ is ℓ-strongly convex, which is a consequence of the fact that it is everywhere a point-wise limit of a ψ_m, which is itself ℓ-strongly convex. □

2.4 Function Class Example: Neural Networks

Another natural idea is to consider classes G of parametrised functions, in particular Neural Networks (NNs) with Lipschitz activation functions. We will consider a relatively general expression for NNs borrowed from [50]. We consider a class G_{NN} of functions $g_\theta = h_N(\theta, \cdot) : \mathbb{R}^k \longrightarrow \mathbb{R}^d$ for a parameter vector $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^p$ is a compact set, and where h_N is the N-th layer of a recursive NN structure defined by

$$
\begin{aligned}
&h_0(\theta, x) = x, \quad \forall n \in [\![1, N]\!],\ h_n = \begin{cases} \mathbb{R}^p \times \mathbb{R}^k & \longrightarrow & \mathbb{R}^{d_n} \\ (\theta, x) & \longmapsto & a_n\left(\sum_{i=0}^{n-1} A_{n,i}(\theta) h_i(\theta, x) + b_n\theta\right) \end{cases}, \\
&N \in \mathbb{N},\ d_0 = k,\ d_N = d,\ \forall n \in [\![1, N]\!],\ d_n \in \mathbb{N}^*, \\
&a_n : \mathbb{R}^{d_n} \longrightarrow \mathbb{R}^{d_n} \text{ Lipschitz},\ b_n \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^{d_n}),\ \forall i \in [\![0, n-1]\!],\ A_{n,i} \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^{d_n \times d_i}),
\end{aligned}
\tag{8}
$$

where $\mathcal{L}(A, B)$ is the space of linear maps from A to B. The terms $A_{n,i}$ and b_n correspond to the weights matrices and biases respectively, and we allow the use of the entire parameter vector $\theta \in \Theta \subset \mathbb{R}^p$ at each layer for generality. The summation over the previous layers allows the inclusion of "skip-connections" in the architecture. Thanks to the assumption that the parameters lie in a compact set, we will show that the class G_{NN} verifies the conditions of our existence theorem Theorem 2.3.

Proposition 2.10. *Let $\Theta \subset \mathbb{R}^p$ be a compact set and G_{NN} the class of functions $\mathbb{R}^p \longrightarrow \mathbb{R}^d$ of the form $g_\theta = h_N(\theta, \cdot)$, with $\theta \in \Theta$ and h_N as in Eq. (8). Then G_{NN} verifies Assumption ii) of Theorem 2.3.*

Proof. An immediate induction over the layers shows that for $g_\theta \in G_{\mathrm{NN}}$, there exists a constant $L > 0$ independent of θ such that g_θ is L-Lipschitz on $\mathbb{R}^k$.
Concerning closedness for the compact-open topology (Definition 2.2), we will show the following stronger property: if $(g_m) \in (G_{\mathrm{NN}})^{\mathbb{N}}$ converges pointwise towards a function $f : \mathbb{R}^k \longrightarrow \mathbb{R}^d$, then there exists $\theta \in \Theta$ such that $f = g_\theta$. For $m \in \mathbb{N}$, we can write $g_m = g_{u_m}$ for $u_m \in \Theta$. Since the sequence (u_m) lies in the compact set Θ, there exists a converging subsequence $(u_{\alpha(m)})$ which converges towards $\theta \in \Theta$. Let $x \in \mathbb{R}^k$, we have the convergence $g_{u_m}(x) \longrightarrow f(x)$. By induction over the layers, the function $v \longmapsto g_v(x)$ is continuous, thus $g_{u_{\alpha(m)}}(x) \longrightarrow g_\theta(x)$. By uniqueness of the limit, $f(x) = g_\theta(x)$, and since $x \in \mathbb{R}^k$ was chosen arbitrarily, we conclude $f \in G$. □

Remark 2.11. *For simplicity, we presented NNs taking $x \in \mathbb{R}^k$ as input, yet the theory holds if $\mathcal{X}$ is a locally compact Polish space, just as in Theorem 2.3. For instance, one could take a Riemannian manifold.*

2.5 On the Necessity of the Lipschitz Constraint for Existence

Beyond the theoretical usefulness of the constraint that g be L-Lipschitz, this constraint may add substantial difficulty to the numerical implementation (see Section 4). As a result, one could consider the map problem Eq. (3) without the Lipschitz assumption on G. Unfortunately, this variant has no solution in general. We illustrate this in the light of the class of functions $\mathcal{F}_{\mathcal{E},L,\ell}$ introduced in Proposition 2.9, in the 1D case and consider G the cone of continuous non-decreasing functions, yielding the problem:

$$
\operatorname*{argmin}_{g \in \mathcal{C}^0(\mathbb{R}),\ \text{non-decreasing}} \mathrm{W}_2^2(g\#\mu, \nu), \tag{9}
$$

where we choose the specific measures $\mu := \mathcal{U}([-1,1])$ and $\nu := \frac{1}{2}\mathcal{U}([-2,-1]) + \frac{1}{2}\mathcal{U}([1,2])$. In this setting, no continuous function $g : \mathbb{R} \longrightarrow \mathbb{R}$ can satisfy $g\#\mu = \nu$. Indeed, suppose that such a continuous function g were to exist. On the one hand, since g is continuous, $\text{supp}(g\#\mu) = g(\text{supp}(\mu)) = g([-1,1])$. On the other hand, by assumption $\text{supp}(g\#\mu) = \text{supp}(\nu) = [-2,-1] \cup [1,2]$. However, since g is continuous and $[-1,1]$ is connected, $g([-1,1])$ is also connected, thus $[-2,-1] \cup [1,2]$ is connected, which is a contradiction.

We now consider a specific function g which satisfies $g\#\mu = \nu$:

$$g := \begin{cases} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \begin{cases} x-1 \text{ if } x < 0 \\ 0 \text{ if } x = 0 \\ x+1 \text{ if } x > 0 \end{cases} \end{cases}, \tag{10}$$

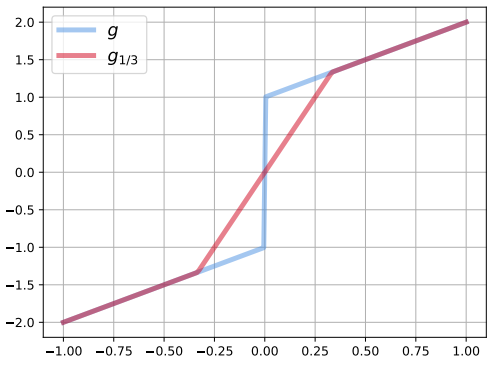
note that the value at 0 can be chosen arbitrarily. This function is not continuous, so we approach it by functions g_ε, with $\varepsilon \in (0,1)$, which are continuous and non-decreasing:

$$g_\varepsilon := \begin{cases} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \begin{cases} x-1 \text{ if } x \leq -\varepsilon \\ \frac{1+\varepsilon}{\varepsilon} x \text{ if } x \in [-\varepsilon, \varepsilon] \\ x+1 \text{ if } x \geq \varepsilon \end{cases} \end{cases}. \tag{11}$$

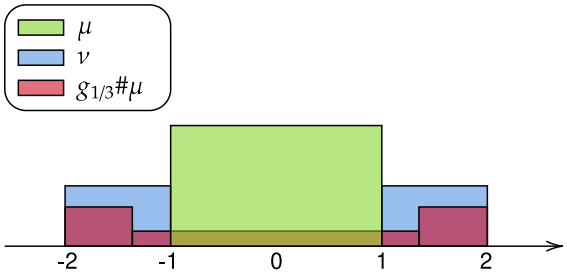
A straightforward computation yields:

$$g_\varepsilon\#\mu = \frac{1-\varepsilon}{2}\mathcal{U}([-2,-1-\varepsilon]) + \varepsilon\mathcal{U}([-1-\varepsilon, 1+\varepsilon]) + \frac{1-\varepsilon}{2}\mathcal{U}([1+\varepsilon,2]), \tag{12}$$

which we illustrate in Fig. 2. It follows that $g_\varepsilon\#\mu$ converges weakly towards ν as $\varepsilon \longrightarrow 0$. As a



(a) Illustration of the maps g from Eq. (10) and g_ε from Eq. (11) with $\varepsilon = 1/3$.



(b) Illustration of the image measure $g_{1/3}\#\mu$ with $\mu = \mathcal{U}([-1,1])$ and g_ε from Eq. (11).

Figure 2: Illustration of the counter-example to existence.

result, since the measures are compactly supported, $\text{W}_2^2(g_\varepsilon\#\mu, \nu) \xrightarrow[\varepsilon \longrightarrow 0]{} 0$, thus the value of Problem Eq. (9) is 0.

To conclude, if Problem Eq. (9) had a solution g, then it would be continuous and verify $\text{W}_2^2(g\#\mu, \nu) = 0$ (since the problem value is 0), thus $g\#\mu = \nu$, which is impossible by the connectivity argument. Therefore, the problem defined in Eq. (9) does not have a solution.

2.6 Discussion on Uniqueness

A natural question is the uniqueness of a solution of the problem

$$\underset{g \in G}{\text{argmin}}\ \mathcal{T}_c(g\#\mu, \nu),$$

in the case where the measures, the cost and the class G satisfy the conditions of Theorem 2.3, guaranteeing existence. A first negative answer concerns the simple case where μ, ν are discrete and at least two-dimensional. For instance, consider

$$\mu := \frac{1}{2}(\delta_{(-1,0)} + \delta_{(1,0)}), \quad \nu := \frac{1}{2}(\delta_{(0,-1)} + \delta_{(0,1)}).$$

Then there are two distinct maps g_1, g_2 both verifying $g_i \# \mu = \nu$, which are characterised in $L^2(\mu)$ by their values on the two points $(\pm 1, 0)$.

$$g_1((-1,0)) = (0,-1),\ g_1((1,0)) = (0,1), \quad g_2((-1,0)) = (0,1),\ g_2((1,0)) = (0,-1),$$

as we illustrate in Fig. 3. The previous example illustrates a potential issue for uniqueness, which is

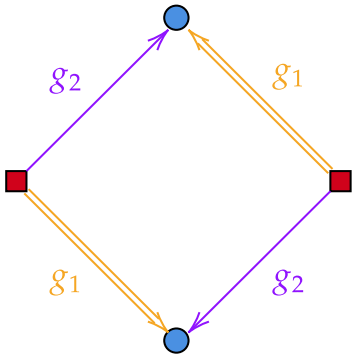


Figure 3: A simple case with two transport maps between 2-point discrete measures in $\mathbb{R}^2$.

the multiplicity of the set $\{g \in G \mid g\#\mu = \nu\}$. Another simple counter-example to uniqueness which stems from this property is for $\mu = \nu = \mathcal{N}(0, I)$ the standard d-variate Gaussian distribution. In this case, any rotation R verifies $R\#\mathcal{N}(0, I) = \mathcal{N}(0, I)$.

More generally, Brenier's polar factorisation theorem [11] sheds a light on our invariance issue. We present the theorem below for completeness, see also [47] Section 1.7.2.

Theorem 2.12 (Brenier's Polar Factorisation [11])**.** *Let $\mathcal{K} \subset \mathbb{R}^d$ be a compact set, and $g : \mathcal{K} \longrightarrow \mathbb{R}^d$. Consider $\mathcal{U}_\mathcal{K}$ the probability measure that is the uniform distribution on $\mathcal{K}$, suppose that $g\#\mathcal{U}_\mathcal{K} \ll \mathscr{L}$, then there exists a unique ($\mathscr{L}$-almost-everywhere) decomposition $g = (\nabla\varphi) \circ s$ such that:*

- *$\varphi : \mathcal{K} \longrightarrow \mathbb{R}^d$ is convex;*
- *$s : \mathcal{K} \longrightarrow \mathcal{K}$ is measure-preserving, which is to say that $s\#\mathcal{U}_\mathcal{K} = U_\mathcal{K}$.*

To fix the ideas, if we consider $\mu = \mathcal{U}_{[0,1]^d}$, we can fix $g \in G$ and assume $g\#\mathcal{U}_\mathcal{K} \ll \mathscr{L}$ (see sufficient conditions for this in Lemma A.1 in the Appendix), then decompose $g = \nabla\varphi \circ s$. Then any map h of the form $\nabla\varphi \circ r$ with r a measure-preserving map will verify $h\#\mathcal{U}_{[0,1]^d} = g\#\mathcal{U}_{[0,1]^d}$. To avoid such potential counter-examples, we will focus on the case where G is a subset of gradients of convex functions.

We provide a uniqueness result for the W_2 case, under the simplifying assumption that $\nu = \mu$. Note that if $L < 1$, the identity map does not belong to G, and there does not exist a $g \in G$ such that $g\#\mu = \mu$.

Proposition 2.13. *Suppose that*

$$G = \left\{ g : \mathbb{R}^d \longrightarrow \mathbb{R}^d : g = \nabla\varphi \ \mathscr{L}-a.e.,\ \varphi \text{ convex},\ g\ L-Lipschitz \right\},$$

and that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mu \ll \mathscr{L}$. Then if g_0 and g_1 are solutions of the problem

$$\operatorname*{argmin}_{g \in G} \mathrm{W}_2^2(g\#\mu, \mu),$$

then $g_0 = g_1$ everywhere on $\mathrm{supp}(\mu)$.

Proof. We will show that if g_0 and g_1 are solutions, then $g_0\#\mu = g_1\#\mu$. First, one may write $g_i = \nabla\varphi_i$ with φ_i convex (for $i = 0, 1$). By [47] Theorem 1.48, since φ_i is convex, g_i is the optimal transport map between μ and $\nabla\varphi_i\#\mu$. Consider for $t \in [0, 1]$ the interpolation $g_t := (1-t)g_0 + tg_1$. Then by definition (see [4], Section 9.2), the curve $(g_t\#\mu)_{t\in[0,1]}$ is a[4] generalised geodesic between $g_0\#\mu$ and $g_1\#\mu$ with respect to the base measure μ. This allows us to apply [4] Lemma 9.2.1, specifically Equation 9.2.7c, which yields

$$\forall t \in [0, 1],\ \mathrm{W}_2^2(g_t\#\mu, \mu) \leq (1-t)\mathrm{W}_2^2(g_0\#\mu, \mu) + t\mathrm{W}_2^2(g_1\#\mu, \mu) - t(1-t)\mathrm{W}_2^2(g_0\#\mu, g_1\#\mu).$$

The curvature of this generalised geodesic will allow us to build a better solution if $g_0\#\mu \neq g_1\#\mu$, as we illustrate in Fig. 4.

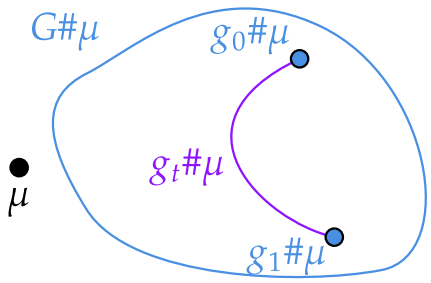


Figure 4: The generalised geodesic based on μ between $g_0\#\mu$ and $g_1\#\mu$.

Taking $t = 1/2$ yields, using the optimality of g_0 and g_1 and writing v for the problem value:

$$\mathrm{W}_2^2(g_{\frac{1}{2}}\#\mu, \mu) \leq v - \frac{1}{4}\mathrm{W}_2^2(g_0\#\mu, g_1\#\mu).$$

Since G is convex, we have $g_{\frac{1}{2}} \in G$, which imposes $\mathrm{W}_2^2(g_0\#\mu, g_1\#\mu) = 0$, since v is the optimal problem value. We conclude $g_0\#\mu = g_1\#\mu$. However, as stated earlier, by [47] Theorem 1.48, g_i is the optimal transport map between μ and $g_i\#\mu$ for $i = 0, 1$. By uniqueness of the optimal transport map in this setting, we conclude $g_0 = g_1$. (The equality holds μ-a.e., then since g_0 and g_1 are assumed Lipschitz, this shows equality everywhere on $\mathrm{supp}(\mu)$.) □

Remark 2.14. *One could replace the set G in Proposition 2.13 by a convex subset of G, the proof of the result would follow verbatim.*

Remark 2.15. *The problem in Proposition 2.13 is related to the problem of the Wasserstein metric projection, which was studied in [17] (see Section 5), from which the curvature argument in our proof was closely inspired. This Wasserstein projection problem was also studied for W_p^p in [1].*

Remark 2.16. *Under some assumptions, it may be possible to find subclasses of gradients of convex functions G such that the set $G\#\mu \subset \mathcal{P}_2(\mathbb{R}^d)$ is geodesically convex (with respect to W_2 geodesics): take $g_0, g_1 \in G$, assume that $g_0\#\mu \ll \mathscr{L}$ (Lemma A.1 provides a sufficient condition on g_0 and μ for this to be the case). Then the W_2 geodesic from $g_0\#\mu$ to $g_1\#\mu$ is*

$$\nu_t := ((1-t)I + tT)\#g\#\mu_0,$$

where T is the optimal transport map from $g_0\#\mu$ to $g_1\#\mu$, which is uniquely defined thanks to Brenier's Theorem (see [47], Theorem 1.22 for a possible reference without compactness assumptions). Since $(T \circ g_0)\#\mu = g_1\#\mu$, under some regularity assumptions, it may be possible to show that $T \circ g_0 = g_1$ using the Monge-Ampère equation, then $((1-t)I + tT) \circ g_0 = (1-t)g_0 + tg_1 \in G$. In this case, the generalised geodesic based on μ coincides with the W_2 geodesic between $g_0\#\mu$ and $g_1\#\mu$.

[4] In this case, since $\mu \ll \mathscr{L}$, there is even uniqueness of *the* generalised geodesic between $g_0\#\mu$ and $g_1\#\mu$, but we do not use that fact.

Unfortunately, $\rho \longmapsto \mathrm{W}_2^2(\rho, \nu)$ is not convex along W_2 geodesics, since it satisfies the opposite inequality ([4], Theorem 7.3.2). As a result, even if we found a convex class G of gradients of convex functions such that $G\#\mu$ were geodesically convex, curvature arguments would not yield uniqueness immediately. Intuition suggests that in some sense, the problem minimises a concave function over a convex set, which bodes poorly with uniqueness.

In Section 3.2, we shall study the case $d = 1$ and show uniqueness and an explicit expression for the minimiser of the map problem for non-decreasing functions g and the squared Euclidean cost. To conclude this discussion, even for the favourable case where $\mu \ll \mathscr{L}^d$, G is a subset of gradients of convex functions and $c(x, y) = \|x - y\|_2^2$, we conjecture that uniqueness is not guaranteed in general for $d \geq 2$.

2.7 The Plan Approximation Problem

In some cases, one may have access to a transport plan between two measures μ, ν, which poses the natural question of finding a map that approximates this transport plan. For instance, one may compute the optimal entropic plan [13], a Gaussian-Mixture-Model optimal plan [18], or an optimal transport plan for a cost that does not verify the twist condition (see [47] Definition 1.16), or more generally an optimal plan when the Monge problem is not equivalent to the Kantorovich problem (see [11, 27, 42] for some known equivalence cases).

Given a cost $C : (\mathbb{R}^k \times \mathbb{R}^d) \times (\mathbb{R}^k \times \mathbb{R}^d) \longrightarrow \mathbb{R}_+$, and measures $\mu \in \mathcal{P}(\mathbb{R}^k), \nu \in \mathcal{P}(\mathbb{R}^d)$, we will want to approximate a plan $\gamma \in \Pi(\mu, \nu)$ by the image measure $(I, g)\#\mu$, where I denotes the identity map of $\mathbb{R}^k$. We define the Constrained Approximate Transport Plan problem as:

$$\operatorname*{argmin}_{g \in G} \mathcal{T}_C((I, g)\#\mu, \gamma). \tag{13}$$

Similarly to Eq. (3), the transport cost in Eq. (13) can be re-written using the change-of-variables formula (Lemma 2.1):

$$\mathcal{T}_C((I, g)\#\mu, \gamma) = \min_{\rho \in \Pi(\mu, \gamma)} \int_{\mathbb{R}^k \times (\mathbb{R}^k \times \mathbb{R}^d)} C\left((x, g(x)), (y_1, y_2)\right) \mathrm{d}\rho(x, y_1, y_2). \tag{14}$$

To begin with, one may cast Eq. (13) as a map problem (Eq. (3)), providing existence automatically under adequate conditions.

Corollary 2.17. *Consider the class of functions*

$$\widetilde{G} := \{\tilde{g} : \mathbb{R}^k \longrightarrow \mathbb{R}^k \times \mathbb{R}^d \ : \ \tilde{g} = (x, y) \longmapsto (x, g(y)),\ g \in G\},$$

the map problem (Eq. (3)) is a particular map problem (Eq. (13)):

$$\min_{g \in G} \mathcal{T}_C((I, g)\#\mu, \gamma) = \min_{\tilde{g} \in \widetilde{G}} \mathcal{T}_C(\tilde{g}\#\mu, \gamma),$$

hence existence holds by Theorem 2.3 if the conditions of the theorem are verified by $C, \widetilde{G}$ and the measures μ, γ.

Remark 2.18. *In the light of Remark 2.6, one could replace the input space $\mathbb{R}^k$ and the target space $\mathbb{R}^d$ by Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ verifying the Heine-Borel property, in which case condition 1) would ask for $(x_1, x_2) \longmapsto C((x_1, x_2), (y_1, y_2))$ to be proper.*

We shall see that in certain cases, the two problems Eq. (13) and Eq. (3) are in fact equivalent.

Proposition 2.19. *Consider a cost C of the separable form $C((x_1, x_2), (y_1, y_2)) = h(c_1(x_1, y_1), c_2(x_2, y_2))$, where $h : \mathbb{R}_+ \times \mathbb{R}_+ \longrightarrow \mathbb{R}_+$, $c_1 : \mathbb{R}^k \times \mathbb{R}^k \longrightarrow \mathbb{R}_+$ and $c_2 : \mathbb{R}^d \times \mathbb{R}^d$ are lower semi-continuous, with $\forall x \in \mathbb{R}^k$, $c_1(x, x) = 0$, and $\forall u, v \in \mathbb{R}_+$, $h(u, v) \geq v$ and $h(0, v) = v$. Let $g : \mathbb{R}^k \longrightarrow \mathbb{R}^d$ be a measurable function, $\nu \in \mathcal{P}(\mathbb{R}^d)$ and $\mu \in \mathcal{P}(\mathbb{R}^k)$. Let $\gamma \in \Pi(\mu, \nu)$ be a plan between μ and ν.*

We assume that the value $\mathcal{T}_C((I,g)\#\mu,\gamma)$ is finite. We have the equality

$$\mathcal{T}_{c_2}(g\#\mu,\nu) = \mathcal{T}_C((I,g)\#\mu,\gamma).$$

Proof. For $\rho \in \Pi(\mu,\gamma)$, let $A(\rho) := \int_{\mathbb{R}^k\times(\mathbb{R}^k\times\mathbb{R}^d)} C((x,g(x)),(y_1,y_2))\mathrm{d}\rho(x,y_1,y_2) < +\infty$, and denote $A^* := \mathcal{T}_C((I,g)\#\mu,\gamma)$. Likewise, for $\pi \in \Pi(\mu,\nu)$, let $B(\pi) := \displaystyle\int_{\mathbb{R}^k\times\mathbb{R}^d} c_2(g(x),y)\mathrm{d}\pi(x,y)$, and $B^* := \mathcal{T}_{c_2}(g\#\mu,\nu)$.

First, we prove $A^* \leq B^*$. By [47], Theorem 1.7, there exists $\pi^* \in \Pi(\mu,\nu)$ such that $B^* = B(\pi^*)$. Define $\rho \in \Pi(\mu,\gamma)$ a measure such that for each test function f,

$$\int_{\mathbb{R}^k\times(\mathbb{R}^k\times\mathbb{R}^d)} f(x,y_1,y_2)\mathrm{d}\rho(x,y_1,y_2) = \int_{\mathbb{R}^k\times\mathbb{R}^d} f(y_1,y_1,y_2)\mathrm{d}\pi^*(y_1,y_2),$$

or symbolically "$\rho(\mathrm{d}x\mathrm{d}y_1\mathrm{d}y_2) = \delta_{y_1}(\mathrm{d}x)\pi^*(\mathrm{d}y_1\mathrm{d}y_2)$". Then, since $h(c_1(y_1,y_1),c_2(g(y_1),y_2)) = c_2(g(y_1),y_2)$, we have

$$A^* \leq A(\rho) = \int_{\mathbb{R}^k\times(\mathbb{R}^k\times\mathbb{R}^d)} h(c_1(y_1,y_1),c_2(g(y_1),y_2))\mathrm{d}\pi^*(y_1,y_2) = \int_{\mathbb{R}^k\times\mathbb{R}^d} c_2(g(y_1),y_2)\mathrm{d}\pi^*(y_1,y_2) = B^*.$$

Now for $A^* \geq B^*$, we let $\rho \in \Pi(\mu,\gamma)$. Using $h(u,v) \geq v$, we have

$$A(\rho) = \int_{\mathbb{R}^k\times(\mathbb{R}^k\times\mathbb{R}^d)} h(c_1(x,y_1),c_2(g(x),y_2))\mathrm{d}\rho(x,y_1,y_2) \geq \int_{\mathbb{R}^k\times(\mathbb{R}^k\times\mathbb{R}^d)} c_2(g(x),y_2)\mathrm{d}\rho(x,y_1,y_2).$$

Again, we can define $\pi \in \Pi(\mu,\nu)$ such that for any test function f,

$$\int_{\mathbb{R}^k\times\mathbb{R}^d} f(x,y_2)\mathrm{d}\pi(x,y_2) = \int_{\mathbb{R}^k\times(\mathbb{R}^k\times\mathbb{R}^d)} f(x,y_2)\mathrm{d}\rho(x,y_1,y_2),$$

and notice

$$\int_{\mathbb{R}^k\times(\mathbb{R}^k\times\mathbb{R}^d)} c_2(g(x),y_2)\mathrm{d}\rho(x,y_1,y_2) = B(\pi) \geq B^*,$$

which yields $A^* \geq B^*$. □

For example, the cost $C(x,y) = \|x-y\|_2^2$ satisfies these conditions (with $h(u,v) = u+v$), and thus the problems Eq. (3) and Eq. (13) are equivalent. This is still the case for costs of the form $C = \|\cdot - \cdot\|_p^{qp}$ for $p \geq 1$ and $q > 0$, in which case one takes $h(u,v) = (u^{1/q}+v^{1/q})^q$. For $C((x_1,x_2),(y_1,y_2)) = \|(x_1,x_2)-(y_1,y_2)\|_\infty^p$, this is also the case with $c_1(x,y) = c_2(x,y) = \|x-y\|_\infty^p$ and $h(u,v) = \max(u,v)$.

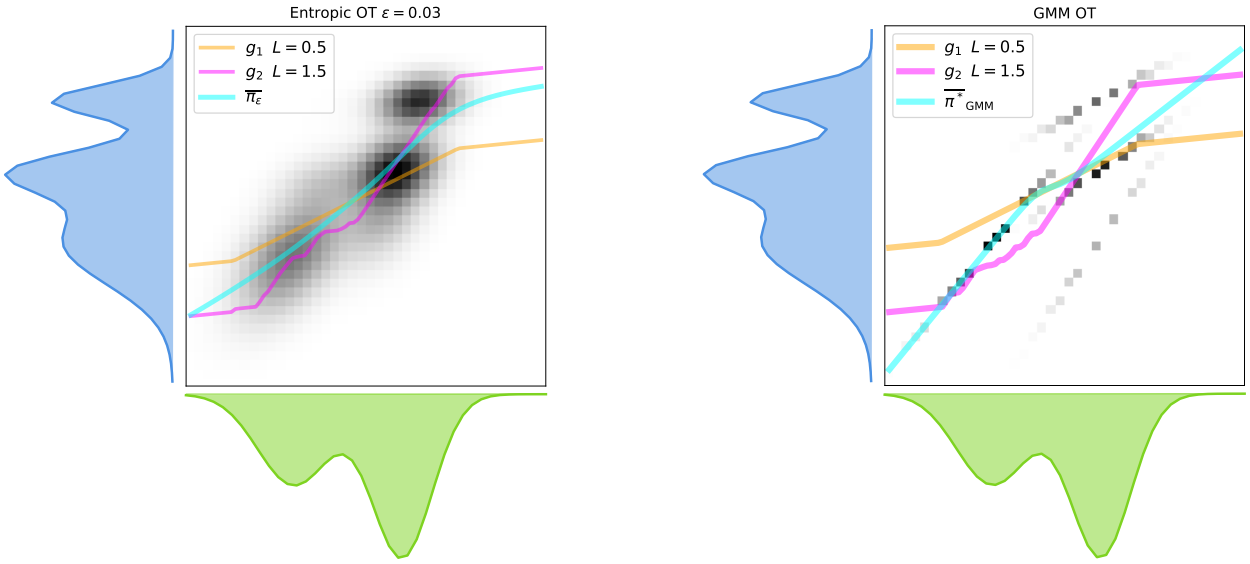
In contrast, a possible choice of norm on the product space is $\|x\|_\Sigma = (x^\top\Sigma^{-1}x)^{1/2}$ for Σ symmetric positive-definite. This choice is of interest since the cost $C((x_1,x_2),(y_1,y_2)) = \|(x_1,x_2)-(y_1,y_2)\|_\Sigma^2$ is quadratic (which is desirable for numerics), but does not satisfy the equivalence condition from Proposition 2.19 as soon as Σ is not block-diagonal.

In Fig. 5, we illustrate the plan approximation problem for the quadratic cost for two different plans: the Entropic Optimal Transport plan [40] and the Gaussian Mixture Model OT plan [18]. The numerics where done using the tools presented in Section 4.2. Note that the plan approximation is equivalent to a map problem in this case, and has a particular structure due to the one-dimensional setting, hence we emphasise that Fig. 5 is merely an illustration of the problem at hand.

3 Alternate Minimisation in the Squared Euclidean Case

The map problem Eq. (3) is a minimisation problem over $\pi \in \Pi(\mu,\nu)$ and $g \in G$:

$$\min_{g\in G}\min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathbb{R}^d} c(g(x),y)\mathrm{d}\pi(x,y).$$



(a) plan approximation solutions for the Entropic-OT plan [40].

(b) Illustration of plan approximation solutions for the GMM-OT plan [18].

Figure 5: Illustration of solutions of plan approximation problems (Eq. (13)), for two different plans between Gaussian Mixtures. We compare the plans with $L = 1/2$ and $L = 3/2$-Lipschitz solutions, as well as to the barycentric projection of the given plans (see Section 3.1).

In this section, we study this alternate minimisation problem in the case where $c(x, y) = \|x - y\|_2^2$ and $\mathcal{X} = \mathbb{R}^d$, thus with maps $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$.

When $\pi \in \Pi(\mu, \nu)$ is fixed, the sub-problem has the particular structure

$$\min_{g \in G} \int_{\mathcal{X} \times \mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y). \tag{15}$$

To ensure the finiteness of the cost, we assume that $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. We shall see in Section 3.1 that the problem in Eq. (15) is equivalent to the L^2 projection of the barycentric map $\overline{\pi}$ onto the set G, provided that G is a convex and closed subset of $L^2(\mu)$.

When $g \in G$ is fixed, the problem reads

$$\min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y), \tag{16}$$

and can be seen from two different viewpoints: either as the squared Euclidean optimal transport problem between $g\#\mu$ and ν (i.e. $\mathrm{W}_2^2(g\#\mu, \nu)$), or as the optimal transport problem with cost $c(x, y) := \|g(x) - y\|_2^2$ between μ and ν. If $g\#\mu$ is absolutely continuous and ν is discrete, then Eq. (16) is a semi-discrete OT problem. We provide sufficient conditions on g for this to be the case in Appendix A.1.

This alternate minimisation viewpoint poses a natural question: if $\pi := \pi^* \in \Pi^*(\mu, \nu)$ is an optimal plan between μ and ν for the quadratic cost $c(y, y') := \|y - y'\|_2^2$, does the following equality holds?

$$\operatorname*{argmin}_{g \in G} \min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y) \overset{?}{=} \operatorname*{argmin}_{g \in G} \int_{\mathcal{X} \times \mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi^*(x, y). \tag{17}$$

In Section 3.2, we prove that this equality holds in the one-dimensional case $\mathcal{X} = \mathbb{R}^d = \mathbb{R}$ and if G is a subclass of non-decreasing functions, thus generalizing a result of [37]. We also provide a counter-example of this property when $d \geq 2$ in Section 3.3.

3.1 Projection of the Barycentric Map

In this section, we will show that the sub-problem with $\pi \in \Pi(\mu, \nu)$ fixed can be written as the following L^2 projection problem:

$$\operatorname*{argmin}_{g \in G} \int_{\mathcal{X} \times \mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y) = \operatorname*{argmin}_{g \in G} \|g - \overline{\pi}\|_{L^2(\mu)}^2,$$

where $\overline{\pi}$ is the barycentric projection of π (defined below), and $L^2(\mu)$ is a shorthand for $L^2(\mu; \mathbb{R}^d)$, the space of measurable functions $T : \mathcal{X} \longrightarrow \mathbb{R}^d$ such that $\int_{\mathcal{X}} \|T(x)\|_2^2 \mathrm{d}\mu(x) < +\infty$. We begin by briefly introducing the notion of barycentric projection.

The *barycentric projection* of π is the map $\overline{\pi} : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ defined for μ-almost all $x \in \mathbb{R}^d$ by

$$\overline{\pi}(x) = \mathbb{E}_{(X,Y) \sim \pi}[Y | X = x]. \tag{18}$$

As illustrated in Fig. 6, if π admits a disintegration with respect to its first marginal μ of the form $\pi(\mathrm{d}x\mathrm{d}y) = \pi_x(\mathrm{d}y)\mu(\mathrm{d}x)$, then

$$\overline{\pi}(x) = \int_{\mathbb{R}^d} y \, \mathrm{d}\pi_x(y).$$

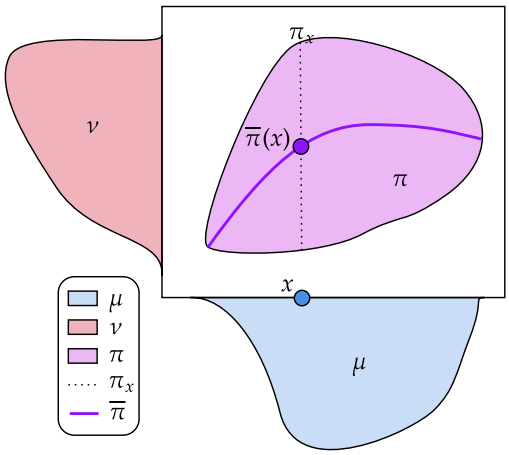


Figure 6: Illustration of a barycentric projection. The disintegration of the coupling π with respect to its first marginal μ at x is the measure π_x concentrated on the dotted line. The barycentric projection of π evaluated at x is the mean of the measure π_x.

Since the conditional expectation minimises the L^2 distance, we also have

$$\overline{\pi} = \operatorname*{argmin}_{T \in L^2(\mu)} \int_{\mathbb{R}^{2d}} \|y - T(x)\|_2^2 \mathrm{d}\pi(x, y), \tag{19}$$

where the equality is to be understood in $L^2(\mu)$. Another interesting property is that if $\pi = \pi^* \in \Pi^*(\mu, \nu)$ is an optimal transport plan between μ and ν with respect to the squared Euclidean distance cost, then by [4], Section 6.2.3, there exists $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$ convex such that for π^*-almost-every $(x, y) \in \mathbb{R}^{2d}$, we have $y \in \partial\varphi(x)$, where $\partial\varphi(x)$ denotes the *Fréchet sub-differential* of φ:

$$y \in \partial\varphi(x) \Longleftrightarrow \liminf_{z \longrightarrow x} \frac{\varphi(z) - \varphi(x) - \langle y, z - x \rangle}{\|x - z\|_2} \geq 0.$$

Since the Fréchet sub-differential of a convex function is convex, it follows that for μ-almost every $x \in \mathbb{R}^d$, $\overline{\pi^*}(x) \in \partial\varphi(x)$.

If we require constraints on T in Eq. (19), we obtain exactly the sub-problem of the map problem with a fixed plan π (Eq. (15)), which we reproduce below:

$$\operatorname*{argmin}_{g \in G} \int_{\mathcal{X} \times \mathbb{R}^d} \|g(x) - y\|_2^2 \mathrm{d}\pi(x, y).$$

For this reason, we call this problem the Constrained Barycentric Map problem. A consequence of the proof of Theorem 2.3 is that this problem has a solution. If G is a convex set and closed in $L^2(\mu)$, then existence and uniqueness are guaranteed by the Hilbert projection Theorem. Since $\overline{\pi}$ minimises the L^2 distance, it is a solution of Eq. (15) if it is in G.

Using the fact that the barycentric projection is an L^2 projection (Eq. (19)), one may re-write the Projected Barycentric Map Problem Eq. (15) as an L^2 minimisation with respect to the barycentric projection. In Proposition 3.1, we need not assume that $\mathcal{X} = \mathbb{R}^d$, but we shall apply it later in Section 3.2 to the case $\mathcal{X} = \mathbb{R}$.

Proposition 3.1. *Let $\pi \in \Pi(\mu, \nu)$ and $f : \mathcal{X} \longrightarrow \mathbb{R}^d$ be a measurable function. Then one has*

$$\int_{\mathcal{X}\times\mathbb{R}^d} \|f(x) - y\|_2^2 \mathrm{d}\pi(x,y) = \int_{\mathcal{X}} \|f(x) - \overline{\pi}(x)\|_2^2 \mathrm{d}\mu(x) + \int_{\mathcal{X}\times\mathbb{R}^d} \|y - \overline{\pi}(x)\|_2^2 \mathrm{d}\pi(x,y), \tag{20}$$

and as a result, the Projected Barycentric Map problem Eq. (15) is equivalent to the problem

$$\operatorname*{argmin}_{g\in G} \int_{\mathcal{X}} \|g(x) - \overline{\pi}(x)\|_2^2 \mathrm{d}\mu(x). \tag{21}$$

Moreover, the second term on the right-hand side of Eq. (20) only depends on $\overline{\pi}$ and the measures μ, ν (it doesn't depend on π). More precisely, we have

$$\int_{\mathcal{X}\times\mathbb{R}^d} \|y - \overline{\pi}(x)\|_2^2 \mathrm{d}\pi(x,y) = m_2(\nu) - m_2(\overline{\pi}\#\mu),$$

where $m_2(\rho) := \int \|x\|_2^2 \mathrm{d}\rho(x)$ for a positive measure ρ.

Proof. Denote J the left-hand-side of Eq. (20), and compute (taking the expectation under $(X, Y) \sim \pi$)

$$\begin{aligned} J &= \mathbb{E}\left[\|Y - f(X)\|_2^2\right] = \mathbb{E}\left[\|Y - \overline{\pi}(X) + \overline{\pi}(X) + f(X)\|_2^2\right] \\ &= \mathbb{E}\left[\|Y - \overline{\pi}(X)\|_2^2\right] + \mathbb{E}\left[\|\overline{\pi}(X) + f(X)\|_2^2\right] + 2\mathbb{E}\left[(Y - \overline{\pi}(X))^T(\overline{\pi}(X) + f(X))\right], \end{aligned}$$

then since $\overline{\pi}(X)$ is the orthogonal projection of Y onto the set of random variables that are functions of X, the inner product $\mathbb{E}\left[(Y - \overline{\pi}(X))^T(\overline{\pi}(X) + f(X))\right]$ is zero, yielding Eq. (20).

We can expand the norm in the second term of the right-hand side of Eq. (20) using $m_2(\nu)$ the second moment of ν and get

$$\int_{\mathcal{X}\times\mathbb{R}^d} \|y - \overline{\pi}(x)\|_2^2 \mathrm{d}\pi(x,y) = m_2(\nu) - 2\int_{\mathcal{X}\times\mathbb{R}^d} y \cdot \overline{\pi}(x) \mathrm{d}\pi(x,y) + \int_{\mathcal{X}} \|\overline{\pi}(x)\|_2^2 \mathrm{d}\mu(x).$$

Writing the disintegration of π w.r.t. μ as $\pi(\mathrm{d}x, \mathrm{d}y) = \pi_x(\mathrm{d}y)\mu(\mathrm{d}x)$, we re-write the second term as

$$\int_{\mathcal{X}\times\mathbb{R}^d} y \cdot \overline{\pi}(x) \mathrm{d}\pi(x,y) = \int_{\mathcal{X}} \overline{\pi}(x) \cdot \left(\int_{\mathbb{R}^d} y \mathrm{d}\pi_x(y)\right) \mathrm{d}\mu(x) = \int_{\mathcal{X}} \overline{\pi}(x) \cdot \overline{\pi}(x) \mathrm{d}\mu(x) = m_2(\overline{\pi}\#\mu).$$

Putting our computations together yields

$$\int_{\mathcal{X}\times\mathbb{R}^d} \|y - \overline{\pi}(x)\|_2^2 \mathrm{d}\pi(x,y) = m_2(\nu) - m_2(\overline{\pi}\#\mu).$$

□

Remark 3.2 (Ties to the Convex Least Squares Estimator [32].)**.** *In [32], Manole et al. study the statistical properties of various estimators of Optimal Transport maps, assuming some regularity on the input distributions. Specifically, they introduce the so-called Convex Least Squares Estimator: given* $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ *with the* (x_i) *being i.i.d. samples of* μ *and* $\hat{\nu}_m := \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$ *with the* (y_j) *i.i.d. samples of* ν*, the estimator is defined as*

$$\hat{T}_{n,m} = \nabla\hat{\varphi}, \quad \hat{\varphi} \in \operatorname*{argmin}_{\varphi\in\Phi_L} \sum_{i=1}^n\sum_{j=1}^m \hat{\pi}^*_{i,j}\|\nabla\varphi(x_i) - y_j\|_2^2, \tag{22}$$

where $\hat{\pi}^*$ *is an optimal transport plan between* $\hat{\mu}_n$ *and* $\hat{\nu}_m$*, and where* Φ_L *is the set of* $\mathcal{C}^1$ *convex functions from* $\Omega \subset \mathbb{R}^d$ *to* $\mathbb{R}$ *with a* L*-Lipschitz gradient. Notice that Eq. (22) is a Constrained Barycentric Projection problem Eq. (15) with a specific (discrete) transport plan* $\hat{\pi}^*$*, chosen to be the optimal transport plan between* $\hat{\mu}_n$ *and* $\hat{\nu}_m$*, and with the particular class* $G := \mathcal{F}_{\mathcal{E},L,\ell}$ *(introduced in Section 2.3).*

3.2 Equivalence to a Constrained Barycentric Projection in Dimension 1

In this section, we shall prove that the Constrained Approximate Transport Map problem (Eq. (3)) is equivalent to the Constrained Barycentric Projection Problem (Eq. (15)) for the quadratic cost in dimension 1. This provides a positive answer to the question raised in Eq. (17) in this particular case. The idea behind this equivalence stems from the fact that in dimension one, optimal transport maps are non-decreasing, and the composition of two optimal transport maps remains an optimal transport map.

Proposition 3.3. *For* $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$*, and* G *a subclass of the non-decreasing functions* $g : \mathbb{R} \longrightarrow \mathbb{R}$ *such that* $g\#\mu \in \mathcal{P}_2(\mathbb{R})$*, we have the equality*

$$\operatorname*{argmin}_{g\in G} \mathrm{W}_2^2(g\#\mu, \nu) = \operatorname*{argmin}_{g\in G} \|g - \overline{\pi^*}\|^2_{L^2(\mu)}, \tag{23}$$

where π^* *is an optimal transport plan between* μ *and* ν *for the squared Euclidean cost.*

Proposition 3.3 generalises [37] Proposition 1, which proves the same equivalence for a specific class of functions G, and assuming μ to be either discrete or absolutely continuous with respect to the Lebesgue measure.

The proof of Proposition 3.3 hinges on Lemma 3.4, which is intuitive in the absolutely continuous or discrete case, but a bit more technical in full generality. We write below the cumulative distribution function of a probability measure ρ as $F_\rho := x \longmapsto \rho((-\infty, x])$. Since it is non-decreasing, we can define its **right-inverse** as (using the notation $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$):

$$F_\rho^{\leftarrow} : \mathbb{R} \longrightarrow \overline{\mathbb{R}} : \quad \forall p \in \mathbb{R},\ F_\rho^{\leftarrow}(p) := \inf\left\{x \in \mathbb{R} \mid\ F_\rho(x) \geq p\right\}.$$

Lemma 3.4. *Let* $\mu \in \mathcal{P}(\mathbb{R})$ *and* $g : \mathbb{R} \longrightarrow \mathbb{R}$ *be a non-decreasing function, we have the following almost-everywhere change of variables formula for the quantile functions of* $g\#\mu$ *and* μ*:*

$$F^{\leftarrow}_{g\#\mu} = g \circ F^{\leftarrow}_\mu, \quad \mathscr{L}_{[0,1]}\text{-almost-everywhere.}$$

Proof. The proof is provided in Appendix A.2. □

Proof of Proposition 3.3 . Let $g \in G$. By [47] Proposition 2.17 and by Lemma 3.4 successively, we have

$$\mathrm{W}_2^2(g\#\mu, \nu) = \int_0^1 |F^{\leftarrow}_{g\#\mu}(p) - F^{\leftarrow}_\nu(p)|^2 \mathrm{d}p = \int_0^1 |g \circ F^{\leftarrow}_\mu(p) - F^{\leftarrow}_\nu(p)|^2 \mathrm{d}p = \int_{\mathbb{R}^2} |g(x) - y|^2 \mathrm{d}\pi(x, y),$$

where $\pi := (F_\mu^{\leftarrow}, F_\nu^{\leftarrow})\#\mathscr{L}_{[0,1]}$, which by [47] Theorem 2.9 is the unique optimal plan between μ and ν for the squared Euclidean cost. We apply Proposition 3.1, which yields

$$\mathrm{W}_2^2(g\#\mu,\nu) = \int_{\mathbb{R}^2} |g(x)-y|^2 \mathrm{d}\pi(x,y) = \int_{\mathbb{R}} |g(x)-\overline{\pi}(x)|^2 \mathrm{d}\mu(x) + m_2(\nu) - m_2(\overline{\pi}\#\mu).$$

Given the expression of the right-hand-side above, we conclude that

$$\operatorname*{argmin}_{g\in G} \mathrm{W}_2^2(g\#\mu,\nu) = \operatorname*{argmin}_{g\in G} \|g-\overline{\pi^*}\|_{L^2(\mu)}^2,$$

(for any optimal transport plan π^* between μ and ν for the squared Euclidean cost, and we have even remarked that such a plan is in fact unique) since the costs are equal up to a constant independent of g. □

3.3 Counter-Example to Equivalence to Constrained Barycentric Projection in Dimension 2

In this section, we provide a negative example to the question formulated in Eq. (17), namely that

$$\operatorname*{argmin}_{g\in G} \mathrm{W}_2^2(g\#\mu,\nu) \neq \operatorname*{argmin}_{g\in G} \|g-\overline{\pi^*}\|_{L^2(\mu)},$$

where π^* is an optimal transport plan (for the squared Euclidean cost) between μ and ν, in dimension $d \geq 2$. We take G to be the class of *monotone* continuous functions $g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$, which is to say that

$$\forall x, y \in \mathbb{R}^2,\ \langle g(x)-g(y), x-y\rangle \geq 0.$$

Note that gradients of convex functions are monotone, but the converse does not hold. For $(a,b,x) \in (0,+\infty)^3$, we consider the following measures:

$$\mu := \frac{2}{3}\delta_{(0,0)} + \frac{1}{3}\delta_{(x,0)} \text{ and } \nu := \frac{2}{3}\delta_{(0,0)} + \frac{1}{3}\delta_{(-a,b)}.$$

There is a unique optimal transport plan π^* between μ and ν, given by

$$\pi^* = \frac{1}{3}\delta_{(0,0)\otimes(0,0)} + \frac{1}{3}\delta_{(0,0)\otimes(-a,b)} + \frac{1}{3}\delta_{(x,0)\otimes(0,0)}.$$

Its barycentric projection is characterised by the following equation

$$\overline{\pi^*}(0,0) = (-a/2, b/2) \text{ and } \overline{\pi^*}(x,0) = (0,0).$$

We now consider the problem $\min_{g\in G} \|g-\overline{\pi^*}\|_{L^2(\mu)}$. A solution of this problem is characterised by its values on the support of μ, and one may reduce the problem to an optimisation over $g(0,0)$ and $g(x,0)$, with the monotonicity constraint $\langle g(0,0)-g(x,0), (0,0)-(x,0)\rangle \geq 0$. Since $\overline{\pi^*}$ itself verifies this condition, it is the only solution (in the sense of $L^2(\mu)$). We conclude

$$\operatorname*{argmin}_{g\in G} \|g-\overline{\pi^*}\|_{L^2(\mu)}^2 = \left\{\overline{\pi^*}\right\}.$$

We now show that the problem $\operatorname*{argmin}_{g\in G} \mathrm{W}_2^2(g\#\mu,\nu)$ has a different solution set. First, we compute

$$\mathrm{W}_2^2(\overline{\pi^*}\#\mu,\nu) = \frac{a^2+b^2}{6}.$$

However, if we introduce $g \in G$ such that $g(0,0) = (0,0)$ and $g(x,0) = (0,b)$, we have

$$\mathrm{W}_2^2(g\#\mu,\nu) = \frac{a^2}{3}.$$

For instance, $(a,b,x) := (1,10,1)$ yields

$$\mathrm{W}_2^2(\overline{\pi^*}\#\mu,\nu) = \frac{a^2+b^2}{6} = \frac{101}{6} > \mathrm{W}_2^2(g\#\mu,\nu) = \frac{a^2}{3} = \frac{1}{3}.$$

We illustrate the point configurations for $(a,b,x) := (1,3,1)$ in Fig. 7.

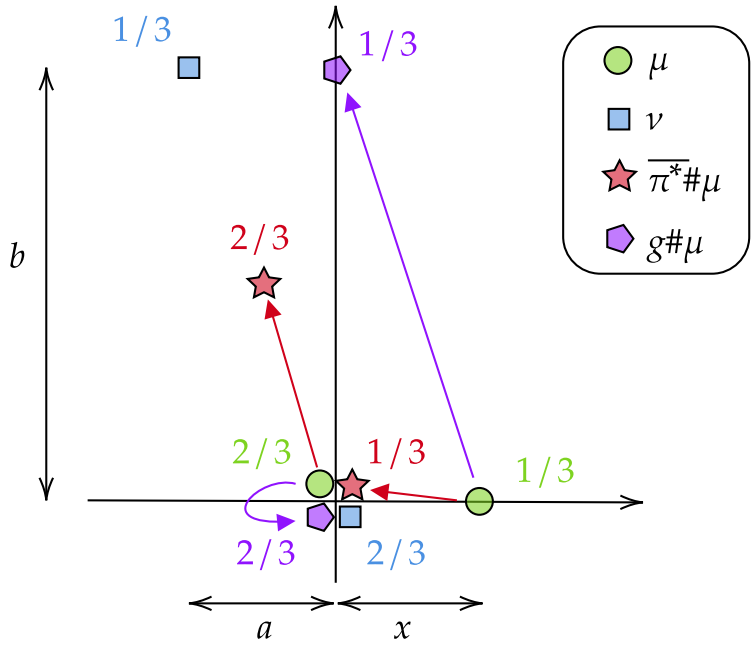


Figure 7: Illustration of the two-dimensional counter-example to the equivalence of the map problem to the L^2 projection of the barycentric projection. The four points close to $(0,0)$ are represented with an offset for legibility, and represent four points equal to $(0,0)$ exactly.

4 Discrete Measures and Numerical Methods

In this section, we consider some numerical methods to solve the approximate map problem for some specific function classes. To prepare for convergence results, we dedicate Section 4.1 to regularity properties of the transport cost in the discrete case. In Section 4.2, we present methods in the case where G is the class of L-Lipschitz gradients of ℓ-strongly convex potentials (presented in Section 2.3). For the squared Euclidean cost, these methods were introduced in [37], using convex interpolation results from [52]. In Section 4.3, we consider a simple kernel method which solves a regularised version of Eq. (3). This type of method hinges on the fact that kernel methods yield a finite-dimensional parametrisation of the function g, and allows for provably convergent gradient descent methods. In Section 4.4, we consider a Stochastic Gradient Descent method for the case where the map g is a Neural Network. Finally, in Section 4.5, we illustrate the use of the methods presented in this section on the problem of colour transfer.

4.1 Regularity of Discrete Optimal Transport Costs

To study the convergence of sub-gradient descent methods theoretically, we will introduce standard notions from non-smooth non-convex analysis, in particular a specific generalisation of sub-gradients, which are in practice computed by automatic differentiation. A central notion in this analysis will be the notion of semi-algebraicity, which we remind in Definition 4.1 (and refer to [56] and [10] for more details).

Definition 4.1. *A set $S \subset \mathbb{R}^d$ is said to be semi-algebraic if it can be written under the form $S = \cup_{n=1}^{N} \cap_{m=1}^{M} S_{n,m}$, where each $S_{n,m}$ is either of the form $\{x \in \mathbb{R}^d : p_{n,m}(x) = 0\}$ or $\{x \in \mathbb{R}^d : p_{n,m}(x) \geq 0\}$, where $p_{n,m}$ is a d-variate polynomial with real coefficients.*

A function $f : \mathbb{R}^{d_1} \longrightarrow \mathbb{R}^{d_2}$ is semi-algebraic if its graph $\{(x, f(x)) : x \in \mathbb{R}^{d_1}\}$ is a semi-algebraic set.

A multifunction $f : \mathbb{R}^{d_1} \rightrightarrows \mathbb{R}^{d_2}$ is semi-algebraic if its graph $\{\{x\} \times f(x) : x \in \mathbb{R}^{d_1}\}$ is a semi-algebraic set.

Another central notion will be a generalisation of the notion of gradient for locally Lipschitz functions

called the Clarke differential.

Definition 4.2. *Given a locally Lipschitz function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, its Clarke sub-gradient at $x \in \mathbb{R}^d$ is the set*

$$\partial_C f(x) = \text{conv}\left\{ \lim_{t \longrightarrow +\infty} \nabla f(x_t) : x_t \xrightarrow[t \longrightarrow +\infty]{} x,\ x_t \in D_f \right\},$$

where D_f is the set of differentiability of f and conv *denotes the convex envelope.*

In Proposition 4.3, we show that the discrete OT cost is semi-algebraic and locally Lipschitz as a function of the cost matrix, and that its Clarke sub-gradient is itself semi-algebraic.

Proposition 4.3. *Consider weights $a \in \Delta_n$ (the n-simplex) and $b \in \Delta_m$ (the m-simplex), and the discrete Kantorovich cost function*

$$W(a,b,\cdot) := \left\{ \begin{array}{ccc} \mathbb{R}^{n \times m} & \longrightarrow & \mathbb{R} \\ M & \longmapsto & \min\limits_{\pi \in \Pi(a,b)} \pi \cdot M \end{array} \right. .$$

Then the map $W(a,b,\cdot)$ is semi-algebraic, Lipschitz, and its Clarke sub-gradient is semi-algebraic and writes for $M \in \mathbb{R}^{n \times m}$:

$$\partial_C W(a,b,\cdot)(M) = \underset{\pi \in \Pi(a,b)}{\text{argmin}}\ \pi \cdot M.$$

Proof. Writing the extremal points of the polytope $\Pi(a,b)$ as $(\pi_i)_{i=1}^N$, the function $W(a,b,\cdot)$ can be written as a finite minimisation over $\pi \in (\pi_i)_{i=1}^N$ of linear functions of M, hence $W(a,b,\cdot)$ is semi-algebraic. The fact that $W(a,b,\cdot)$ is Lipschitz is also its consequence of its expression as a minimum of a finite amount of linear maps. The expression of sub-gradients of $W(a,b,\cdot)$ is a consequence of Danskin's Theorem [14]. Finally, the semi-algebraic property of $\partial_C W(a,b,\cdot)$ is a consequence of the fact that $W(a,b,\cdot)$ is semi-algebraic and locally Lipschitz, or can alternatively be seen using the extremal point method as done for $W(a,b,\cdot)$. □

In Lemma 4.4 we remind some useful properties of semi-algebraic maps that we will use later on. In particular, semi-algebraic maps are *generalised differentiable* (see [22] Definition 3.1), which can be understood as a generalised first-order Taylor expansion.

Lemma 4.4. *Any locally Lipschitz semi-algebraic map $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is generalised differentiable (see [22] Definition 3.1) and its set of critical values $f\{x \in \mathbb{R}^d : 0 \in \partial_C f(x)\}$ is finite.*

Proof. Since f is semi-algebraic and locally Lipschitz, by [8] Theorem 3.6 it is semi-smooth, which in turn implies generalised differentiability (by [35] Theorem 1.4). Next, by definable Morse-Sard (from [10] Theorem 5), the set of critical values $f\{x \in \mathbb{R}^d : 0 \in \partial_C f(x)\}$ is finite. □

4.2 Numerical Method for Gradients of Convex Functions

In this section, we present numerical methods to solve the approximate problem in the case of the function class $\mathcal{F}_{\mathcal{E},L,\ell}$ of functions $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ that is L-Lipschitz and gradient of an ℓ-strongly convex function $\varphi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$, on each part E_k of the fixed partition $\mathcal{E}$. We already introduced this class in Section 2.3, and it was first considered in the context of map problems by [37]. The numerical methods will aim to solve the problem

$$\underset{\varphi \in \mathcal{F}_{\mathcal{E},L,\ell}}{\text{argmin}}\ \mathcal{T}_c(g\#\mu, \nu), \tag{24}$$

with a particular emphasis on the case where c is quadratic, i.e. $c(x,y) = (x-y)^T Q(x-y) + b^T(x-y)$, where $Q \in S_d^+(\mathbb{R})$ is a positive-semi-definite matrix, and $b \in \mathbb{R}^d$. For our numerical questions, we consider the discrete case

$$\mu = \sum_{i=1}^n a_i \delta_{x_i}, \quad \nu = \sum_{j=1}^m b_j \delta_{y_j}.$$

Obviously, we need to assume that the measure μ is compatible with the partition, which is to say the the x_i are never at the boundary of a part E_k: $\forall i \in [\![1,n]\!],\ x_i \in (\cup_k \partial E_k)^c$. The objective in Eq. (24) only depends on the values $\varphi_i := \varphi(x_i)$ and $g_i := g(x_i)$, the immediate question is that given a candidate $(\varphi_i, g_i) \in (\mathbb{R} \times \mathbb{R}^d)^n$, does there exist a function $g \in \mathcal{F}_{\mathcal{E},L,\ell}$ of the form $\nabla\varphi$ which interpolates these values, i.e. $g(x_i) = g_i$ and $\varphi(x_i) = \varphi_i$? This question, which is called $\mathcal{F}_{\mathcal{E},L,\ell}$-interpolation, was studied by Taylor [52][5]. We write $\mathcal{F}_{L,\ell} := \mathcal{F}_{\mathcal{E},L,\ell}$ in the case $\mathcal{E} = \{\mathbb{R}^d\}$, and present the results in the restricted case where the space is $\mathbb{R}^d$, as opposed to any vector space.

Proposition 4.5. *(Multiple results from Taylor [52, 53]). Let $S = (x_i, g_i, \varphi_i)_{i\in[\![1,n]\!]} \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^n$. The set S is said to be $\mathcal{F}_{L,\ell}$-interpolable ([52], Definition 3.1) if there exists $\varphi \in \mathcal{F}_{L,\ell}$ such that $\forall i \in [\![1,n]\!],\ \nabla\varphi(x_i) = g_i$ and $\varphi(x_i) = \varphi_i$. Consider the quadratic function*

$$Q(x, x', \varphi, \varphi', g, g') := \varphi - \varphi' - \langle g', x - x'\rangle - c_1\|g - g'\|_2^2 - c_2\|x - x'\|_2^2 + c_3\langle g' - g, x' - x\rangle, \tag{25}$$

for $x, x' \in \mathbb{R}^d,\ \varphi, \varphi' \in \mathbb{R},\ g, g' \in \mathbb{R}^d$, with

$$c_1 := \frac{1}{2L(1-\ell/L)}, \quad c_2 := \frac{\ell}{2(1-\ell/L)}, \quad c_3 := \frac{\ell}{L(1-\ell/L)}.$$

- *([52], Theorem 3.8) The set S is $\mathcal{F}_{L,\ell}$-interpolable if and only if for all $i, j \in [\![1,n]\!]$,*

$$Q(x_i, x_j, \varphi_i, \varphi_j, g_i, g_j) \geq 0. \tag{26}$$

- *([52], Theorem 3.14) For $x \in \mathbb{R}^d$, let:*

$$\varphi_l(x) = \min_{t\in\mathbb{R},\ g\in\mathbb{R}^d} t, \quad \text{s.t. } \forall j \in [\![1,n]\!],\ Q(x, x_j, t, \varphi_j, g, g_j) \geq 0; \tag{27}$$

$$\varphi_u(x) = \max_{t\in\mathbb{R},\ g\in\mathbb{R}^d} t, \quad \text{s.t. } \forall i \in [\![1,n]\!],\ Q(x_i, x, \varphi_i, t, g_i, g) \geq 0. \tag{28}$$

If S is $\mathcal{F}_{L,\ell}$-interpolable, then any interpolating function φ satisfies $\varphi_l \leq \varphi \leq \varphi_u$, and the potentials φ_l, φ_u are valid interpolations.

Proposition 4.5 shows that the constraint on $(\varphi_i, g_i)_i$ can be written as a set of quadratic constraints. It follows immediately that any problem that only depends on the values $g(x_i)$ for a variable $G \in \mathcal{F}_{\mathcal{E},L,\ell}$ can be written as a problem over $(\varphi_i, g_i)_i$ under quadratic constraints, as stated in Corollary 4.6.

Corollary 4.6. *Consider an objective $J : \mathcal{F}_{\mathcal{E},L,\ell} \longrightarrow \mathbb{R}_+$ such that for $g \in \mathcal{F}_{\mathcal{E},L,\ell}$, the value $J(g)$ can be written $J(g(x_1), \cdots, g(x_n))$. Then the problem*

$$\min_{g\in\mathcal{F}_{\mathcal{E},L,\ell}} J(g) \tag{29}$$

is equivalent to the problem

$$\min_{\substack{\varphi_1,\cdots,\varphi_n\in\mathbb{R}\\ g_1,\cdots,g_n\in\mathbb{R}^d}} J(g(x_1), \cdots, g(x_n)) \tag{30}$$
$$s.t.\ \forall k \in [\![1,K]\!],\ \forall i,j \in I_k : Q(x_i, x_j, \varphi_i, \varphi_j, g_i, g_j) \geq 0,$$

where $I_k := \{i \in [\![1,n]\!] \mid x_i \in E_k\}$, and Q is defined in Eq. (25). Given a solution $(\varphi_i^*, g_i^*)_i$ of Eq. (30), any solution $\nabla\varphi^*$ of Eq. (29) satisfies $\varphi_l \leq \varphi^* \leq \varphi_u$ on $\cup_k \mathring{E}_k$, where for $x \in \mathring{E}_k$, the bounding potentials and their gradients are solutions of:

$$(\varphi_l(x), \nabla\varphi_l(x)) = \operatorname*{argmin}_{t\in\mathbb{R},\ g\in\mathbb{R}^d} t, \quad \text{s.t. } \forall j \in I_k,\ Q(x, x_j, t, \varphi_j^*, g, g_j^*) \geq 0; \tag{31}$$

[5] Note that ([52], Theorem 3.14) writes an erroneous argmin for φ_u: in the light of ([52], Remark 3.13), it should instead read argmax, especially given the fact that the minimisation problem is unbounded.

$$(\varphi_u(x), \nabla\varphi_u(x)) = \underset{t\in\mathbb{R},\, g\in\mathbb{R}^d}{\operatorname{argmax}}\ t, \quad \text{s.t. } \forall i \in I_k,\ Q(x_i, x, \varphi_i^*, t, g_i^*, g) \geq 0. \tag{32}$$

The potentials (φ_l, φ_u) themselves are both solutions of Eq. (29).

Note that the values of the potentials can be chosen arbitrarily on the boundaries ∂E_k.

We can now provide an algorithm for $\operatorname{argmin}_{g\in\mathcal{F}_{\mathcal{E},L,\ell}} \mathcal{T}_c(g\#\mu, \nu)$ (Eq. (24)) using Corollary 4.6: the objective is

$$J(g(x_1), \cdots g(x_n)) = \min_{\pi\in\Pi(a,b)} \sum_{i,j} \pi_{i,j} c(g(x_i), y_j), \tag{33}$$

and the resulting problem defined in Eq. (30) can be solved by alternating over π (solving a discrete Kantorovich problem, using `ot.emd` from the PythonOT library, for instance [26]), and over (φ_i, g_i), for which the constraints are quadratic and the objective depends on the cost c. For smooth cost, one may use projected gradient descent, and for (convex) quadratic costs, the problem becomes a (convex) Quadratically Constrained Quadratic Program (QCQP). In the case $c(x, y) = \|x - y\|_2^2$, this method is already known, and is the core contribution of [37] summarised in Algorithm 1, with a generalisation to convex quadratic costs $c_P(x, y) := (x - y)^T P(x - y)$ with $P \in S_d^{++}(\mathbb{R})$ (a positive-definite symmetric matrix). We remind the notation $\Pi^*_{c_P}(g\#\mu, \nu)$ as the set of optimal couplings between $g\#\mu$ and ν for the cost c_P.

Algorithm 1: Alternate Minimisation for the Gradient of Strongly Convex Functions.

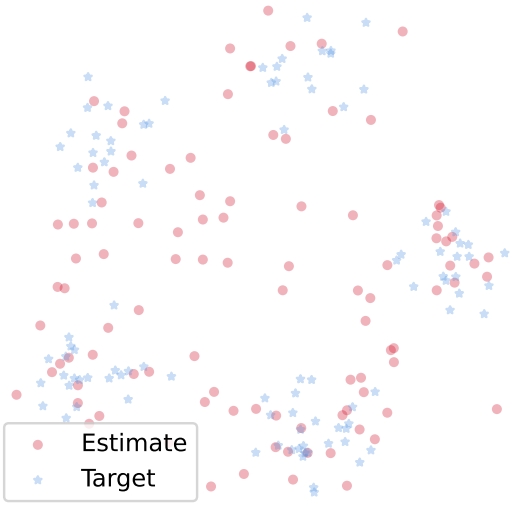
Data: Strongly convex constant $\ell \geq 0$, Lipschitz constant $L \geq \ell$, disjoint point classes $I_k \subset [\![1, n]\!]$ and discrete probability distributions $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$.

1 **Initialisation:** Compute $\pi \in \Pi^*_{c_P}(\mu, \nu)$.;
2 **for** $t \in [\![0, T_{\max} - 1]\!]$ **do**
3 Update $(\varphi_i, g_i)_{i\in[\![1,n]\!]}$ by solving the QCQP:
4 $\min_{\substack{\varphi_1,\cdots,\varphi_n\in\mathbb{R} \\ g_1,\cdots,g_n\in\mathbb{R}^d}} \sum_{i,j}(g_i - y_j)^T P(g_i - y_j)\pi_{i,j}$
5 s.t. $\forall k \in [\![1, K]\!],\ \forall i, j \in I_k : Q(x_i, x_j, \varphi_i, \varphi_j, g_i, g_j) \geq 0.$
6 Update π by solving the discrete Kantorovich problem: $\pi \in \Pi^*_{c_P}(g\#\mu, \nu)$.
7 **end**
8 **Return**: $(\varphi_i, g_i)_{i\in[\![1,n]\!]}$.

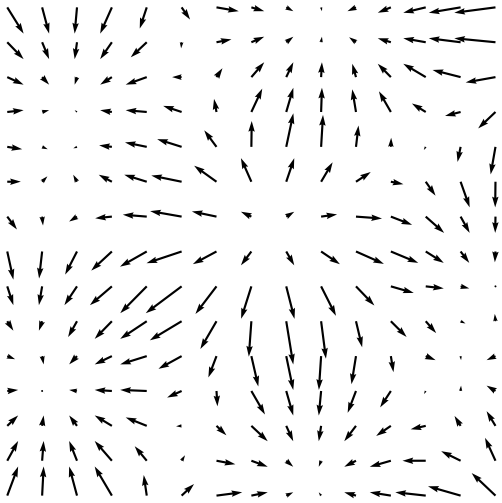
Time complexity. From a time complexity standpoint, the QCQP problem at lines 3-4-5 bears a substantial cost. As a coarse analysis, standard methods such as [57] have $\mathcal{O}(L^2N^4)$ complexity, where N is the dimension of variables, here $N = (d + 1)n$, and where $L = N^2 + NM + R$, where M is the number of constraints, here $M = \sum_k \#I_k^2$, and $R = [\log |T|]$, with T the sum of the non-zero integers in the float representation of P and the constraint matrix. For simplicity, we will continue with $K = 1$ and thus $M = n^2$. This yields the final (prohibitive) complexity: $\mathcal{O}\left((n(d + 1) + n^3 + R)^2(d + 1)^4n^4\right)$. For the transport cost, using the network simplex (see the explanation in [39] Section 3.5), omitting multiplicative logarithmic terms, the time complexity of solving the linear Kantorovich problem between measures with n and m points and cost matrix M is $\mathcal{O}((n + m)nm \log(n + m) \log((n + m)\|M\|_\infty))$ [51].

In Fig. 8, we present a numerical example of the method for a map fitting a two-dimensional standard Gaussian to a two-dimensional Gaussian Mixture. Since no public implementation of the QCQP problem from [37] is available, we contributed a solver for Algorithm 1 for the squared-Euclidean cost in the Python OT library [26], with an example [6].

[6] https://pythonot.github.io/auto_examples/others/plot_SSNB.html#sphx-glr-auto-examples-others-plot-ssnb-py



(a) Positions of the optimal value positions (g_i^*) for the map estimated with Algorithm 1.



(b) Evaluation of the learned map $g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ from Algorithm 1 on a grid.

Figure 8: Illustration of the method described in Algorithm 1 for the map problem from samples of a standard Gaussian distribution to samples of a Gaussian mixture. The map g is constrained to be 2-Lipschitz and the gradient of a 1/2-strongly convex function. The cost is taken as $c(x, y) = \|x - y\|_2^2$. Note that due to the constraints, we obtain an inexact matching, with in particular leakage between the modes of the target distribution.

4.3 Numerical Method for Maps in a RKHS

We introduce a relatively straightforward kernel method to solve the map problem (Eq. (3)). We fix a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions $\mathcal{X} \longrightarrow \mathbb{R}^d$ of kernel $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}^{d\times d}$. We denote by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ the inner product on $\mathcal{H}$, and $\| \cdot \|_{\mathcal{H}}$ the associated RKHS norm on $\mathcal{H}$.

Given discrete measures $\mu = \sum_{i=1}^n a_i \delta_{x_i} \in \mathcal{P}(\mathcal{X})$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j} \in \mathcal{P}(\mathbb{R}^d)$, we will solve a regularised variant of the map problem (Eq. (3)):

$$\operatorname*{argmin}_{h\in\mathcal{H}} \mathcal{T}_c(h\#\mu, \nu) + \lambda \|h\|_{\mathcal{H}}^2, \tag{34}$$

for some constant $\lambda > 0$ that penalises the norm of h, which equates to imposing regularity on the function h. Given the support of μ, the cost $\mathcal{T}_c(h\#\mu, \nu)$ only depends on $(h(x_1), \cdots, h(x_n))$. A well known reduction method in RKHS theory (detailed in Appendix A.3 for completeness) then allows to look for solutions in an n-dimensional linear subspace V of $\mathcal{H}$:

$$V := \left\{ \sum_{k=1}^n K(\cdot, x_k) u_k \ : \ \forall k \in [\![1, n]\!],\ u_k \in \mathbb{R}^d \right\}, \quad \text{of} \quad \operatorname*{argmin}_{h\in V} \mathcal{T}_c(h\#\mu, \nu) + \lambda \|h\|_{\mathcal{H}}^2. \tag{35}$$

Since any element $h \in V$ is characterised by its coefficients $(u_1, \cdots, u_n) \in (\mathbb{R}^d)^n$, we can formulate Eq. (35) as a problem over the (u_i). First, using the kernel reproducing property, we compute

$$\left\| \sum_{k=1}^n K(\cdot, x_k) u_k \right\|_{\mathcal{H}}^2 = \sum_{k=1}^n \sum_{l=1}^n u_k^T K(x_k, x_l) u_l. \tag{36}$$

Concerning the transport cost term, we remind the notation for the value of the Kantorovich discrete problem

$$\mathrm{W}(a, b, M) := \min_{\pi\in\Pi(\mu,\nu)} M \cdot \pi,$$

and in this case, the cost matrix M can be computed using the expression

$$\forall (i, j) \in [\![1, n]\!] \times [\![1, m]\!],\ M_{i,j} = c\left(\sum_{k=1}^n K(x_i, x_k) u_k, y_j \right). \tag{37}$$

The dependency in the (u_i) lies in the cost M. Numerically, provided that c is sufficiently regular, this allows a minimisation through classical algorithms such as gradient descent, using differentiable implementations of the discrete Kantorovich cost, such as `ot.emd2` [26]. By introducing the $nd \times nd$ matrix $\mathbf{K}$ defined by $n \times n$ blocks $K(x_i, x_j)$ of size $d \times d$:

$$\mathbf{K} = \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix}, \tag{38}$$

and the stacked vector $\mathbf{u} \in \mathbb{R}^{nd}$, Eqs. (36) and (37) can be re-written as matrix products. This yields our final expression for Eq. (34):

$$\min_{\mathbf{u} \in \mathbb{R}^{nd}} \mathrm{W}(a, b, M(\mathbf{u})) + \lambda \mathbf{u}^T \mathbf{K} \mathbf{u}, \quad M(\mathbf{u})_{i,j} := c\left(\mathbf{K}_{[i,:]}\mathbf{u}, y_j\right), \tag{39}$$

where $\mathbf{K}_{[i,:]}$ denotes the sub-matrix of $\mathbf{K}$ with the n lines $((i-1)d + 1, \cdots, id)$, which corresponds to the i-th $d \times d$ block line of $\mathbf{K}$. Given optimal coefficients $\mathbf{u} = (u_1, \cdots, u_n) \in (\mathbb{R}^d)^n$, a solution h is defined everywhere in $\mathcal{X}$ using the kernel:

$$\forall x \in \mathcal{X},\ h(x) = \sum_{i=1}^{n} K(x, x_i) u_i.$$

Remark 4.7. *The only constraints that are imposed upon a solution of Eq. (34) come from the choice of the kernel K (or equivalently of the space $\mathcal{H}$) and of the regularisation coefficient $\lambda > 0$. A natural idea would be to add a regularisation term $R(h)$, for instance to enforce h to be a gradient of a convex function. For Lemma A.5 to apply, one would need to have a regularisation which only depends on the values $(h(x_i))$, which is very restrictive. A possible heuristic would be to look for $h \in V$ regardless of this property on R, however the resulting problem would have no theoretical link to the problem over $h \in \mathcal{H}$, unlike in our case. Finally, a regularisation which depends on an infinite amount of values $h(x)$ are numerically challenging, in the specific case of dense inequality constraints, we refer to [44] as a useful tool.*

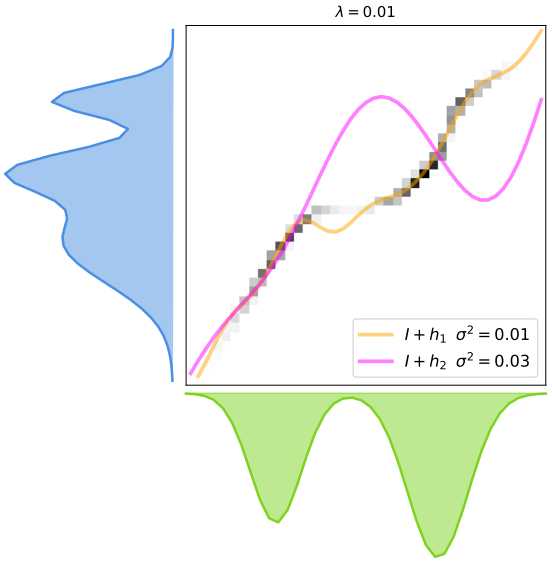
Remark 4.8. *A natural idea is to consider class of functions that are perturbations of a simple map, for instance $g = sI + h$, where h is in a RKHS $\mathcal{H}$, and $s > 0$ is a scale factor. Given Lemma A.5, this tweak comes without numerical or theoretical cost.*

We illustrate this kernel method in Fig. 9 with the Gaussian kernel $K(x, y) = \exp\left(-\|x - y\|_2^2/(2\sigma^2)\right) I_d$ and maps of the form $g = I + h$, where h is in the RKHS generated by the Gaussian kernel.

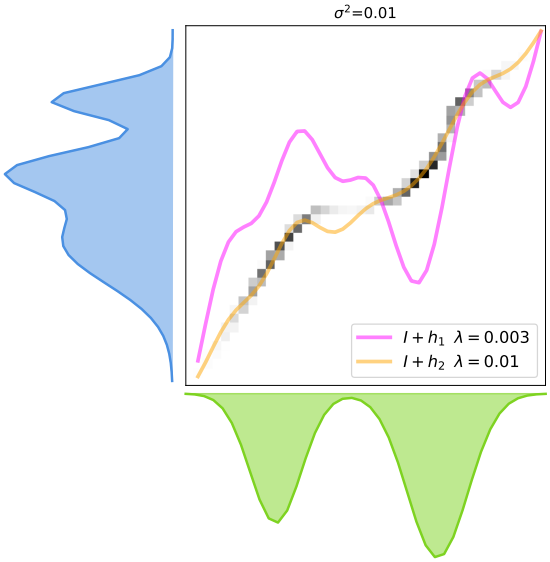
From an algorithmic viewpoint, we propose in Algorithm 2 a simple (sub-)gradient descent method (GD) for the discrete kernel map problem Eq. (39), and provide a convergence result in Proposition 4.9 using results from Section 4.1 and Ermoliev-Norkin [22].

Concerning time complexity, there are two main bottlenecks: the matrix-vector computations $\mathbf{Ku}$ which incur a $\mathcal{O}(n^2 d^2)$ cost, and solving the discrete Kantorovich problem, which is in $\mathcal{O}((n + m)nm \log(n + m) \log((n + m)\|M\|_\infty))$, as discussed in Section 4.2. Note that for memory efficiency, one may use map-reduce methods such as proposed in [12] to avoid storing the matrix $\mathbf{K}$, at the cost of a higher time complexity. For scalar kernels $K(x, x') = k(x, x')I_d$, it suffices to store $\mathbf{K} := (k(x_i, x_j))_{i,j} \in \mathbb{R}^{n \times n}$, reducing the memory complexity to $\mathcal{O}(n^2 + nd)$, and matrix-vector products to $\mathcal{O}(n^2 d)$.

Proposition 4.9 (Convergence of GD for the Kernel Method, application of [22] Theorem 4.1)**.** *Take a locally Lipschitz and semi-algebraic (see Definition 4.1) cost function c, and gradients steps $\alpha_t > 0$ such that $\alpha_t \to 0$ and $\sum_t \alpha_t = +\infty$. The iterates $(\mathbf{u}_t)$ of Algorithm 2 are such that any accumulation point $\mathbf{v}$ is Clarke critical: $0 \in \partial_C J(\mathbf{v})$, with $J(\mathbf{u}) := W(a, b, M(\mathbf{u})) + \lambda \mathbf{u}^T \mathbf{K} \mathbf{u}$.*



(a) Kernel map solution for a regularisation $\lambda = 0.01$ and multiple scales σ^2.



(b) Kernel map solution for a scale $\sigma^2 = 0.01$ and multiple regularisations λ.

Figure 9: Illustration of the kernel method for the map problem between two Gaussian mixtures, using the Gaussian kernel. In greyscale, the OT plan is represented for reference.

Algorithm 2: GD on the Kernel Map Parameters.

Data: Gradient steps $\alpha_t > 0$, kernel regularisation $\lambda > 0$, discrete probability distributions $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$, kernel function K.

Pre-Processing: Compute the matrix $\mathbf{K}$ from Eq. (38);

Initialisation: Draw $\mathbf{u}_0 \in \mathbb{R}^{nd}$;

for $t \in [\![0, T_{\max} - 1]\!]$ **do**

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \alpha_t \left[\frac{\partial}{\partial \mathbf{u}} \Big(W(a, b, M(\mathbf{u})) + \lambda \mathbf{u}^T \mathbf{K} \mathbf{u} \Big) \right]_{\mathbf{u}=\mathbf{u}_t} \quad \text{(see Eq. (39))}.$$

end

Proof. First, by Proposition 4.3 and by semi-algebraicity and local Lipschitzness of c, J is locally Lipschitz and semi-algebraic. Using now Lemma 4.4, we have the sufficient regularity conditions to use the convergence result of [22] (Theorem 4.1). □

4.4 Gradient Descent for Neural Networks

We now consider the case where the function class G is the class of neural networks introduced in Eq. (8), with parameters in a compact set $\Theta \subset \mathbb{R}^p$, which we also assume to be convex. We will introduce a technical modification of the neural network from Eq. (8) and consider the parametrised function:

$$h := (\theta, x) \longmapsto g_{P_\Theta(\theta)}(x), \tag{40}$$

with g the map defined Eq. (8), and where the map $P_\Theta : \mathbb{R}^p \longrightarrow \Theta$ denotes the orthogonal projection onto Θ. This re-writing allows us to define the NN h on all of $\mathbb{R}^p \supset \Theta$. In practice, SGD with this network is essentially equivalent to projecting the parameters after each gradient step (with the technicality that in our formalism, the gradient of P_Θ is included in the backpropagation).

To solve the map problem of minimising $\mathcal{T}_c(h(\theta, \cdot)\#\mu, \nu)$ in practice, we consider a commonly used minibatch surrogate loss $F(\theta)$, which we define in Eq. (41). Given a dataset $X^{(n)} \in \mathbb{R}^{n \times k} = (x_1, \cdots, x_n)$, we will denote abusively $h(\theta, X^{(n)}) \in \mathbb{R}^{n \times d} := (h(\theta, x_1), \cdots, h(\theta, x_n))$. Given a dataset $X^{(n)} \in \mathbb{R}^{n \times k}$, the measure $\delta_{X^n} \in \mathcal{P}(\mathbb{R}^k)$ denotes $\frac{1}{n} \sum_i \delta_{x_i}$. Similarly, we will denote a target dataset

$Y^{(m)}$. The loss F we consider is

$$F(\theta) := \int \mathcal{T}_c(\delta_{h(\theta,X^{(n)})}, \delta_{Y^{(m)}}) \mathrm{d}\mu^{\otimes n}(X^{(n)}) \mathrm{d}\nu^{\otimes m}(Y^{(m)}). \tag{41}$$

The loss F will be minimised by Stochastic Gradient Descent over θ, where the stochasticity is on the data batches $X^{(n)}$ and $Y^{(m)}$, as described in Algorithm 3.

Algorithm 3: Training a NN map for the cost $\mathcal{T}_c$.

Data: Gradient steps $\alpha_t > 0$, probability distributions $\mu \in \mathcal{P}(\mathbb{R}^k)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, NN $h(\theta, .)$.

1 **Initialisation:** Draw $\theta_0 \in \Theta$;
2 **for** $t \in [\![0, T_{\max} - 1]\!]$ **do**
3 Draw $X^{(n)} \sim \mu^{\otimes n}$, $Y^{(m)} \sim \nu^{\otimes m}$.
4 $\theta_{t+1} = \theta_t - \alpha_t \left[\frac{\partial}{\partial \theta} \mathcal{T}_c(\delta_{h(\theta,X^{(n)})}, \delta_{Y^{(m)}}) \right]_{\theta=\theta_t}$.
5 **end**

An important remark is that this formalism bears strong similarities to the alternate minimisation framework studied in Section 3 for the squared-Euclidean cost. Indeed, it can be seen as an alternation of the map parameters θ and the (minibatch) OT plan π in $\mathcal{T}_c$ (line 4): the optimisation over π is done by solving the linear program when computing the cost $\mathcal{T}_c$, and then one gradient step of optimisation over θ is performed. To study Algorithm 3 theoretically, we will give precise meaning to the partial derivative at line 4 using the notions introduced in Section 4.1. Numerically, the sub-gradients in question are computed by automatic differentiation. Note that P_Θ is Lipschitz and semi-algebraic.

Thanks to the regularity result on the OT cost proved in Proposition 4.3, we can use recent SGD convergence results by Bolte, Le and Pauwels [9] to show that the iterates of Algorithm 3 converge in a certain sense. First, to give sense to the gradient in Algorithm 3, we remark that for locally Lipschitz semi-algebraic activation functions, the map $h(\cdot, \cdot)$ is semi-algebraic and locally Lipschitz. By composition using Proposition 4.3, the sample loss function:

$$f(\cdot, X^{(n)}, Y^{(m)}) := \theta \longmapsto \mathcal{T}_c(\delta_{h(\theta,X^{(n)})}, \delta_{(Y^{(m)})}),$$

is locally Lipschitz and semi-algebraic. We can select a semi-algebraic sub-gradient $\varphi : \mathbb{R}^p \times \mathbb{R}^{n\times k} \times \mathbb{R}^{m\times d} \longrightarrow \mathbb{R}^p$ such that

$$\forall \theta \in \mathbb{R}^p,\ \forall X^{(n)} \in \mathbb{R}^{n\times k},\ \forall Y^{(m)} \in \mathbb{R}^{m\times d},\ \varphi(\theta, X^{(n)}, Y^{(m)}) \in \partial_C f(\theta, X^{(n)}, Y^{(m)}),$$

where the selection can be done by lexicographic order on coordinates, for example. Note that $f(\cdot, X^{(n)}, Y^{(m)})$ is differentiable almost-everywhere, and that at differentiable points, φ equates its usual gradient. The choice of sub-gradient performed by automatic differentiation satisfies this condition (see a discussion on this procedure in [10, 15].) We remind that the population loss function is $F = \theta \longmapsto \int f(\theta, X^{(n)}, Y^{(m)}) \mathrm{d}\mu^{\otimes n}(X^{(n)}) \mathrm{d}\nu^{\otimes m}(Y^{(m)})$ in this setting.

Proposition 4.10 (Convergence of SGD for NN maps, application of [9] Theorem 3)**.** *Assume that μ, ν are discrete measures or compactly supported measures with semi-algebraic densities with respect to the Lebesgue measure. Assume that the NN h is defined as in Eq. (40), with locally Lipschitz semi-algebraic activation functions. Assume that Θ is compact, convex and semi-algebraic. Suppose that the cost function $c : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}_+$ is locally Lipschitz and semi-algebraic. Take gradient steps $(\alpha_t)_{t\in\mathbb{N}} \in (0, +\infty)^{\mathbb{N}}$ such that $\sum_t \alpha_t = +\infty$ with $\alpha_t = o(1/\log(t))$.*

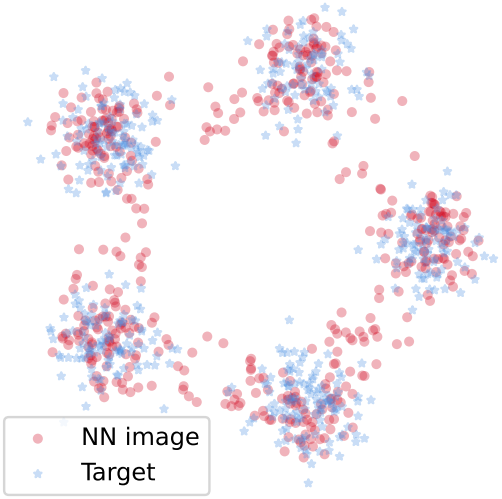
Then there exists a set of possible steps $A \subset (0, +\infty)$ whose complement is finite, and a set of possible initialisations $\Theta_0 \subset \Theta$ of full measure, such that for each step sequence $(\alpha_t) \in A^{\mathbb{N}}$ verifying the conditions, the stochastic gradient descent iterates:

$$\theta_0 \in \Theta_0,\ \forall t \in \mathbb{N},\ \theta_{t+1} = \theta_t - \alpha_t \varphi(\theta_t, X_t^{(n)}, Y_t^{(m)}),\ X_t^{(n)} \sim \mu^{\otimes n},\ Y_t^{(m)} \sim \nu^{\otimes m},$$

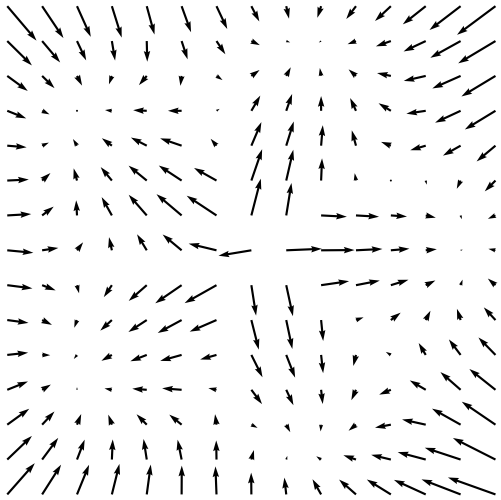
verify that almost-surely, $(F(\theta_t))$ converges, and almost-surely any accumulation point $\overline{\theta}$ of (θ_t) is such that $0 \in \partial_C F(\overline{\theta})$, under the (mild) additional assumption that that the trajectories (θ_t) are almost-surely bounded.

Proof. We apply Bolte-Le-Pauwels [9], Theorem 3 to the NN h with the discrete OT loss from Proposition 4.3. Thanks to the assumptions formulated in the result statement, to Proposition 4.3 and to the construction of a semi-algebraic sub-gradient selection φ, we have collected all the conditions for Bolte-Le-Pauwels, Theorem 3, yielding the result. For the case where one or both of $\{\mu, \nu\}$ is/are discrete, we applied their Remark 3. □

In Fig. 10, we present a numerical example of the method for a map fitting a two-dimensional standard Gaussian to a two-dimensional Gaussian Mixture.



(a) Comparison of images of samples of the source Gaussian by the learned NN map to the target Gaussian Mixture.



(b) Evaluation of the learned NN map $g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ on a grid.

Figure 10: Illustration of the method described in Algorithm 3 for the map problem from samples of a standard Gaussian distribution to samples of a Gaussian mixture. The map g is of the form $g = I + h$, where h is a small 4-layer NN with ReLU activation functions, and weights constrained to $[-1/2, 1/2]$. The cost is taken as $c(x, y) = \|x - y\|_2^2$. Note that due to the (indirect) constraint on the Lipschitz constant, we obtain an inexact matching, with in particular leakage between the modes of the target distribution.

4.5 Illustrative Application to Colour Transfer

In this section, we consider the task of colour transfer, which consists in transforming the colour distribution of a source image onto the colour distribution of a target image. An $(n \times m)$ RGB image is seen as a 3-tensor $\mathbf{I} \in [0, 1]^{n \times m \times 3}$, and its colour distribution is then a discrete measure $\mu = \frac{1}{nm} \sum_{i,j} \delta_{\mathbf{I}_{i,j}}$ on $\mathbb{R}^3$.

We illustrate that a learned map $g : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ which is optimised to transfer the colours of a source image $\mathbf{I}_s$ onto a target image $\mathbf{I}_t$ can be used on a new image $\mathbf{I}$ to transfer its colours. This is made possible since the map g is defined everywhere, and not only at the points of μ. We consider the cost $c(x, y) = \|x - y\|_2^2$, and a simple NN map g using Algorithm 3 on the source and target images, and then apply the map to new images. We present the results in Fig. 11 for three different training tasks. Notice how the constraint on the map g allows us to have a colour transfer that is robust to outliers. In Fig. 12 in Appendix A.4, we present the results in the RGB space, seeing the images as pixel point clouds.

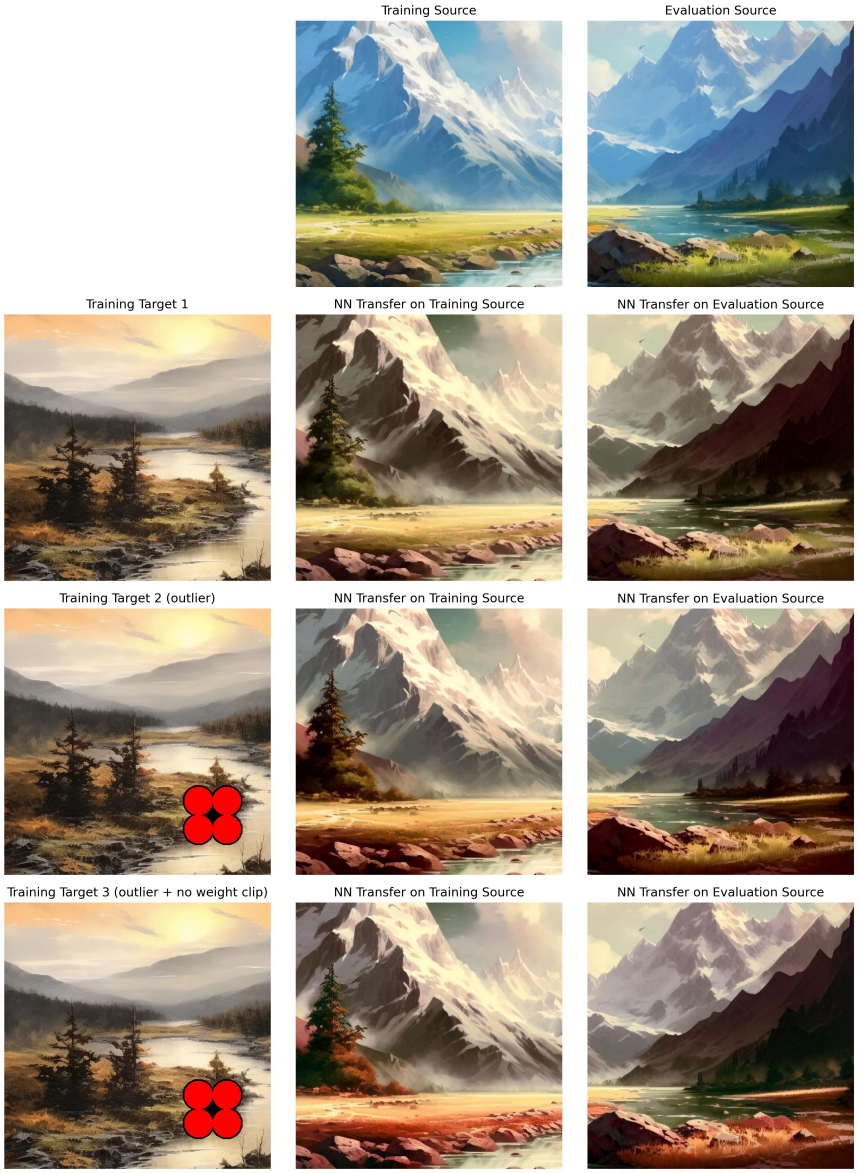


Figure 11: Colour transfer using a NN map trained on the task of transferring the colour distribution of the Training Source image onto three different Training Target images (column 1, rows 2 to 4). In the second training (row 3), the target image presents an outlier in the colour distribution. In the third training (row 4), the same target image with outliers is used, this time with no weight clipping. For all trainings, the learned map is applied to the training image (second column) and to the test image "Evaluation Source" (column 3).

5 Conclusion and Outlook

In this paper, we have considered the problem of finding an optimal transport map g between two probability measures μ and ν under the constraint that $g \in G$, where G is a given set of functions (L-Lipschitz, gradient of a convex function, for instance). We have given general assumptions to ensure the existence of an optimal map g, we have studied the relationship between our problem and many other concepts in Optimal Transport, and also the link with kernel methods. We have also explained how to solve the problem from a practical point a view with convergence guarantees, and an application to colour transfer.

We believe that there are two important but difficult questions that should be investigated in future work. The first is the question of the uniqueness of an optimal map. We have given a partial answer to this question, but it seems to be a difficult question in its whole generality. Having a result of

uniqueness would then open the way to new questions, such as the use of g to compare measures in a way similar to Linearised Optimal Transport, or the study of the statistical properties of g (related to the sample complexity). The second important question is the addition of constraints in the kernel method, more precisely: how to translate a set of functions G (like the set of gradients of convex functions for instance) into a RKHS representation?

Acknowledgements

We extend our warmest thanks to Nathaël Gozlan for his valuable input regarding technical assumptions for the existence result. We thank Joan Glaunès for the fruitful time we spent working together on kernel problems. We also want to thank Tam Le for providing references and insight for non-smooth and non-convex optimisation questions. We are grateful for the remarks and feedback of two anonymous reviewers, which in particular allowed significant improvement of the map existence results.

This research was funded in part by the Agence nationale de la recherche (ANR), Grant ANR-23-CE40-0017 and by the France 2030 program, with the reference ANR-23-PEIA-0004.

References

[1] Anshul Adve and Alpár Mészáros. On nonexpansiveness of metric projection operators on Wasserstein spaces. *arXiv preprint arXiv:2009.01370*, 2020.

[2] David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–1879. PMLR, 2019.

[3] Luigi Ambrosio, Elia Brué, Daniele Semola, et al. *Lectures on optimal transport*, volume 130. Springer, 2021.

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[6] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[7] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 111–121, 2020.

[8] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. Tame functions are semismooth. *Mathematical Programming*, 117(1):5–19, 2009.

[9] Jérôme Bolte, Tam Le, and Edouard Pauwels. Subgradient sampling for nonsmooth nonconvex minimization. *SIAM Journal on Optimization*, 33(4):2542–2569, 2023.

[10] Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188:19–51, 2021.

[11] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[12] Benjamin Charlier, Jean Feydy, Joan Alexis Glaunes, François-David Collin, and Ghislain Durif. Kernel operations on the gpu, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[14] John M Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

[15] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

[16] Lucas De Lara, Alberto González-Sanz, and Jean-Michel Loubes. A consistent extension of discrete optimal transport maps for machine learning applications. *arXiv preprint arXiv:2102.08644*, 2021.

[17] Guido De Philippis, Alpár Richárd Mészáros, Filippo Santambrogio, and Bozhidar Velichkov. BV estimates in optimal transportation and applications. *Archive for Rational Mechanics and Analysis*, 219:829–860, 2016.

[18] Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

[19] Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 2. Springer, 1992.

[20] Theo Dumont, Théo Lacombe, and François-Xavier Vialard. On the Existence of Monge Maps for the Gromov-Wasserstein Distance. working paper or preprint, October 2022.

[21] Paul Embrechts and Marius Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 77:423–432, 2013.

[22] Y. M. Ermoliev and V. I. Norkin. Stochastic generalized gradient method with application to insurance risk management. *IIASA Interim Report, IR-97-021, Laxenburg, Austria*, April 1997.

[23] LawrenceCraig Evans. *Measure theory and fine properties of functions*. Routledge, 2018.

[24] Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

[25] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[26] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[27] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177:113–161, 1996.

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[29] Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166 – 1194, 2021.

[30] John L Kelley. *General topology*. Courier Dover Publications, 2017.

[31] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *Stat*, 1050:1, 2014.

[32] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

[33] Quentin Mérigot, Filippo Santambrogio, and Clément Sarrazin. Non-asymptotic convergence bounds for Wasserstein approximation using point clouds. *Advances in Neural Information Processing Systems*, 34:12810–12821, 2021.

[34] Quentin Merigot and Boris Thibert. Optimal transport: discretization and algorithms. working paper or preprint, February 2020.

[35] VS Mikhalevich, AM Gupal, and VI Norkin. Methods of nonconvex optimization. *arXiv preprint arXiv:2406.10406*, 2024.

[36] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations*, 6, 2018.

[37] François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as regularization: Smooth and strongly convex Brenier potentials in optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232. PMLR, 2020.

[38] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29, 2016.

[39] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 51(1):1–44, 2019.

[40] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[41] Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps, 2024.

[42] Aldo Pratelli. On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 43, pages 1–13. Elsevier, 2007.

[43] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[44] Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. *Mathematical Programming*, pages 1–82, 2024.

[45] Antoine Salmona, Valentin De Bortoli, Julie Delon, and Agnes Desolneux. Can push-forward generative models fit multimodal distributions? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10766–10779. Curran Associates, Inc., 2022.

[46] Antoine Salmona, Julie Delon, and Agnès Desolneux. Gromov-Wasserstein-like distances in the gaussian mixture models space. *arXiv preprint arXiv:2310.11256*, 2023.

[47] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[48] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2020.

[50] Eloi Tanguy. Convergence of SGD for training neural networks with sliced Wasserstein losses. *Transactions on Machine Learning Research*, 2023.

[51] Robert E Tarjan. Dynamic trees as search trees via euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177, 1997.

[52] Adrien B Taylor. *Convex interpolation and performance estimation of first-order methods for convex optimization.* PhD thesis, Catholic University of Louvain, Louvain-la-Neuve, Belgium, 2017.

[53] Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.

[54] Cédric Villani. *Optimal transport : old and new / Cédric Villani.* Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[55] Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

[56] Seiichiro Wakabayashi. Remarks on semi-algebraic functions, January 2008. Online Notes.

[57] Yinyu Ye and Edison Tse. An extension of karmarkar's projective algorithm for convex quadratic programming. *Mathematical programming*, 44:157–179, 1989.

A Appendix

A.1 Continuous-to-Discrete Case: Semi-discrete OT

In the alternate optimisation scheme proposed in Section 3, the step with g fixed can be seen as semi-discrete Optimal Transport, whenever the target measure ν is discrete, and when the measure $g\#\mu$ is absolutely continuous with respect to the Lebesgue measure. The condition $g\#\mu \ll \mathscr{L}$ arises naturally whenever the source measure μ is itself absolutely continuous, which we will assume for this section.

Specifically, the sub-problem of computing

$$\mathcal{T}_c(g\#\mu, \nu)$$

can be seen as a semi-discrete optimal transport problem between $g\#\mu$ and ν (see [34] for a course with a detailed section on semi-discrete OT). To apply semi-discrete optimal transport methods to this sub-problem, we need to verify $g\#\mu \ll \mathscr{L}$. First, it follows from the definition that if $g\#\mathscr{L} \ll \mathscr{L}$, then, since we assume μ is absolutely continuous, $g\#\mu \ll \mathscr{L}$ would follow. In Lemma A.1, we provide relatively general sufficient conditions on the map $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$.

Lemma A.1. *Let $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ locally Lipschitz such that for $\mathscr{L}$-a.e. $x \in \mathbb{R}^d$, $\det \partial g(x) \neq 0$. Then $g\#\mathscr{L} \ll \mathscr{L}$.*

Remark A.2. *By Rademacher's theorem ([23], Theorem 3.2), a locally Lipschitz function is differentiable $\mathscr{L}$-a.e..*

Proof. First, we remind that $J_g := x \longmapsto |\det \partial g(x)|$ (defined $\mathscr{L}$-almost-everywhere) is locally integrable since g is locally Lipschitz. We now prove that $g\#\mathscr{L} \ll \mathscr{L}$ by considering the intersection

of compact sets and $\mathscr{L}$-null sets. Let $\mathcal{K} \subset \mathbb{R}^d$ a compact set and $E \subset \mathbb{R}^d$ a Borel set such that $\mathscr{L}(E) = 0$. By the area formula ([23], Theorem 3.8), the following equality holds

$$\int_{g^{-1}(E)\cap\mathcal{K}} J_g(x)\mathrm{d}x = \int_{\mathbb{R}^d} \mathcal{H}^0(g^{-1}(\{y\}) \cap \mathcal{K} \cap g^{-1}(E))\mathrm{d}y = \int_E \mathcal{H}^0(g^{-1}(\{y\}) \cap \mathcal{K})\mathrm{d}y, \tag{42}$$

where $\mathcal{H}^0$ denotes the 0-dimensional Hausdorff measure (the counting measure). The left-side expression in Eq. (42) is finite. Since $\mathscr{L}(E) = 0$, it follows that the right-most term in Eq. (42) is 0, thus

$$\int_{g^{-1}(E)\cap\mathcal{K}} J_g(x)\mathrm{d}x = 0.$$

Since by assumption J_g is positive almost-everywhere, it follows that $\mathscr{L}(g^{-1}(E) \cap \mathcal{K}) = 0$. Since the compact set $\mathcal{K}$ was chosen arbitrarily, we conclude that $\mathscr{L}(g^{-1}(E)) = 0$, which shows $g\#\mathscr{L} \ll \mathscr{L}$. □

A.2 Lemmas on Pseudo-inverses and Quantile Functions

To begin with, we introduce some notions regarding pseudo-inverses of non-decreasing functions.

Definition A.3. *For $\psi : \mathbb{R} \longrightarrow \mathbb{R}$ non-decreasing, its **right-inverse** is defined as the function:*

$$\psi^{\leftarrow} : \mathbb{R} \longrightarrow \overline{\mathbb{R}} : \quad \forall p \in \mathbb{R},\ \psi^{\leftarrow}(p) := \inf\{x \in \mathbb{R} \mid \psi(x) \geq p\}.$$

*For $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ non-decreasing, its **left-inverse** is defined as the function:*

$$\phi^{\rightarrow} : \mathbb{R} \longrightarrow \overline{\mathbb{R}} : \quad \forall p \in \mathbb{R},\ \phi^{\rightarrow}(p) := \sup\{x \in \mathbb{R} \mid \phi(x) \leq p\}.$$

These notions are particularly useful for the definition of the right-inverse of the cumulative distribution function of a probability measure μ: $F_\mu := x \longmapsto \mu((-\infty, x])$, and for the left-inverse of the function $G_\mu := x \longmapsto \mu((-\infty, x))$. We recall and prove some well-known properties of pseudo-inverses (see [21] for a detailed presentation of right-inverses). For a non-decreasing function ψ, we define $\psi(-\infty) := \lim_{x\searrow-\infty} \psi(x) \in \mathbb{R} \cup \{-\infty\}$ and $\psi(+\infty) := \lim_{x\nearrow+\infty} \psi(x) \in \mathbb{R} \cup \{+\infty\}$.

Lemma A.4. *1. Let $\psi : \mathbb{R} \longrightarrow \mathbb{R}$ non-decreasing and right-continuous. Then:*

(a) For all $(x, p) \in \mathbb{R}^2$, $\psi(x) \geq p \iff x \geq \psi^{\leftarrow}(p)$.

(b) If $\psi^{\leftarrow}(p) < +\infty$, $\psi(\psi^{\leftarrow}(p)) \geq p$.

2. Let $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ non-decreasing and left-continuous. Then:

(a) For all $(x, p) \in \mathbb{R}^2$, $\phi(x) \leq p \iff x \leq \phi^{\rightarrow}(p)$.

(b) If $\phi^{\rightarrow}(p) > -\infty$, $\phi(\phi^{\rightarrow}(p)) \leq p$.

3. Under the assumptions above, if additionally $\phi \leq \psi$, then $\phi^{\rightarrow} \geq \psi^{\leftarrow}$.

Proof. We detail the proofs for claims 1.(a) and 1.(b), the arguments for 2.(a) and 2.(b) being essentially the same. First, we let $p \in \mathbb{R}$ such that $\psi^{\leftarrow}(p) < +\infty$, which is equivalent to supposing $A_p \neq \varnothing$, with $A_p := \{x \in \mathbb{R} \mid \psi(x) \geq p\}$. We also suppose $\psi^{\leftarrow}(p) > -\infty$, which is equivalent to assuming that A_p is lower-bounded. Since $A_p \neq \varnothing$, we can choose a decreasing sequence $(x_n) \in A_p^{\mathbb{N}}$ such that $x_n \xrightarrow[n\longrightarrow+\infty]{} \psi^{\leftarrow}(p)$. Since ψ is right-continuous and $\psi^{\leftarrow}(p) \in \mathbb{R}$, we have $\psi(x_n) \xrightarrow[n\longrightarrow+\infty]{} \psi(\psi^{\leftarrow}(p))$. However, since each $x_n \in A_p$, we have $\psi(x_n) \geq p$, and by taking the limit in the inequality we deduce $\psi(\psi^{\leftarrow}(p)) \geq p$. If $\psi^{\leftarrow}(p) = -\infty$, then the same argument with $x_n \xrightarrow[n\longrightarrow+\infty]{} -\infty$ and $\psi(-\infty) := \lim_{x\searrow-\infty} \psi(x) \in \mathbb{R} \cup \{-\infty\}$ also shows $\psi(\psi^{\leftarrow}(p)) \geq p$, which concludes the proof of 1.(b).

For 1.(a), we first assume $\psi^{\leftarrow}(p) < +\infty$. In this case, by 1b) we have $\phi(\psi^{\leftarrow}(p)) \geq p$, thus $[\psi^{\leftarrow}(p), +\infty) \subset A_p$. Yet by definition of $\psi^{\leftarrow}(p)$, $x \in A_p \implies x \geq \psi^{\leftarrow}(p)$, thus we conclude

$A_p = [\psi^{\leftarrow}(p), +\infty)$, which is exactly the same statement as $\psi(x) \geq p \iff x \geq \psi^{\leftarrow}(p)$. If $\psi^{\leftarrow}(p) = +\infty$, then the equivalence still holds, since $\psi(x) \geq p \iff x \in A_p$, with $A_p = \varnothing$.

Regarding 3., let $p \in \mathbb{R}$ such that $\phi^{\leftarrow}(p) > -\infty$. Then $\{x \in \mathbb{R} \mid \phi(x) \leq p\} = (-\infty, \phi^{\leftarrow}(p)]$ by 2.a), thus $\phi^{\leftarrow}(p) = \inf\{x \in \mathbb{R} \mid \phi(x) > p\}$. The previous equality also holds when $\phi^{\leftarrow}(p) = -\infty$. Now since $\phi \leq \psi$, we have $\{x \in \mathbb{R} \mid \phi(x) > p\} \subset \{x \in \mathbb{R} \mid \psi(x) \geq p\}$, and taking the infimum yields $\phi^{\rightarrow}(p) \geq \psi^{\leftarrow}(p)$. □

Using this result, we can now prove Lemma 3.4.

Proof of Lemma 3.4 . First, notice that as a cumulative distribution function, F_μ is non-decreasing and right-continuous. Since g is non-decreasing, we have for $p \in (0, 1)$:

$$F_{g\#\mu}(g \circ F_\mu^{\leftarrow}(p)) = \mathbb{P}_{X\sim\mu}\left(g(X) \leq g \circ F_\mu^{\leftarrow}(p)\right) \geq \mathbb{P}_{X\sim\mu}(X \leq F_\mu^{\leftarrow}(p)) = F_\mu \circ F_\mu^{\leftarrow}(p).$$

Now if $F_\mu^{\leftarrow}(p) < +\infty$, we have $F_\mu \circ F_\mu^{\leftarrow}(p) \geq p$ by Lemma A.4 1.b). We now turn to the case $F_\mu^{\leftarrow}(p) = +\infty$, which implies that $\forall x \in \mathbb{R}$, $F_\mu(x) < p$. Since F_μ is a cumulative distribution function, this implies $p \geq 1$, which we excluded. We have shown that $F_{g\#\mu}(g \circ F_\mu^{\leftarrow}(p)) \geq p$, thus by definition of $F_{g\#\mu}^{\leftarrow}(p)$, we have $F_{g\#\mu}^{\leftarrow}(p) \leq g \circ F_\mu^{\leftarrow}(p)$.

Regarding the converse inequality, we will show that the set $N := \left\{p \in (0,1) \ : \ F_{g\#\mu}^{\leftarrow}(p) < g \circ F_\mu^{\leftarrow}(p)\right\}$ is Lebesgue-null. Let $p \in N$ and $\alpha \in \left[F_{g\#\mu}^{\leftarrow}(p), g \circ F_{g\#\mu}^{\leftarrow}(p)\right)$. As done earlier with F_μ, using Lemma A.4 and the fact that $F_{g\#\mu}$ is a c.d.f. and that $p < 1$, we have $F_{g\#\mu} \circ F_{g\#\mu}^{\leftarrow}(p) \geq p$. Since $F_{g\#\mu}$ is non-decreasing, we obtain $p \leq F_{g\#\mu}(\alpha)$. We re-write $F_{g\#\mu}(\alpha)$ using its definition, then use the fact that g is non-decreasing:

$$p \leq F_{g\#\mu}(\alpha) = \mathbb{P}_{X\sim\mu}\left(g(X) \leq \alpha\right) \leq \mathbb{P}_{X\sim\mu}\left(g(X) < g \circ F_\mu^{\leftarrow}(p)\right) \leq \mathbb{P}_{X\sim\mu}\left(X < F_\mu^{\leftarrow}(p)\right) =: G_\mu(F_\mu^{\leftarrow}(p)). \tag{43}$$

We now want to show that $G_\mu(F_\mu^{\leftarrow}(p)) \leq p$. Since $G_\mu \leq F_\mu$ and since they are non-decreasing and G_μ is left-continuous, and F_μ is right-continuous (by the axiomatic properties of μ), by Lemma A.4 item 3, we have $G_\mu^{\rightarrow} \geq F_\mu^{\leftarrow}$. In particular, since G_μ is non-decreasing, we have

$$G_\mu(F_\mu^{\leftarrow}(p)) \leq G_\mu(G_\mu^{\rightarrow}(p)) \leq p,$$

where the final inequality comes from Lemma A.4 item 2b), with $\phi^{\rightarrow}(p) > -\infty$ since we chose $p > 0$.

We have shown that $G_\mu(F_\mu^{\leftarrow}(p)) \leq p$, thus every equality in Eq. (43) is an equality, and as a result, for any $\alpha \in \left[F_{g\#\mu}^{\leftarrow}(p), g \circ F_{g\#\mu}^{\leftarrow}(p)\right)$, we have $F_{g\#\mu}(\alpha) = p$, thus the right-inverse $F_{g\#\mu}^{\leftarrow}$ has a jump-discontinuity at p:

$$\sup_{q<p} F_{g\#\mu}^{\leftarrow}(q) = F_{g\#\mu}^{\leftarrow}(p) < \inf_{p<q} F_{g\#\mu}^{\leftarrow}(q).$$

We conclude that N is a subset of the set J of jump-discontinuities of $F_{g\#\mu}^{\leftarrow}$, and since $F_{g\#\mu}^{\leftarrow}$ is non-decreasing, J is countable and thus of Lebesgue measure 0. As a result, we have for almost-every $p \in [0, 1]$, $F_{g\#\mu}^{\leftarrow}(p) = g \circ F_\mu^{\leftarrow}(p)$. □

A.3 Reminder on Reduction in RKHS methods

The reduction method in RKHS is known since [6] (Section 3), but given the simplicity of the arguments and for the sake of self-completeness, we provide a proof and presentation in Lemma A.5.

Lemma A.5. *Consider a cost function $J : \mathcal{H} \longrightarrow \mathbb{R}_+$ which can be written as $J(h) = J\left(h(x_1), \cdots, h(x_n)\right)$, then if $h^* \in \mathcal{H}$ is a solution of*

$$\operatorname*{argmin}_{h\in\mathcal{H}} J(h) + \lambda\|h\|_{\mathcal{H}}^2,$$

then h_V, the orthogonal projection of h^* onto V (defined in Eq. (35)) verifies :

$$\forall i \in [\![1, n]\!],\ h_V(x_i) = h^*(x_i),$$

and as a result $J(h_W) = J(h^*)$, which leads to the following problem reduction:

$$\operatorname*{argmin}_{h \in \mathcal{H}}\ J(h) + \lambda \|h\|_{\mathcal{H}}^2 = \operatorname*{argmin}_{h \in V}\ J(h) + \lambda \|h\|_{\mathcal{H}}^2. \tag{44}$$

Proof. To show that $\forall i \in [\![1, n]\!],\ h_V(x_i) = h^*(x_i)$, we will show that

$$V^\perp = H_0 := \{h \in \mathcal{H} \mid \forall i \in [\![1, n]\!],\ h(x_i) = 0\}.$$

Indeed,

$$\begin{aligned}
h \in H_0 &\iff \forall i \in [\![1, n]\!],\ g(x_i) = 0 \\
&\iff \forall i \in [\![1, n]\!],\ \forall u \in \mathbb{R}^d,\ g(x_i) \cdot u = 0 \\
&\iff \forall i \in [\![1, n]\!],\ \forall u \in \mathbb{R}^d,\ \delta_{x_i}^u g = 0 \\
&\iff \forall i \in [\![1, n]\!],\ \forall u \in \mathbb{R}^d,\ \langle g, K(\cdot, x_i)u \rangle_{\mathcal{H}} = 0 \\
&\iff f \in V^\perp,
\end{aligned}$$

where δ_x^u is the linear form $h \longmapsto h(x) \cdot u$, whose Riesz representation in $\mathcal{H}$ is $K(\cdot, x)u$ by the kernel reproducing property. We conclude the proof with the fact that as an orthogonal projection, $\|h_V\|_{\mathcal{H}}^2 \leq \|h^*\|_{\mathcal{H}}^2$, which shows that the cost of h_V is less than the cost of h^*. □

A.4 Colour Transfer: RGB Point Cloud Viewpoint

In Fig. 12, we provide a visualisation of the colour transfer from Fig. 11 in the RGB space.

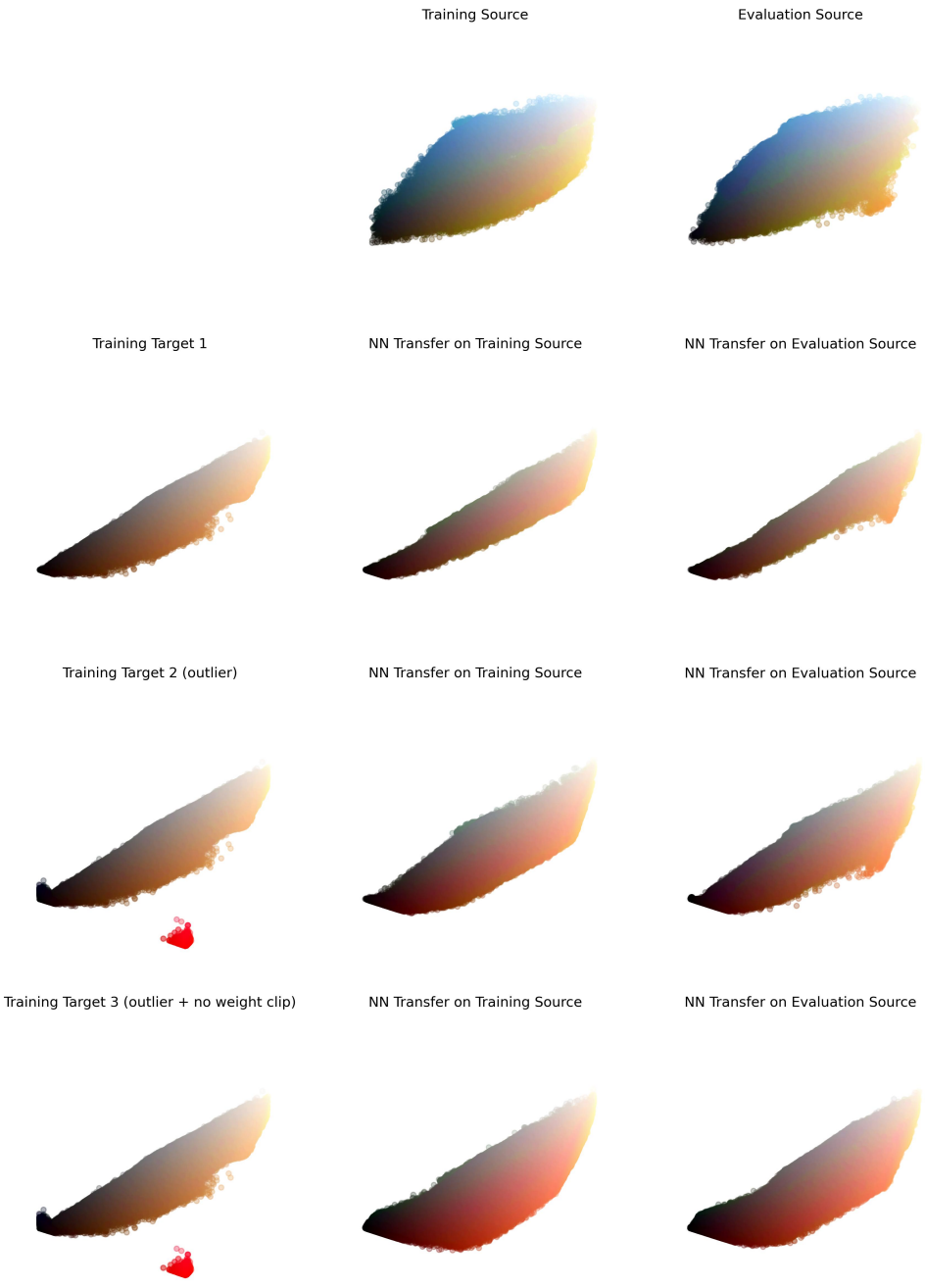


Figure 12: RGB space visualisation of the colour transfer from Fig. 11.