ENS Paris Saclay (MVA) & Mines Paris
Rapport de stage de fin d'études

Existence of Monge maps for the Gromov–Wasserstein distance

Théo Dumont

Laboratoire d'Informatique Gaspard Monge (ligm)

Supervisor 1: François-Xavier Vialard, ligm
Supervisor 2: Théo Lacombe, ligm
Referent at MVA: Gabriel Peyré, cnrs, dma, ens
Referent at Mines: Olivier Hermant, Mines Paris

April 15th 2022 – September 30th 2022

Contents

Introduction

Optimal transport

The Optimal Transport (OT) problem traces back to 1781, with Gaspard Monge and his *Mémoire sur la théorie des déblais et des remblais* [Mon81]. It can be described as the following: given two probability distributions μ and ν, how can we transfer all the mass of μ to ν while minimizing the overall *effort* to do so; the idea being originally to move dirt (*déblais*) from one place to another (*remblais*) in the most efficient way. Since then, it has matured a lot and has given birth to a prolific field of study. The interest of optimal transport, that, broadly speaking, uses *transport plans* to go from a probability measure to another, lies in both its ability to provide correspondences between sets of points and its ability to induce a geometric notion of distance between probability distributions, both having an impressive amount of applications and connections with pure mathematics.

Applications of optimal transport, to name a few, are vision (image registration and morphing [HT01], image retrieval [RTG98]), machine learning (domain adaptation [CFHR17], signal processing [KPT+17], unsupervised learning [GPC18], fairness [GDBFL19]), economics (equilibration of supply with demand [CMN10], structure of cities [CE07], social welfare [FKM11]), engineering (optimal shape / material design [BB01], reflector antenna design [GO03]), atmosphere and ocean dynamics (the semigeostrophic theory [Cul06]), biology [Xia07, SST+19], astrophysics [FMMS02] and statistics [Rac98]. OT also proved very useful in pure mathematics, with connections to inequalities [McC94, MV05, FMP10], geometry [Loe09, LV09], nonlinear partial differential equations [Bre87], and dynamical systems (weak KAM theory [BB07]; gradient flows [AGS05]).

Gromov–Wasserstein

However, optimal transport in its classical formulation is restricted to applications where there exists a direct way of comparing the samples of the data. It is therefore quite limited to the cases where the samples live in a same metric space, which prevents its use for a variety of machine learning tasks where the samples lie in different, seemingly not related, metric spaces or when a meaningful notion of distance between the samples can not be easily defined. The Gromov–Wasserstein (GW) problem [Mém11] solves the issue of comparing probability measures whose supports dwell in different, *incomparable* spaces by only considering comparisons *within* each space and not *between* them. Additionally, the GW problem mods isometries out, in the sense that it allows for the comparison of two shapes modulo isometries (rotations, translations), which proves very useful in contexts such as shape matching or shape analysis. This naturally finds many applications in machine learning: surfaces [BBK06] or graph matching [XLC19], regression problems in quantum chemistry [GSR+17], biology [DSS+20], or natural language processing [GJB19, AMJ18]. Extensions of the GW problem such as *unbalanced GW* [SVP21, Chap. 5] also demonstrate direct applications to biology [DSSS22].

Monge maps

A challenge in the field of optimal transport is to characterize the situations in which the optimal transport plan between two probability measures is a map. It is a theoretical challenge which can have fruitful consequences on the computational and algorithmic side: reduction of the optimization problem from plans to mappings and characterization of this mapping by optimality conditions, similarly to Brenier's theorem [Bre87] in optimal transport which gives the existence and uniqueness of the solution to the OT problem, this map being the gradient of a convex function.

Conditions on the cost function that guarantee the existence (and sometimes uniqueness) of Monge maps exist in the literature for the classical optimal transport problem. However, this conditions are very specific to the *linearity* of the OT problem, and there is no straightforward extension of them to the

GW problem, which is *quadratic*. Proving the existence of Monge maps for the Gromov–Wasserstein problem is therefore *a priori* much more complicated than for the linear OT problem and only little literature exist [Stu12, Vay20].

Contributions and outline

We start by defining some notions of optimal transport theory in Chapter 1: we recall some basic notions, the Gromov–Wasserstein problem and some important results on the existence and uniqueness of Monge maps for the linear OT problem. In Chapter 2, we detail our contributions to the field, starting with a summary of the current state-of-the-art and stating the main theorems of this work. In the appendix can be found some additional material, such as definitions (Appendix B) and the proofs of some of our claims that we did not put in the main corpus for the sake of clarity (Appendix A).

For $\mu, \nu \in \mathcal{P}(\mathbb{R}^n) \times \mathcal{P}(\mathbb{R}^d)$ of compact support with $\mu \ll \mathcal{L}^n$, we consider the GW problem with two different cost functions: the inner product cost (GW-IP) and the quadratic cost (GW-Q). The main contributions of this work are the two following theorems:

(i) (GW-IP) admits a map as a solution. (Th. 2.11)
(ii) (GW-Q) either admits a map, a bimap or a map/anti-map as a solution. (Th. 2.14)

We also provide some insights on the tightness of (ii):

(iii) There exists μ and ν for which no map is an optimal solution of (GW-Q). (Conj. 2.16)

On a different note, we study the optimality of the monotone non-decreasing or non-increasing plans $\pi^{\oplus}_{\text{mon}}$ and $\pi^{\ominus}_{\text{mon}}$ for (GW-Q):

(iv) There exists μ and ν for which neither $\pi^{\oplus}_{\text{mon}}$ nor $\pi^{\ominus}_{\text{mon}}$ is optimal for (GW-Q); (Alg. 1)
and having $\pi^{\oplus}_{\text{mon}}$ or $\pi^{\ominus}_{\text{mon}}$ as optimal is not stable by perturbations of μ and ν, even in the symmetric case. (Prop. 2.17)
(v) If μ and ν are composed of two distant parts, $\pi^{\oplus}_{\text{mon}}$ or $\pi^{\ominus}_{\text{mon}}$ is optimal for (GW-Q). (Prop. 2.18)

This last claim has been proven by my supervisors during the internship.

Details on these claims—motivations and what is at stake—can be found in Section 2.1.2. A preprint with our results can be found online[1], and the code for our experiments is available on GitHub[2].

[1] link: https://arxiv.org/abs/2210.11945.
[2] link: https://github.com/theodumont/monge-gromov-wasserstein.

List of symbols

Linear algebra

$\nabla f, \tilde{\nabla} f, \nabla^2 f$	gradient, approximate gradient (see Definition B.13), Hessian of the function f
Df	differential of the function f
$\|\cdot\|_p^p$ or $\|\cdot\|^p$	p-norm on $\mathbb{R}^d$, often written $\|\cdot\|^p$ for clarity
$\langle\cdot,\cdot\rangle_F, \|\cdot\|_F$	Frobenius inner product $\mathrm{tr}(A^\top B)$ and associated norm
$\langle\cdot,\cdot\rangle$	depending on the context, standard scalar product on $\mathbb{R}^d$ (also written $\cdot$), Frobenius inner product $\mathrm{tr}(A^\top B)$ or duality bracket $\langle x,\, x^*\rangle$

Measure theory

$\mathcal{P}(\mathcal{X})$	set of probability measures on $\mathcal{X}$
$\mathcal{P}_p(\mathcal{X})$	set of probability measures on $\mathcal{X}$ with finite p-moment
$\hat{\mathcal{P}}_n(\mathcal{X})$	set of empirical probability measures on $\mathcal{X}$ with n points: $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$
$\mathcal{B}(\mathcal{X})$	Borel sets of $\mathcal{X}$
μ, ν	(often) measures on $\mathcal{X}$ and $\mathcal{Y}$
π, γ	(often) measures on the product space $\mathcal{X} \times \mathcal{Y}$ (couplings)
$\mathcal{L}^d$	Lebesgue measure on $\mathbb{R}^d$
vol_M	volume measure on M
$\ll$	absolute continuity (see Definition B.3)
ρ_μ	density function of μ
$\mathrm{supp}(\mu)$	support of the measure μ (see Definition B.2)
$\mu \llcorner A$	restriction of μ to the set A (see Definition B.5)
T_*	pushforward operator induced by the function T
$\mu \otimes \nu$	product measure of the measures μ and ν: $(\mu\otimes\nu)(A\times B) = \mu(A)\nu(B)$
$(\mathrm{id}, T)_*\mu$	pushforward of μ by $x \mapsto (x, T(x))$; equivalently, the measure $\mu(\mathrm{d}x) \otimes \delta_{T(x)}(\mathrm{d}y)$
$A\mu$	for A a matrix and μ a measure, pushforward of μ by A

Convex analysis

φ^*, φ^c	Legendre transform and c-transform of φ (see Appendix B.2)
∂, ∂_c	subdifferential and c-subdifferential operators (see Appendix B.2)

Optimal transport

$\Pi(\mu,\nu)$	set of couplings of the measures μ and ν (see Definition 1.2)
Π_n	set of bistochastic matrices of size $n \times n$
$\Pi^\star(\mu,\nu)$	set of optimal couplings of the measures μ and ν (see Definition 1.2)
W_p	p-Wasserstein distance between measures (see Definition 1.4)
GW_p	p-Gromov–Wasserstein distance between mm-spaces (see Definition 1.5)
$\pi^\oplus_{\mathrm{mon}}, T^\oplus_{\mathrm{mon}}$	monotone non-decreasing transport plan, map (see Proposition 1.8)
$\pi^\ominus_{\mathrm{mon}}, T^\ominus_{\mathrm{mon}}$	monotone non-increasing transport plan, map (see Proposition 1.8)

Differential geometry

M	(often) a manifold
$\exp_p$	Riemannian exponential map at $p \in M$ (see Definition B.10)

Sets

$\bar{\mathbb{N}}, \bar{\mathbb{R}}$	$\mathbb{N} \cup \{+\infty\}, \mathbb{R} \cup \{+\infty\}$
C^k	k-times continuously differentiable functions
L^p	functions f such that $\lvert f \rvert^p$ is Lebesgue integrable
$\mathfrak{S}(X)$, $\mathfrak{S}_n$	set of permutations of a set X, of $[\![n]\!]$
P_σ	$n \times n$ matrix associated to a permutation $\sigma \in \mathfrak{S}_n$
$\lvert S \rvert$ or card S	cardinality of a set S
$\mathbb{1}_S$	indicator function of a set S
H	Hausdorff distance between sets (see Sec. 1.3.1)
GH	Gromov–Hausdorff distance between metric spaces (see Sec. 1.3.1)
$\mathrm{diam}_p(\mathbb{X})$	p-diameter of a metric measure space $\mathbb{X}$ (see Sec. 1.3.1)

Acronyms

OT	optimal transport
W, GW	Wasserstein, Gromov–Wasserstein
MP, KP	Monge problem, Kantorovich problem
AP, QAP	assignment problem, quadratic assignment problem
a.e., w.r.t.	almost everywhere, with respect to
mm-space	metric measure space

Chapter 1

Notions of optimal transport theory

Contents

Optimal transport (OT) is a long-standing mathematical problem that first arose in the late XVIII$^{\text{th}}$ century [Mon81] and has matured a lot since then. A good introduction to this theory can be found in [San15], while the more mathematical and especially geometric aspects can be found in [Vil08]; the most complete document about the computational aspects of OT is [PC19]. In this chapter, we put an emphasis on the tools we need for our study. After a quick introduction to the OT problems (Sec. 1.1), we will see that in a certain number of cases, the optimal solution of the problem has a deterministic structure (Sec. 1.2). We will then define the Gromov–Wasserstein transportation problem (Sec. 1.3) for which we refer the reader to [Mém11] for its first introduction and to [Stu12] for an in-depth analysis of its mathematical properties.

1.1 Basics of optimal transport

1.1.1 Monge and Kantorovich problems

The OT problem traces back to 1781, with Gaspard Monge and his *Mémoire sur la théorie des déblais et des remblais* [Mon81]. It can be described as the following: given two probability distributions μ and ν, how can we transfer all the mass of μ to ν while minimizing the overall *effort* to do so; the idea being originally to move dirt (*déblais*) from one place to another (*remblais*) in the most efficient way. Let us define these notions rigorously. Given two Polish spaces[1] $\mathcal{X}$ and $\mathcal{Y}$, a cost is a function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ which takes two points $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and outputs a value $c(x, y)$ evaluating how far is x from y, quantifying the effort of moving x to y. Let now $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ two measures. Given a Borel measurable function $T : \mathcal{X} \to \mathcal{Y}$, we define the *pushforward* of μ by T, written $T_*\mu$, by

$$\text{for all Borel set } A, \quad T_*\mu(A) \triangleq \mu(T^{-1}(A)) = \mu(\{x \in \mathcal{X} \mid T(x) \in A\}), \tag{1.1}$$

or equivalently by

$$\text{for all measurable function } f, \quad \int_{\mathcal{Y}} f(y)\, \mathrm{d}(T_*\mu)(y) = \int_{\mathcal{X}} f(T(x))\, \mathrm{d}\mu(x). \tag{1.2}$$

[1] a Polish space is a separable completely metrizable topological space.

This simply means that we are "pushing" the probability measure μ using T to obtain a probability measure on $\mathcal{Y}$. From a random variable perspective, this means that if $X \sim \mu$, then $T(X) \sim T_*\mu$. In the discrete case, *i.e.* when $\mu = \sum_{i=1}^N a_i \delta_{x_i}$, we simply get $T_*\mu = \sum_{i=1}^N a_i \delta_{T(x_i)}$. Back to our *déblais* and *remblais* problem, we see now that sending μ onto ν can be restated as finding a map T such that $T_*\mu = \nu$, and among all these maps, finding one that minimizes the total cost of sending every x onto $T(x)$, namely $\int_{\mathcal{X}} c(x, T(x))\, \mathrm{d}\mu(x)$. Hence the following formulation of the Monge problem (MP):

Definition 1.1 (Monge problem)**.** *Given two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$, we consider the problem*

$$\min_{T_*\mu = \nu} \int_{\mathcal{X}} c(x, T(x))\, \mathrm{d}\mu(x)\,. \tag{MP}$$

A map T satisfying $T_*\mu = \nu$ is called a transport map*. If it realizes* (MP)*, it is called an* optimal transport map*, or a* Monge map*.*

Monge analyzed some questions on the geometric properties of the solution to this problem but did not really solve it, as the question of the existence of a minimizer was not even addressed. In the following 150 years, the optimal transport problem mainly remained French and little progress was made. Indeed, this formulation raises a certain number of problems. First, transport maps may fail to exist; see for instance Fig. 1.1, where μ is composed of one dirac at x and ν of two diracs at y_1 and y_2. Sending μ to ν requires to "split" δ_x in half, which cannot be done by a map.

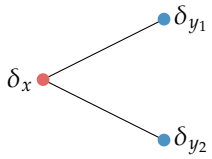


Figure 1.1: Optimal transport between δ_x and $\frac{1}{2}(\delta_{y_1} + \delta_{y_2})$.

Another issue arises when one tries to make the condition $T_*\mu = \nu$ more explicit: considering μ and ν with densities f and g w.r.t. the Lebesgue measure, condition (1.2) becomes (if f, g and T are nice enough) the partial differential equation

$$g(T(x))\, |\det(DT(x))| = f(x)\,, \tag{1.3}$$

which is highly non-linear in T, a major issue preventing from an easy analysis of the Monge problem.

In order to get rid of these difficulties, Kantorovich relaxed the Monge problem [Kan42] using the notion of transport plans, that we define now.

Definition 1.2 (Transport plan)**.** *Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$. A* transport plan *(or* coupling*) between μ and ν is a (probability) measure $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ of marginals μ and ν, i.e. noting P^1 and P^2 the projections on $\mathcal{X}$ and $\mathcal{Y}$ respectively, the set of transport plans $\Pi(\mu, \nu)$ is*

$$\Pi(\mu, \nu) \triangleq \left\{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid P^1_*\pi = \mu, P^2_*\pi = \nu\right\}\,, \tag{1.4}$$

or equivalently $\Pi(\mu, \nu) = \left\{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \forall A \subset \mathcal{X}, \forall B \subset \mathcal{Y}, \pi(A \times \mathcal{Y}) = \mu(A), \pi(\mathcal{X} \times B) = \nu(B)\right\}$. See Fig. 1.2 for an illustration.

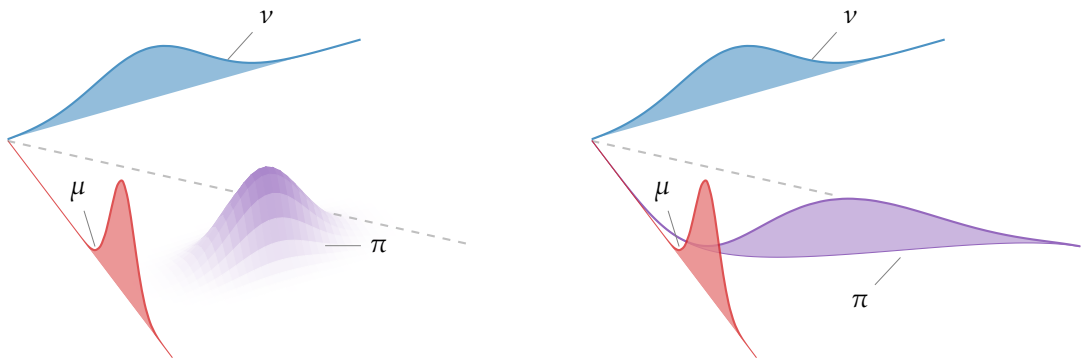


Figure 1.2: Transport plans are probability measures on $\mathcal{X} \times \mathcal{Y}$. **(Left)** π is the product measure $\mu \otimes \nu$. **(Right)** π is induced by a map.

This notation introduced, we are ready to write the Kantorovich formulation of the optimal transport problem:

Definition 1.3 (Kantorovich problem)**.** *Given two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$, we consider the problem*

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

A transport plan π realizing (KP) *is called an* optimal transport plan*, or* optimal coupling*. We denote by $\Pi^\star(\mu,\nu)$ the set of all optimal couplings between μ and ν.*

Remark 1.1. The quantity $\pi(x,y)$ can be understood as the proportion of mass of μ originally located in x which is transported to y. The marginal constraint states that for all y, the total mass of μ transported to y must equal $\nu(y)$, and conversely that the total mass coming from a position x must equal $\mu(x)$. We already see the benefits of this relaxation: the new problem is linear in π, $\Pi(\mu,\nu)$ is always non-empty, since it always contains the product measure $\mu \otimes \nu$, and by arguments of compactness there always exists a minimizer of (KP).

This formulation is indeed a relaxation: from a transport map T, one can construct a transport plan by simply considering $(\mathrm{id}, T)_*\mu$, which has the right marginals. A crucial interrogation is whether this useful relaxation is *tight* or not: is the optimal of (KP) optimal for (MP) (*i.e.* is it a map)? Under some assumptions on the cost c (*e.g.* if $c(x,y) = |x-y|^p$ in $\mathbb{R}^d$) and on the measure μ (*e.g.* if $\mu \ll \mathcal{L}^d$), it is indeed the case, as illustrated by Brenier's theorem (Theorem 1.3). Section 1.2 gives more details on these specific situations that are of particular interest to us.

1.1.2 Wasserstein distances

An interesting application of optimal transport with the cost $c(x,y) = |x-y|^p$ in $\mathbb{R}^d$ is that it allows to define a distance over the space of probability measures:

Definition 1.4 (Wasserstein distance)**.** *Let $\Omega \subset \mathbb{R}^d$ and $p \geq 1$. Let $\mu, \nu \in \mathcal{P}(\Omega)$ with finite p-moment, that is $\int_\Omega |x|^p \,\mathrm{d}\mu + \int_\Omega |y|^p \,\mathrm{d}\nu < +\infty$. The* p-Wasserstein distance *between μ and ν is defined by*

$$\mathrm{W}_p(\mu,\nu) = \left(\min_{\pi \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} |x-y|^p \,\mathrm{d}\pi(x,y) \right)^{1/p}. \tag{1.5}$$

Proposition 1.1. W_p defines a distance over $\mathcal{P}_p(\Omega)$, the set of probability measures with finite p-moment.

Proof. See [San15, Ch. 5.1]. □

The set $\mathcal{P}_p(\Omega)$ equipped with the distance W_p is called the *Wasserstein space*, noted $\mathbb{W}_p(\Omega)$. It satisfies a lot of very interesting and useful properties, such as the fact that W_p metrizes the weak* convergence, but we do not need them for our study and will therefore refer the reader to [San15] for more information on this matter.

Remark 1.2. All of the above stays valid in a Polish space Ω by replacing the usual p-norm $|\cdot|^p$ on $\mathbb{R}^d$ by the distance function $d(x,y)^p$.

Remark 1.3 (Why Wasserstein?)**.** One could wonder why we use the Wasserstein distance to evaluate the similarity between probability measures, as it requires solving an optimization problem and as we already have other means to do so such as the KL divergence or the standard L^2 difference, both straightforward to use. The thing is that they rely on measuring $\mu(x)-\nu(x)$, *i.e.* how the measures differ at any given point x without working at "ground level" like OT does. As an illustration, let us consider two dirac measures δ_x and δ_y. Their L^1 distance is 0 when $x = y$ and 2 otherwise, independently of the values of x and y; their p-Wasserstein distance however behaves like $|x-y|^p$. This will prove useful in many applications such as computer vision: two images can have very different values pixel-wise but still be very similar in terms of Wasserstein distance, which can be of major interest regarding our goals.

Remark 1.4 (∞-Wasserstein distance)**.** The natural limit of Definition 1.4 when $p \longrightarrow \infty$ is the following:

$$W_\infty(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \sup_{(x,y)\in\operatorname{supp}\pi} \|x-y\|_\infty \,. \tag{1.6}$$

In this work, we will only use this notion to draw a parallel with the Hausdorff distance in Sec. 1.3.1.

1.1.3 Discrete case

Given two sets $\{x_1,\dots,x_N\}$ and $\{y_1,\dots,y_M\}$ of $\mathbb{R}^d$ and two probability vectors[2] a and b, we consider the two discrete probability measures $\mu = \sum_{i=1}^N a_i\delta_{x_i}$ and $\nu = \sum_{j=1}^M b_j\delta_{y_j}$. The set of couplings between μ and ν is the *transport polytope*

$$U(a,b) \triangleq \left\{P \in \mathbb{R}_+^{N\times M} \mid P\mathbb{1}_M = a, P^\top \mathbb{1}_N = b\right\} ,$$

where $P\mathbb{1}_M \triangleq (\sum_j P_{i,j})_i \in \mathbb{R}^N$ and $P^\top\mathbb{1}_N \triangleq (\sum_i P_{i,j})_j \in \mathbb{R}^m$, namely the set of probability matrices of "marginals" μ and ν, equivalent of the set of transport plans in the continuous case. Given a cost matrix $C = (c(x_i,y_j))_{i,j}$ (equivalent of the cost function c in the continuous case), the Kantorovich problem (KP) reads

$$\min_{P\in U(a,b)} \langle C,\, P\rangle \,, \tag{$\hat{\text{KP}}$}$$

where $\langle\cdot,\cdot\rangle$ is the Frobenius inner product between matrices $\langle A,\, B\rangle \triangleq \operatorname{tr}(A^\top B) = \sum_{i,j} A_{i,j}B_{i,j}$. In the case where $N = M$ and $a = b = \mathbb{1}_N/N$, the set of couplings $U(a,b)$ is the set of bistochastic matrices[3] Π_N (up to a factor n). A fundamental theorem of linear programming states that the minimum of a linear objective in a nonempty polyhedron, if finite, is reached at an extremal point of the polyhedron, and Birkhoff's theorem states that the extremal points of Π_N are the permutation matrices, hence the following result:

[2] that satisfies $a_i \geq 0$ for all $i \in [\![N]\!]$ and $\sum_{i=1}^N a_i = 1$.
[3] square matrices of nonnegative real numbers, each of whose rows and columns sum to 1.

Proposition 1.2 (Tight relaxation for discrete KP)**.** If $N = M$ and $a = b = \mathbb{1}_N/N$, there exists a solution $P_{\sigma^\star}$ of (K̂P) that is the permutation matrix associated to a permutation $\sigma^\star$. Hence in this case, the relaxation of (K̂P) to the following *assignment problem* (AP)

$$\min_{\sigma \in \mathfrak{S}_N} \langle C, P_\sigma \rangle = \frac{1}{n} \sum_{i=1}^{N} C_{i,\sigma(i)} \tag{AP}$$

is tight. (AP) is the equivalent of the Monge problem (MP) in the discrete case.

1.2 Solutions of OT: maps and monotonicity

1.2.1 Deterministic structure of optimal plans

In many cases, it is very convenient to dispose of an optimal transport problem for which the optimal transport plan is unique, and even more when it is induced by a transport map, as it can have fruitful consequences on the computational and algorithmic side: it reduces the optimization problem from plans to mappings and can also make explicit a characterization of this mapping by optimality conditions. Brenier's theorem, stated below, is the most well-known of such cases where the optimal plan is a map.

Theorem 1.3 (Brenier's theorem)**.** Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ such that the optimal cost between μ and ν is finite and $c(x, y) = |x - y|^2$. If $\mu \ll \mathcal{L}^d$, then there exists a unique[4] solution to (KP) and it is induced by a map T. This map is characterized by being the unique gradient of a convex function $T = \nabla f$ such that $(\nabla f)_* \mu = \nu$.

This is very useful in practice: if one finds a function f whose gradient sends μ to ν, then it is the only solution of (KP), as illustrated in the following example.

Example 1.1 (OT in the Gaussian case)**.** Let $\mu = \mathcal{N}(m_1, s_1^2)$ and $\nu = \mathcal{N}(m_2, s_2^2)$ be two Gaussians in $\mathbb{R}$. Then one can check that

$$T : x \mapsto \frac{s_2}{s_1}(x - m_1) + m_2$$

satisfies $T_* \mu = \nu$ and is the the derivative of the convex function $f(x) = \frac{s_2}{2s_1}(x - m_1)^2 + m_2 x$. By Brenier's theorem, T is therefore the unique optimal transport for the cost $c(x - y) = |x - y|^2$, and the associated Wasserstein distance is, after computation $W_2^2(\mu, \nu) = (m_1 - m_2)^2 + (s_1 - s_2)^2$. The same can be done when $\mu = \mathcal{N}(m_1, \Sigma_1)$ and $\nu = \mathcal{N}(m_2, \Sigma_2)$ are Gaussians of $\mathbb{R}^d$, where the optimal map is

$$T : x \mapsto A(x - m_1) + m_2 \,,$$

with $A = \Sigma_1^{-\frac{1}{2}} (\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}})^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}}$.

For the sake of completeness, we state here a version of Brenier's theorem in the context of complete Riemannian manifolds with the squared distance:

Proposition 1.4 ([Vil08, Thm. 10.41])**.** Let M be a Riemannian manifold, and $c(x, y) = d(x, y)^2$. Let $\mu, \nu \in \mathcal{P}(M)$ such that the optimal cost between μ and ν is finite. If $\mu \ll \mathrm{vol}_M$, then there is a unique solution of the Monge problem between μ and ν and it can be written as

$$y = T(x) = \exp_x(\tilde{\nabla} f(x)) \,,$$

where f is some $d^2/2$-convex function. The approximate gradient can be replaced by a true gradient

[4] by "unique", we mean "unique up to a set of μ-measure zero".

if any of the following conditions is satisfied:

(a) μ and ν are compactly supported;
(b) M has nonnegative sectional curvature[5];
(c) ν is compactly supported and M has asymptotically nonnegative curvature.

Brenier's theorem can be extended in a few directions. The condition that μ has a density can be weakened to the fact that it does not give mass to sets of Hausdorff dimension smaller than $d-1$ (*e.g.* hypersurfaces), and c can actually be a bit more general, as long as it satisfies the *twist condition*, that we define now together with its variants. In the following, let $\mathcal{X} = \mathcal{Y}$ be complete Riemannian manifolds and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous cost function. We refer to [MG11, CMN10, Vil08] for more information on the twist condition, to [AKM11, McC12] on the subtwist condition and to [Moa16] on the m-twist and generalized twist conditions.

Proposition 1.5 (Twist)**.** We say that c satisfies the *twist condition* if c is differentiable w.r.t. x and

$$\text{for all } x_0 \in \mathcal{X}, \quad y \mapsto \nabla_x c(x_0, y) \in T_{x_0}\mathcal{X} \text{ is injective.} \tag{Twist}$$

Suppose that c satisfies (Twist) and assume that any c-concave function is differentiable μ-a.e. on its domain. If μ and ν have finite transport cost, then (KP) admits a unique optimal transport plan $\pi^\star$ and it is induced by a map which is the gradient of a c-convex function $f : \mathcal{X} \to \mathbb{R}$:

$$\pi^\star = (\mathrm{id}, c\text{-}\exp_x(\nabla f))_* \mu \,.$$

Remark 1.5. Following [MG11, Vil08], we recall that the *c-exponential map* is defined on the image of $-\nabla_x c$ by the formula $c\text{-}\exp_x(p) = (\nabla_x c)^{-1}(x, -p)$, *i.e.* $c\text{-}\exp_x(p)$ is the unique y such that $\nabla_x c(x, y) + p = 0$. This notion particularizes into the usual Riemannian exponential map when $c = d^2/2$.

Proof. We refer to [Vil08, Chap. 10] as well as [San15, Sec. 1.3] for an intuition behind this property, although it relies on the dual formulation of optimal transport, which we do not discuss in this work. □

Remark 1.6. Costs of the form $c(x, y) = h(x - y)$ with h strictly convex, and in particular the costs $c(x, y) = |x - y|^p$ for $p > 1$, do satisfy the twist condition.

This formulation of the twist condition is equivalent to the fact that for all $y_1 \neq y_2 \in \mathcal{Y}$, the function $x \in \mathcal{X} \mapsto c(x, y_1) - c(x, y_2)$ has no critical point. Remark that on a compact manifold, if the cost is C^1, this condition can never be satisfied. Note also that the squared Riemannian distance is not C^1 everywhere and one can still prove the existence of Monge map; see Proposition 1.4. This justifies the introduction of two weaker notions, that turn out to remain sufficient to obtain some (but less) structure on the optimal plans:

Proposition 1.6 (Subtwist)**.** We say that c satisfies the *subtwist condition* if

$$\text{for all } y_1 \neq y_2 \in \mathcal{Y}, \quad x \in \mathcal{X} \mapsto c(x, y_1) - c(x, y_2) \quad \text{has at most 2 critical points.} \tag{Subtwist}$$

Suppose that c satisfies (Subtwist) and assume that any c-concave function is differentiable μ-a.e on its domain. If μ and ν have finite transport cost, then (KP) admits a unique optimal transport plan $\pi^\star$ and it is induced by the union of a map and an anti-map:

$$\pi^\star = (\mathrm{id}, G)_* \bar{\mu} + (H, \mathrm{id})_* (\nu - G_* \bar{\mu})$$

for some (Borel) measurable maps $G : \mathcal{X} \to \mathcal{Y}$ and $H : \mathcal{Y} \to \mathcal{X}$ and non-negative measure $\bar{\mu} \leq \mu$ such that $\nu - G_* \bar{\mu}$ vanishes on the range of G.

[5] see Appendix B.3.3.

Proof. The proof relies on the notion of *numbered limb systems* (NLS), which is formally the alternation of maps f_{2k} and anti-maps f_{2k+1} with conditions on their domains and ranges, introduced in [HW95] and applied to optimal transport by [AKM11, CMN10] where it is shown that in such a case, the support of the optimal transport plan is included in a numbered limb system with at most two limbs. See Definition B.15 for the definition of NLS and [AKM11] for the proof of Proposition 1.6. □

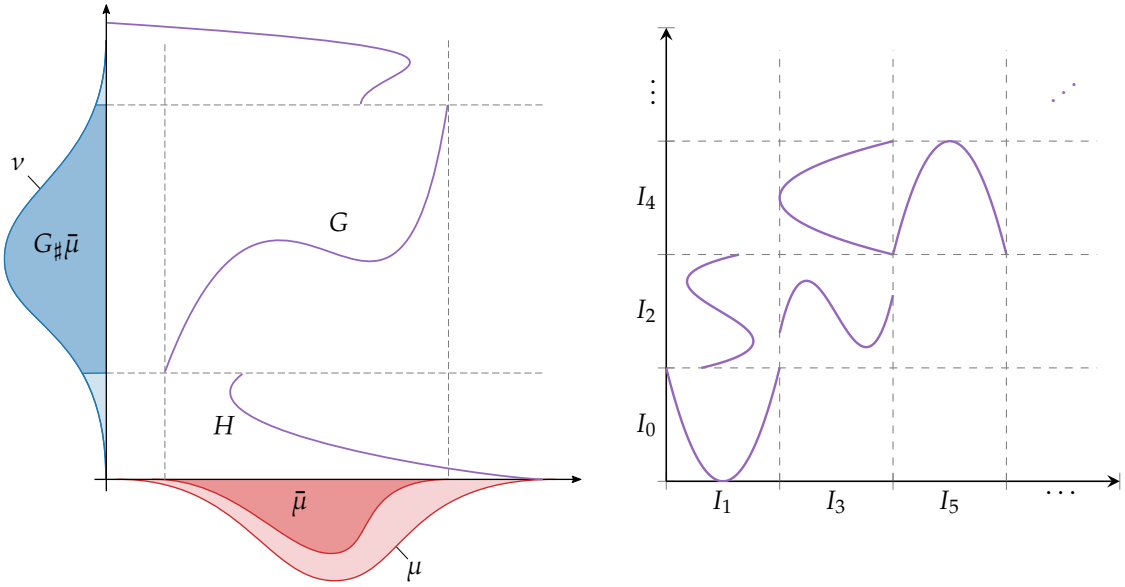


Figure 1.3: Optimal plans for a subtwisted cost. **(Left)** A map/anti-map structure (*i.e.* a numbered limb system with two limbs). **(Right)** Representation of a numbered limb system with N limbs. The subsets I_k are represented connected for visual convenience but do not need to be.

Proposition 1.7 (m-twist)**.** We say that c satisfies a *m-twist* (resp. *generalized twist*) *condition* if c is differentiable w.r.t. x and

$$\text{for all } x_0 \in \mathcal{X}, y_0 \in \mathcal{Y}, \text{ the set } \big\{y \mid \nabla_x c(x_0, y) = \nabla_x c(x_0, y_0)\big\} \text{ has at most } m \text{ elements} \qquad (m\text{-twist})$$

(resp. is a finite subset of $\mathcal{Y}$). Suppose that c is bounded, satisfies (m-twist) and assume that any c-concave function is differentiable μ-almost surely on its domain. If μ has not atom and μ and ν have finite transport cost, then each optimal plan $\pi^\star$ of (KP) is supported on the graphs of $k \in [\![m]\!]$ (resp. in $\mathbb{N} \cup \{\infty\}$) measurable maps, *i.e.* there exists non-negative functions $\alpha_i : \mathcal{X} \to [0,1]$ and (Borel) measurable maps $T_i : \mathcal{X} \to \mathcal{Y}$ such that

$$\pi^\star = \sum_{i=1}^{k} \alpha_i (\mathrm{id}, T_i)_* \mu \,,$$

in the sense $\pi^\star(S) = \sum_{i=1}^k \int_{\mathcal{X}} \alpha_i(x) \mathbb{1}_S(x, T_i(x)) \,\mathrm{d}\mu$ for any Borel $S \subset \mathcal{X} \times \mathcal{Y}$.

Remark 1.7. Notice that although the m-twist condition is a generalization of the twist condition (which is the 1-twist condition[6]), it is not a generalization of the subtwist condition. To make things

[6] y_0 is always in the set.

more clear, it can be useful to reformulate both twist and subtwist conditions in the following way:

$$\begin{aligned}
&\text{(Twist)}: && \text{for all } y_1 \neq y_2 \in \mathcal{Y}, && \operatorname{card}\left\{x \in \mathcal{X} \mid \nabla_x c(x, y_1) = \nabla_x c(x, y_2)\right\} = 0 \\
&\text{(Subtwist)}: && \text{for all } y_1 \neq y_2 \in \mathcal{Y}, && \operatorname{card}\left\{x \in \mathcal{X} \mid \nabla_x c(x, y_1) = \nabla_x c(x, y_2)\right\} \leq 2 \\
&\text{(m-twist)}: && \text{for all } x_0, y_0 \in \mathcal{X} \times \mathcal{Y}, && \operatorname{card}\left\{y \in \mathcal{Y} \mid \nabla_x c(x_0, y) = \nabla_x c(x_0, y_0)\right\} \leq m\,.
\end{aligned}$$

Remark 1.8. Following [Vil08, Rem. 10.33], when measures μ and ν have compact support and μ has a density—which are assumptions that we make in the following—, all conditions of Propositions 1.5 to 1.7 are satisfied.

Figure 1.6 illustrates the relationships between these notions as well as some others that concern the monotonicity of the optimal transport maps, that we tackle now.

1.2.2 Monotonicity of the optimal map

We will now see that in the one-dimensional real case and under some assumptions on the cost function c, there exists an optimal map, which is monotone non-decreasing. Given a probability distribution $\mu \in \mathcal{P}(\mathbb{R})$, we define its *cumulative distribution function* (CDF) $F_\mu : \mathbb{R} \cup \{+\infty\} \to [0,1]$ by

$$F_\mu(x) = \mu((-\infty, x])\,.$$

The CDF cannot always be inverted as it is not always strictly increasing, but we can define its pseudo-inverse $F_\mu^{[-1]} : [0,1] \to \mathbb{R} \cup \{+\infty\}$, which is the *quantile function* of μ, as

$$F_\mu^{[-1]}(r) = \inf\left\{x \in \mathbb{R} \mid F_\mu(x) \geq r\right\}\,.$$

This will prove very useful since we now have the following proposition:

Proposition 1.8 (Monotone plan, monotone map)**.** For any $\mu, \nu \in \mathcal{P}(\mathbb{R})$, we have that $(F_\mu^{[-1]})_*(\mathcal{L}^1 \llcorner [0,1]) = \mu$, which means that $\pi^\oplus_{\text{mon}} \triangleq (F_\mu^{[-1]}, F_\nu^{[-1]})_*(\mathcal{L}^1 \llcorner [0,1])$ is a transport plan between μ and ν, that we call the *monotone non-decreasing transport plan*. If μ is atomless, then the map $T^\oplus_{\text{mon}} \triangleq F_\nu^{-1} \circ F_\mu$ is well-defined, monotone, and sends μ to ν. We call it the *monotone non-decreasing map*.

Proof. See [San15, Sec. 2.1]. □

Intuitively, the monotone non-decreasing plan sends the "beginning" of μ on the "beginning" of ν, and so on, matching the quantiles with each other (see Fig. 1.5, right). We define similarly the *monotone non-increasing plan* (resp. map) $\pi^\ominus_{\text{mon}}$ (resp. $T^\ominus_{\text{mon}}$), that sends the "beginning" of μ on the "end" of ν Now, under some assumptions on the cost function c, these plans will be optimal, as stated in the following proposition [Car08, San15]:

Proposition 1.9 (Submodularity)**.** Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$. We say that a function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ is *(strictly) submodular* if

$$\text{for all } y_1 < y_2 \in \mathcal{Y}, \quad x \in \mathcal{X} \mapsto c(x, y_1) - c(x, y_2) \quad \text{is (strictly) increasing.}$$

If c is twice-differentiable, this writes

$$\text{for all } x, y \in \mathcal{X} \times \mathcal{Y}, \quad \partial_{xy} c(x, y) \leq 0\,. \tag{Submod}$$

Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ of finite transport cost. If c satisfies (Submod), then $\pi^\oplus_{\text{mon}}$ is an optimal plan for (KP), with uniqueness if the submodularity is strict. If μ is atomless, this optimal plan is induced by the optimal map $T^\oplus_{\text{mon}}$. Similarly, *supermodularity* is defined with the reversed inequality and induces the optimality of $\pi^\ominus_{\text{mon}}$ and $T^\ominus_{\text{mon}}$.

Remark 1.9. Let $x_1 \leq x_2$ and $y_1 \leq y_2$. Intuitively, the submodularity of c says that the "monotone non-decreasing" couple of affectations $(x_1 \leftrightarrow y_1), (x_2 \leftrightarrow y_2)$ has a better (smaller) transport cost than the "monotone non-increasing" one $(x_1 \leftrightarrow y_2), (x_2 \leftrightarrow y_1)$ (see Fig. 1.4). A more general definition of submodular functions can be found in Appendix B.4.2.

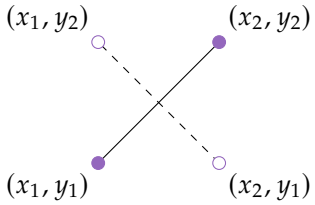


Figure 1.4: Submodularity and monotonicity of the optimal transport plan.

Remark 1.10 (OT in 1d is sorting)**.** This means that in some cases where the cost function is nice and the measures are discrete—*e.g.* when using the quadratic cost $|x-y|^2$ on empirical data, very common in various applications—, solving the OT problem resolves to sorting the data points in increasing order $x_1 \leq \cdots \leq x_N$ and $y_1 \leq \cdots \leq y_N$ and matching x_1 with y_1, x_2 with y_2, *etc* (see Fig. 1.5, left).

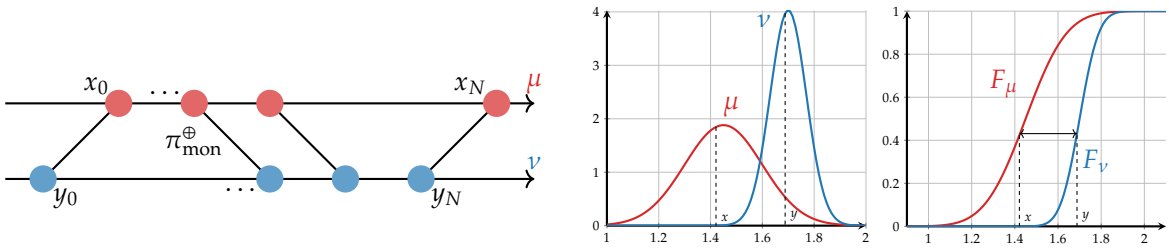


Figure 1.5: Optimal transport in 1D resolves to sorting. **(Left)** Optimal plan between two discrete probability measures with n uniform weights. **(Right)** In the general case, the optimal plan associates horizontally the points w.r.t. the cumulative distributions.

This property is *highly specific to the one-dimensional case*! One could indeed expect to obtain similar monotonicity results when assuming a similar property on c defined on $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, say for instance $\det \nabla^2_{xy} c(x,y) > 0$, but in such a case, we only have the following weaker proposition [MPW12]:

Proposition 1.10 (Non-degeneracy)**.** Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$. We say that a twice-differentiable function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ satisfies the *non-degeneracy condition* if

$$\text{for all } x, y \in \mathcal{X} \times \mathcal{Y}, \quad \det \nabla^2_{xy} c(x,y) \neq 0\,. \tag{Non-deg}$$

Suppose that c is C^2, satisfies (Non-deg), and that μ and ν are compactly supported. Then any solution of (KP) is supported by a d-dimensional Lipschitz submanifold (while its support is *a priori* $2d$-dimensional).

Remark 1.11. In such a case, we do not have the uniqueness of the solution of KP; see [MPW12].

All the properties above which relate to the cost function are linked; for the sake of clarity, we report these dependencies on a diagram that can be seen on Fig. 1.6.

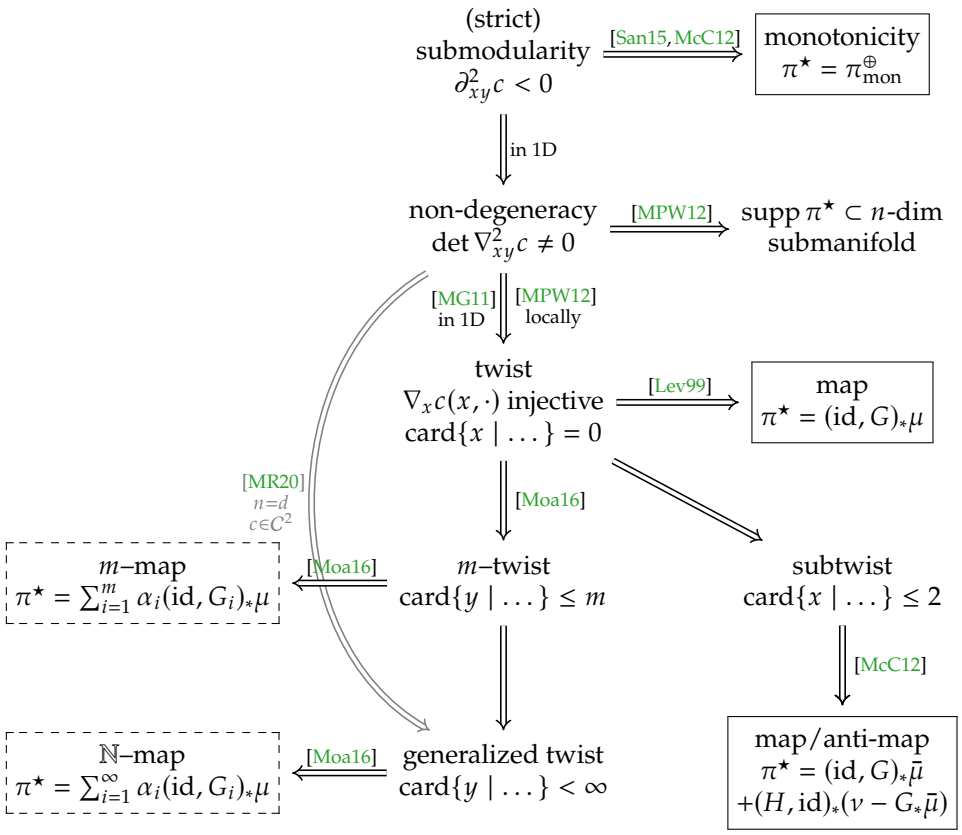


Figure 1.6: Links between the conditions for the deterministic structure of optimal plans and their uniqueness for a cost function c between measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$. Boxed plans are unique, dashed-boxed plans are unique up a condition. Each implication ($\Rightarrow$) has its own conditions on c and μ; see propositions above for reference.

1.3 Gromov–Wasserstein

This section aims at presenting the Gromov–Wasserstein (GW) problem, its basic properties and how it allows to get rid of some limitations of the classical OT problem. We refer the reader to [Mém11] where GW is introduced and to [Stu12] where its mathematical properties, in particular its geodesic structure and gradient flows, are analyzed.

1.3.1 The GW problem

In numerous applications, modeling data as a metric measure (mm) space[7] is very natural: the data is then represented by a space on which there is a metric and a probability measure. For instance, a cloud of points in the Euclidean space can be represented as a mm-space with the empirical measure and the standard distance between points in $\mathbb{R}^d$. When comparing two mm-spaces (*e.g.* in the context of shape analysis and shape matching), we often want to do it while modding translations and rotations out, since they do not change the intrinsic geometry and structure of the data. The idea of GW consists in using the metric and the probability measure of both ambient spaces $\mathcal{X}$ and $\mathcal{Y}$ to define the correspondence between them modulo isometries: intuitively, one wants to rotate and translate

[7] see Definition B.4.

both spaces in order to find a correspondence that associates nearby points in $\mathcal{X}$ to nearby points in $\mathcal{Y}$. The Gromov–Wasserstein (GW) problem, initially introduced in [Mém11], can be seen as an extension of the Gromov–Hausdorff distance [GKPS99], to the context of measure spaces $(\mathcal{X}, \mu)$ equipped with a cost function $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (typically, $c_{\mathcal{X}}$ can be a distance on $\mathcal{X}$); see also [Stu06] for a similar extension. Given $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ equipped with costs $c_{\mathcal{X}}$, $c_{\mathcal{Y}}$ respectively, and random variables $X, X' \sim \mu$ and $Y, Y' \sim \nu$, the GW problem seeks a correspondence (*i.e.* a joint law) between μ and ν that would make the distribution $c_{\mathcal{X}}(X, X')$ as close as possible to $c_{\mathcal{Y}}(Y, Y')$, in a L^p sense.

In the following, we build up some notions that will justify the theoretical introduction of the GW distance, by extending the parallel between comparing sets (Hausdorff dist.) and measures (Wasserstein dist.) to comparing metric spaces (Gromov–Hausdorff dist.) and metric measure spaces (Gromov–Wasserstein dist.), as done in [Mém11].

Hausdorff distance. Let A and B be two sets in a metric space $(\mathcal{Z}, d)$. The *Hausdorff distance* between A and B is defined by

$$\mathrm{H}_{\mathcal{Z}}(A, B) \triangleq \max\left\{\sup_{a\in A}\inf_{b\in B} d(a, b),\ \sup_{b\in B}\inf_{a\in A} d(a, b)\right\},$$

and $\mathrm{H}_{\mathcal{Z}}$ defines a distance between compact sets of $\mathcal{Z}$. Following [Mém11] and [PC19], we remark that we can define the Hausdorff distance between sets in a similar fashion to the Wasserstein distance between measures. Replacing the measures couplings (1.4) by *set couplings*, or *correspondences*,

$$\mathcal{R}(A, B) \triangleq \left\{R \subset A \times B \mid P^1(R) = A, P^2(R) = B\right\},$$

one has that

$$\mathrm{H}_{\mathcal{Z}}(A, B) = \inf_{R\in\mathcal{R}(A,B)} \sup_{(a,b)\in R} d(a, b), \tag{1.7}$$

Remark 1.12. This formulation is similar to the Kantorovich problem (KP) with a supremum instead of an integration, related to the ∞-Wasserstein distance (1.6).

Gromov–Hausdorff distance. The *Gromov–Hausdorff (GH) distance* [GKPS99] is a way of measuring how close two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are to being isometric to each other. It reads:

$$\mathrm{GH}(\mathcal{X}, \mathcal{Y}) \triangleq \inf_{\mathcal{Z}, f, g} \mathrm{H}_{\mathcal{Z}}(f(\mathcal{X}), g(\mathcal{Y})), \tag{GH}$$

where the infimum is taken on metric spaces $\mathcal{Z}$ and $f : \mathcal{X} \to \mathcal{Z}$ and $g : \mathcal{Y} \to \mathcal{Z}$ isometric embeddings (distance-preserving transformations) into $\mathcal{Z}$. One can show that GH defines a distance between compact metrics spaces up to isometries (bijective distance-preserving transformation). Similarly to (1.7), we can rewrite the GH distance using set couplings:

$$\mathrm{GH}(\mathcal{X}, \mathcal{Y}) = \frac{1}{2} \inf_{R\in\mathcal{R}(\mathcal{X},\mathcal{Y})} \sup_{(x,y),(x',y')\in R} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|. \tag{1.8}$$

Motivated by Remark 1.12, one could try to replace the maximization by an integration, which turns out to define the Gromov–Wasserstein distance, that we introduce now.

Gromov–Wasserstein distance. Given two mm-spaces $\mathbb{X} = (\mathcal{X}, d_{\mathcal{X}}, \mu_{\mathcal{X}})$ and $\mathbb{Y} = (\mathcal{Y}, d_{\mathcal{Y}}, \mu_{\mathcal{Y}})$, the *Gromov–Wasserstein (GW) distance* between $\mathbb{X}$ and $\mathbb{Y}$ is defined as the L^p-analogue of (1.8):

$$\mathrm{GW}_p(\mathbb{X}, \mathbb{Y}) = \inf_{\pi\in\Pi(\mu_{\mathcal{X}},\mu_{\mathcal{Y}})} \left(\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^p \,\mathrm{d}\pi(x, y)\,\mathrm{d}\pi(x', y')\right)^{1/p}.$$

In the Gromov–Wasserstein framework, there is no notion of transport, but rather a notion of *correspondence*. The goal is to find a correspondence, a matching plan between the two spaces $\mathbb{X}$ and $\mathbb{Y}$, which minimizes the *distortion* between the pairwise distances of couples of points. The most important property of the GW distance is the following:

Proposition 1.11. Let $\mathbb{X} = (\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X})$ and $\mathbb{Y} = (\mathcal{Y}, d_\mathcal{Y}, \mu_\mathcal{Y})$ be two mm-spaces. Then $\mathrm{GW}(\mathbb{X}, \mathbb{Y}) = 0$ if and only if $\mathbb{X}$ and $\mathbb{Y}$ are *strongly isomorphic*, *i.e.* if there exists an isometry $\varphi : \mathcal{X} \to \mathcal{Y}$ such that $\varphi_* \mu_\mathcal{X} = \mu_\mathcal{Y}$.

This property allows one to endow the space of metric measure spaces with a distance defined by GW, as long as the finiteness of GW is ensured; defining the *p-diameter* of a mm-space $\mathbb{X} = (\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X})$ as $\mathrm{diam}_p(\mathbb{X}) \triangleq \left(\iint d_\mathcal{X}(x, x')^p \, \mathrm{d}\mu_\mathcal{X} \, \mathrm{d}\mu_\mathcal{X}\right)^{1/p}$, we have:

Proposition 1.12 (GW is a distance)**.** GW_p is a distance on the space of finite p-diameter mm-spaces quotiented by the strong isomorphisms.

As a bonus, the GW problem has the advantage of allowing for the comparison of measures living in different spaces. Indeed, the classical OT problem (KP) presupposes that one has at their disposal a function c that takes an element x of $\mathcal{X}$ and an element y of $\mathcal{Y}$, and outputs a "distance" corresponding to the cost of sending x on y. When $\mathcal{X}$ and $\mathcal{Y}$ contain radically different objects, such as images, graphs, *etc*, it can be quite difficult to find a relevant function c of this sort. The Gromov–Wasserstein (GW) problem solves the issue of comparing probability measures whose supports dwell in different, *incomparable* spaces by only considering comparisons *within* each space and not *between* them. Similarly to the linear OT problem where $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ can be any function (not necessarily a distance), we consider the following problem:

Definition 1.5 (generalized GW)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $p \geq 1$. Given two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, two continuous functions $c_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the (generalized)* p-Gromov–Wasserstein problem *aims at finding*

$$\mathrm{GW}_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} |c_\mathcal{X}(x, x') - c_\mathcal{Y}(y, y')|^p \, \mathrm{d}\pi(x, y) \, \mathrm{d}\pi(x', y') \right)^{1/p}. \tag{GW}$$

A correspondence plan π realizing (GW) *is called an* optimal correspondence plan. *If it is induced by a map, we call this map a* Monge map, *with a slight overloading of the OT designation.*

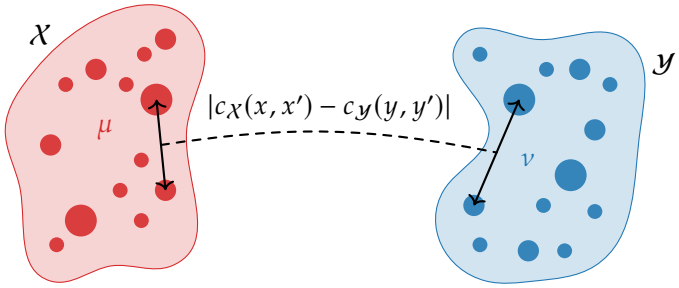


Figure 1.7: The Gromov–Wasserstein distance compares distances between couples of points within $\mathcal{X}$ and $\mathcal{Y}$ (continuous case).

From a random variable perspective, given $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ equipped with costs $c_\mathcal{X}, c_\mathcal{Y}$ respectively, and random variables $X, X' \sim \mu$ and $Y, Y' \sim \nu$, the GW problem seeks a correspondence (*i.e.* a joint law) between μ and ν that would make the distribution $c_\mathcal{X}(X, X')$ as close as possible to $c_\mathcal{Y}(Y, Y')$, in a L^p sense.

As for the linear OT problem, (GW) always admits a solution. This can be shown using the compactness of $\Pi(\mu,\nu)$ and the lower semi-continuity of $\pi \mapsto \iint L \,\mathrm{d}\pi\,\mathrm{d}\pi$ for the weak convergence, given by the l.s.c. of $L : (x,x',y,y') \mapsto |c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y')|^p$ itself [Vay20].

Remark 1.13 (GW with arbitrary cost functions)**.** When considering arbitrary continuous $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ (provided the finiteness or their p-moment w.r.t. $\mu_{\mathcal{X}} \otimes \mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}} \otimes \mu_{\mathcal{Y}}$ respectively), we are *a priori* losing the property that GW defines a distance between strong isomorphism classes of mm-spaces, and even its invariance by isometries. However, we still have the following properties, summarized in Theorem 2.2.1 of [Vay20] and whose proofs can be found in [Stu12, CM19]:

(i) GW_p is symmetric, positive and satisfies the triangle inequality;
(ii) for any $q \geq 1$ and $c_{\mathcal{X}} = d_{\mathcal{X}}^q$ and $c_{\mathcal{Y}} = d_{\mathcal{Y}}^q$, $\mathrm{GW}_p(\mu_{\mathcal{X}},\mu_{\mathcal{Y}}) = 0$ if and only if $(\mathcal{X}, c_{\mathcal{X}}, \mu_{\mathcal{X}})$ and $(\mathcal{Y}, c_{\mathcal{Y}}, \mu_{\mathcal{Y}})$ are "strongly isomorphic";
(iii) for arbitrary $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$, $\mathrm{GW}_p(\mu_{\mathcal{X}},\mu_{\mathcal{Y}}) = 0$ if and only if $(\mathcal{X}, c_{\mathcal{X}}, \mu_{\mathcal{X}})$ and $(\mathcal{Y}, c_{\mathcal{Y}}, \mu_{\mathcal{Y}})$ are *weakly isomorphic*, *i.e.* if there exists $(\mathcal{Z}, c_{\mathcal{Z}}, m)$ with $\mathrm{supp}(m) = \mathcal{Z}$ and "strong isomorphisms"[8] $\varphi : \mathcal{Z} \to \mathcal{X}$ and $\psi : \mathcal{Z} \to \mathcal{Y}$.

Hence for arbitrary $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$, we still have that GW_p defines a distance on the space

$$\mathbb{N}_p = \{(\mathcal{X}, c_{\mathcal{X}}, \mu_{\mathcal{X}}) \mid \mathrm{diam}_p(\mathcal{X}, c_{\mathcal{X}}, \mu_{\mathcal{X}}) < \infty\}\,,$$

where $\mathcal{X}$ is a Polish space, $\mu \in \mathcal{P}(\mathcal{X})$ and $c_{\mathcal{X}}$ a continuous function, quotiented by the weak isomorphisms.

1.3.2 Discrete case

We consider two discrete probability measures $\mu = \sum_{i=1}^N a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^M b_j \delta_{y_j}$. In the GW problem, we will not have at our disposal a similarity matric C like in Section 1.1.3, but rather two matrices $D^{\mathcal{X}}$ and $D^{\mathcal{Y}}$ that measure similarities *within* $\mathcal{X}$ and $\mathcal{Y}$, *i.e.* between elements of $\mathcal{X}$ and between elements of $\mathcal{Y}$. The Gromov–Wasserstein problem (GW) is then:

$$\min_{P \in U(a,b)} \sum_{i,j,i',j'} |D^{\mathcal{X}}_{i,i'} - D^{\mathcal{Y}}_{j,j'}|^p P_{i,j} P_{i',j'}\,. \tag{GŴ}$$

Similar to the linear OT case, one could ask whether the optimum of this problem is attained by a permutation σ in the case where $N = M$ and $a = b = \mathbb{1}_N/N$, *i.e.* if (GŴ) is a tight relaxation of:

$$\min_{\sigma \in \mathfrak{S}_N} \sum_{i,j} |D^{\mathcal{X}}_{i,\sigma(i)} - D^{\mathcal{Y}}_{j,\sigma(j)}|^p\,. \tag{QAP}$$

Theorem 4.1.2 from [Vay20] gives this result in the case where $p = 2$ and $D^{\mathcal{X}}$, $D^{\mathcal{Y}}$ are squared distance matrices but we can actually state a more general result, given in [ML18]:

Proposition 1.13 (Tight relaxation for discrete GW)**.** In the case $N = M$, $a = b = \mathbb{1}_N/N$ and $p = 2$, if we can write

$$D^{\mathcal{X}}_{i,i'} = h_1(x_i - x_{i'}) \quad \text{and} \quad D^{\mathcal{Y}}_{j,j'} = h_2(x_j - x_{j'})$$

where h_1 and h_2 are conditionally negative (or positive) definite functions (see Definition 1.6 below), then there exists a solution $P_{\sigma^\star}$ of (GŴ) that is the permutation matrix associated to a permutation $\sigma^\star$.

[8] with quotation marks since $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ do not give $\mathcal{X}$ and $\mathcal{Y}$ a metric space structure.

Definition 1.6 (Conditionally negative definite function)**.** *A function $h : \mathcal{X} \to \mathbb{R}$ is said to be* conditionally negative definite *if for all $N \geq 1$, $x_1, \dots, x_N \in \mathcal{X}$ and $\omega_1, \dots, \omega_N \in \mathbb{R}$ such that $\sum_{i=1}^N \omega_i = 0$, we have $\sum_{1 \leq i,j \leq N} \omega_i \omega_j h(x_i - x_j) \leq 0$.*

In particular, $x \mapsto |x|_2$ and $x \mapsto |x|_2^2$ are conditionally negative definite functions, which allows one to obtain the tightness of the relaxation of (GŴ) with (squared) Euclidean norm.

Remark 1.14 (Link with QAP)**.** Notice that we denoted by (QAP) the restriction of (GŴ) to the set of permutation matrices in the case where $N = M$ and $a = b = \mathbb{1}_N/N$. This is a slight abuse of notation, since in full generality this discrete GW formulation differs from the so-called *Quadratic Assignment Problem (QAP)* [KB57]; but when $p = 2$, a connection between the two can be drawn. The QAP reads:

Definition 1.7 (Quadratic Assignment Problem)**.** *Given a weight matrix W and a distance matric D of size $n \times n$, the* Quadratic Assignment Problem (QAP) *consists in finding a permutation $\sigma \in \mathfrak{S}_N$ (an* assignment*) that solves:*

$$\min_{\sigma \in \mathfrak{S}_N} \sum_{i,j=1}^N W_{i,j} D_{\sigma(i),\sigma(j)} \,. \tag{1.9}$$

In matrix notation, this reads

$$\min_{\sigma \in \mathfrak{S}_N} \operatorname{tr}(W P_\sigma D P_\sigma^\top) \,.$$

In the case where $N = M$ and $a = b = \mathbb{1}_N/N$, $U(a, b)$ is the set of bistochastic matrices Π_N (up to a factor n). It is then claimed in [AMJJ19] that (GŴ) with $p = 2$ is equivalent to the problem

$$\max_{P \in \Pi_N} \|APB^\top\|_F^2 \tag{1.10}$$

for some matrices A and B (and this even when $n \neq m$ and a and b are arbitrary). Then,

$$\|APB^\top\|_F^2 = \operatorname{tr}(BP^\top A^\top APB^\top) = \operatorname{tr}(A^\top APB^\top BP^\top) \,,$$

where $\|\cdot\|_F$ is the Frobenius norm. Hence problem (1.10) can be recast into the relaxation of QAP to Π_N with distance and weight matrices $W \triangleq -A^\top A$ and $D \triangleq B^\top B$, and it is therefore the same for (GŴ) with $p = 2$. The QAP is known as NP-hard to solve for arbitrary inputs. To the best of our knowledge, the NP-hardness of (GŴ) itself or of its restriction on $\mathfrak{S}_N$ has not been proved in the literature; but it is to expect. In the following, we will abuse notations and denote by QAP the (GŴ) problem restricted to $\mathfrak{S}_N$. Figure 1.8 illustrates the relationship between all the problems mentioned in this remark as well as the abuse of notation we will use in this report.

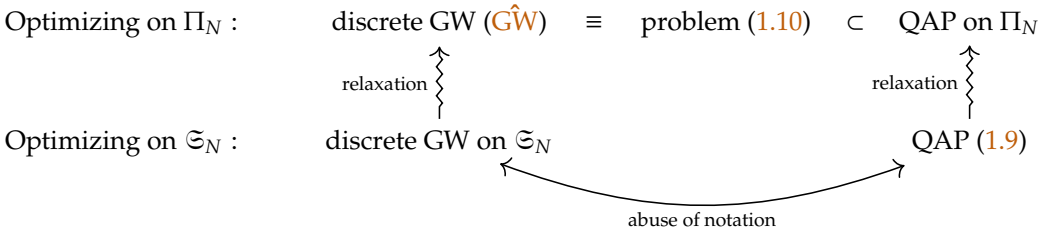


Figure 1.8: Relations between the GW problems and the QAP problem when optimizing either on $\mathfrak{S}_N$ or on Π_N.

1.3.3 Relation with the OT problem: a tight relaxation

The minimization problem in (GW) can be interpreted as the minimization of the map $\pi \mapsto F(\pi,\pi) \triangleq \iint k \,\mathrm{d}\pi\otimes\pi$ where $k((x,y),(x',y')) = |c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y')|^2$, and F is thus a symmetric bilinear map. By first order optimality condition, if $\pi^\star$ minimizes (GW), then it also minimizes $\pi \mapsto 2F(\pi,\pi^\star)$. If we let $C_{\pi^\star}(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} k((x,y),(x',y'))\,\mathrm{d}\pi^\star(x',y')$, we obtain the linear problem

$$\min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} C_{\pi^\star}(x,y)\,\mathrm{d}\pi(x,y)\,, \tag{1.11}$$

which is nothing but the (KP) problem induced by the cost $C_{\pi^\star}$ on $\mathcal{X}\times\mathcal{Y}$. Therefore, we obtain that any optimal *correspondence* plan for (GW) with costs $c_{\mathcal{X}}, c_{\mathcal{Y}}$ must be an optimal *transportation* plan for (KP) with cost $C_{\pi^\star}$. A crucial point, proved in [SVP21, Thm. 3] as a generalization of [Kon76], is that if k is symmetric negative on the set of (signed) measures on $\mathcal{X}\times\mathcal{Y}$ with null marginals, that is $\iint k\,\mathrm{d}\alpha\otimes\alpha \le 0$ for all such α, then the converse implication holds: any solution $\gamma^\star \in \Pi(\mu,\nu)$ of the OT problem with cost $C_{\pi^\star}$ is also a solution of the GW problem, that is $F(\pi^\star,\pi^\star) = F(\gamma^\star,\gamma^\star) = F(\pi^\star,\gamma^\star)$. Since the solutions of (GW) are in this case in correspondence with the solutions of an OT problem, the tools and knowledge from optimal transportation can be used to derive existence and structure of optimal maps.

In the following, we will be considering the GW problem in $\mathbb{R}^n$ and $\mathbb{R}^d$ in two different settings:

(i) the *inner product case*, where $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are the inner products on $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively (both denoted by $\langle\cdot,\cdot\rangle$):

$$\min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \left|\langle x,\,x'\rangle - \langle y,\,y'\rangle\right|^2 \,\mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y')\,; \tag{GW-IP}$$

(ii) the *quadratic case*, where $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are the squared Euclidean distance on $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively:

$$\min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \left||x-x'|^2 - |y-y'|^2\right|^2 \,\mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y')\,. \tag{GW-Q}$$

This linearization holds in particular for these two problems of interest: if α denotes a signed measure on $\mathcal{X}\times\mathcal{Y} \subset \mathbb{R}^n\times\mathbb{R}^d$ with null marginals, observe that

$$\begin{aligned}
&\int \left||x-x'|^2 - |y-y'|^2\right|^2 \,\mathrm{d}\alpha(x,y)\,\mathrm{d}\alpha(x',y') \\
&\quad= \underbrace{\int |x-x'|^2\,\mathrm{d}\alpha\otimes\alpha}_{=0} + \underbrace{\int |y-y'|^2\,\mathrm{d}\alpha\otimes\alpha}_{=0} - 2\int |x-x'|^2|y-y'|^2\,\mathrm{d}\alpha\otimes\alpha \\
&\quad= -2\int (|x|^2 - 2\langle x,x'\rangle + |x|^2)(|y|^2 - 2\langle y,y'\rangle + |y'|^2)\,\mathrm{d}\alpha\otimes\alpha\,.
\end{aligned}$$

Developing the remaining factor involve nine terms, but given that α has zero marginals (in particular, zero mass), we obtain that $\int |x|^2|y|^2\,\mathrm{d}\alpha\otimes\alpha = 0$ (and similarly for the terms involving $|x'|^2|y'|^2$, $|x|^2|y'|^2$ and $|x'|^2|y|^2$), and also that $\int |x|^2\langle y,y'\rangle\,\mathrm{d}\alpha\otimes\alpha = 0$ (and similarly for the other terms). Eventually, the only remaining term is

$$-8\int \langle x,x'\rangle\langle y,y'\rangle\,\mathrm{d}\alpha\otimes\alpha = -8\left\|\int x\otimes y\,\mathrm{d}\alpha(x,y)\right\|_F^2 \le 0\,,$$

where $x\otimes y \in \mathbb{R}^{n\times d}$ is the matrix $(x_i y_j)_{i,j}$, where $x = (x_1,\dots,x_n)$ and $y = (y_1,\dots,y_d)$, and $\|\cdot\|_F$ denotes the Froebenius norm of a matrix. The negativity of this term ensures that solutions of (GW-Q) are exactly the solutions of an OT problem. Computations for (GW-IP) are similar—actually, they immediately boil down to the same last two equalities.

Chapter 2

Optimal maps for Gromov–Wasserstein

Contents

A challenge in the field of optimal transport is to characterize the situations in which the optimal plan is a Monge map. It can have fruitful consequences on the computational and algorithmic side: reduction of the optimization problem from plans to mappings and characterization of this mapping by optimality conditions. However, conditions on the cost function that guarantee the existence of Monge maps (see Sec. 1.2) are very specific to the *linearity* of the OT problem, and there is no straightforward extension of them to the *quadratic* GW problem. Proving the existence of Monge maps for the Gromov–Wasserstein problem is therefore *a priori* much more complicated than for the linear OT problem. This challenge has been proposed by Sturm in [Stu12] and has been little studied since then. Results in the literature exist for two cost functions in Euclidean spaces, the squared distance cost (a standard choice for GW) and the inner product cost (that finds application on the sphere for instance). We study both and improve on the current state-of-the-art.

2.1 Optimal maps for Gromov–Wasserstein

2.1.1 State-of-the-art

Let $n \geq d$. We consider the GW problem in $\mathbb{R}^n$ and $\mathbb{R}^d$ in two different settings:

(i) the *inner product case*, where $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are the inner products on $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively (both denoted by $\langle \cdot, \cdot \rangle$):

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \left| \langle x, x' \rangle - \langle y, y' \rangle \right|^2 \mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y')\,, \tag{GW-IP}$$

which essentially compares distribution of angles in $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$;

(ii) the *quadratic case*, where $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are the squared Euclidean distance on $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively:

$$\min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \left||x-x'|^2 - |y-y'|^2\right|^2 \mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y')\,, \tag{GW-Q}$$

which is a standard choice for $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ and makes GW a distance between strong isometry classes of mm-spaces (see Remark 1.13).

In the inner product case, [Vay20, Thm. 4.2.3] gives a result on the existence of a Monge map under some assumptions:

Proposition 2.1 (Inner product cost: optimal map under condition)**.** Let $n \geq d$, $\mu, \nu \in \mathcal{P}(\mathbb{R}^n) \times \mathcal{P}(\mathbb{R}^d)$ two measures of finite second order moment with $\mu \ll \mathcal{L}^n$. Suppose that there exists $\pi^\star$ solution of (GW-IP) such that $M^\star = \int y \otimes x \,\mathrm{d}\pi^\star(x,y)$ is of full rank. Then there exists an optimal map between μ and ν that can be written as $T = \nabla f \circ M^\star$ with $f : \mathbb{R}^d \to \mathbb{R}$ convex.

For the quadratic case, there are only very few results. In [Vay20] is claimed that in the discrete case in dimension 1 with uniform mass and same number of points N, the optimal solution of (QAP) would either be the identity $\sigma(i) = i$ or the anti-identity $\sigma(i) = N + 1 - i$ (Thm. 4.1.1). However, a counter-example to this claim has been recently provided by [BHS22].
To the best of our knowledge, the only positive results on the existence of Monge maps for the quadratic cost are the following.

Proposition 2.2 ([Stu12, Thm. 9.21])**.** Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$. Assume that $\mu, \nu \ll \mathcal{L}^n$ and that both measures are rotationally invariant around their barycenter. Then every $\pi \in \Pi(\mu,\nu)$ which minimizes (GW-Q) is induced by a transport map T, unique up to composition with rotations. The transport map is constructed as follows: let s_μ be the radial distribution of μ around its barycenter z_μ, and let F_μ be the corresponding distribution function, *i.e.*

$$F_\mu(r) \triangleq s_\mu([0,r]) \triangleq \mu(\bar{B}_r(z_\mu))\,,$$

and similarly for ν. Then the monotone rearrangement $F_\nu \circ F_\mu^{-1} : \mathbb{R}_+ \to \mathbb{R}_+$ pushes forward μ to ν.

Proposition 2.3 ([Vay20, Prop. 4.2.4])**.** Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^n) \times \mathcal{P}(\mathbb{R}^d)$ with compact support, with $n \geq d$. Assume that $\mu \ll \mathcal{L}^n$ and that both μ and ν are centered. Suppose that there exists $\pi^\star$ solution of (GW-Q) such that $M^\star = \int y \otimes x \,\mathrm{d}\pi^\star(x,y)$ is of full rank. Then there exists $f : \mathbb{R}^d \to \mathbb{R}$ convex such that $T = \nabla f \circ M^\star$ pushes μ to ν. Moreover, if there exists a differentiable convex $F : \mathbb{R} \to \mathbb{R}$ such that $|T(x)|_2^2 = F'(|x|_2^2)$ μ-a.e., then T is optimal for (GW-Q).

2.1.2 Contributions

Contributions. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^n) \times \mathcal{P}(\mathbb{R}^d)$ two measures of compact support. Suppose $\mu \ll \mathcal{L}^n$. The main contributions of this work are the two following theorems:

(i) The (GW-IP) problem admits a map as a solution. (Th. 2.11)
(ii) The (GW-Q) problem either admits a map, a bimap or a map/anti-map as a solution. (Th. 2.14)

Both follow from a general theorem defined for costs that are invariant on the fibers a a certain function φ that we will state and prove first. We also ask if the second claim is tight, in the sense that there are cases where the optimal solution of (GW-Q) is not a map. Supported by numerical results, we believe that the following conjecture holds:

(iii) There exists measures μ and ν for which no map is an optimal solution of (GW-Q). (Conj. 2.16)

On a different note, Theorem 4.1.1 in [Vay20] states that in the discrete case, (GW-Q) is solved either by the monotone non-decreasing or monotone non-increasing plan, but [BHS22] provided a counter-example to this claim recently. We show numerically that:

(iv) There indeed exists measures μ and ν for which neither of the monotone non-decreasing or non-increasing plans is optimal for (GW-Q). (Alg. 1)
Additionally, having a monotone plan as optimal is not stable by small perturbations of μ and ν, even in the symmetric case. (Prop.2.17)

Yet, we state a new positive result for the existence of a Monge map for the quadratic cost:

(v) When measures μ and ν are composed of two distant parts, the monotone non-decreasing or non-increasing plan is optimal for (GW-Q). (Prop. 2.18)

This last claim has been proved by my supervisors during my internship. I still chose to include it in this report since it is the only positive result about the optimality of $\pi^{\oplus}_{\text{mon}}$ or $\pi^{\ominus}_{\text{mon}}$ other than the symmetric case given by [Stu12], and since it gives some insight on the fact that a monotone map is often optimal

Outline. The outline of this chapter is the following:

- Section 2.2 is focused on the statement and proof of our general theorem for the existence of Monge maps;
- In Section 2.3, we apply it to both GW problems:
  - to the inner product cost in Section 2.3.1;
  - to the quadratic cost in Section 2.3.2.
- Section 2.4 includes all additional results on the quadratic cost:
  - Section 2.4.1 details the non-optimality of the monotone plans and describes a procedure for finding bimaps and map/anti-maps, supporting our tightness conjecture;
  - Section 2.4.2 details our no-stability result;
  - Section 2.4.3 states our positive result on the existence of Monge maps in the context of two-components measures.

2.2 A general existence theorem

2.2.1 Statement of the results

An intuitive statement of the main theorem of this section is the following:

"Let $\mu, \nu \in \mathcal{P}(E)$. If one can send μ and ν in a space B by a function φ, such that $c(x, y) = \tilde{c}(\varphi(x), \varphi(y))$ for all $x, y \in E$ with $\tilde{c}$ a *twisted* cost on B, then *we can construct an optimal map between μ and ν*."

More precisely, let μ, ν be two probability measures supported on a measurable space (E, Σ_E) and consider a measurable map $\varphi : E \to B$, for some measurable space (B, Σ_B). We shall omit to mention the σ-algebra afterwards. We use the name *base space* for the space B. Let $(\mu_u)_{u\in B}$ (resp. $(\nu_u)_{u\in B}$) denote a disintegration of μ (resp. ν) with respect to φ (see Appendix B.1.2 for a definition). Consider a cost $c : E \times E \to \mathbb{R}$ that is invariant on the fibers of φ (that are simply the pre-images of points in the base B by φ), that is $c(x, y) = \tilde{c}(\varphi(x), \varphi(y))$ for all $x, y \in E$ and some cost function $\tilde{c}$ on $B \times B$. Solving the OT problem between μ and ν for c boils down to the OT problem between $\varphi_*\mu$ and $\varphi_*\nu$ on $B \times B$ for $\tilde{c}$. If we can ensure that there exists a Monge map t_B between $\varphi_*\mu$ and $\varphi_*\nu$ (for instance, if we can use Theorem 1.3), we may try to build a Monge map T between μ and ν by (i) transporting each fiber μ_u onto $\nu_{t_B(u)}$ *using a map* T_u, and (ii) gluing the $(T_u)_{u\in B}$ together to define a *measurable* map T satisfying $T_*\mu = \nu$ that will be optimal as it coincides with t_B on B and the cost c does not depend on the fibers

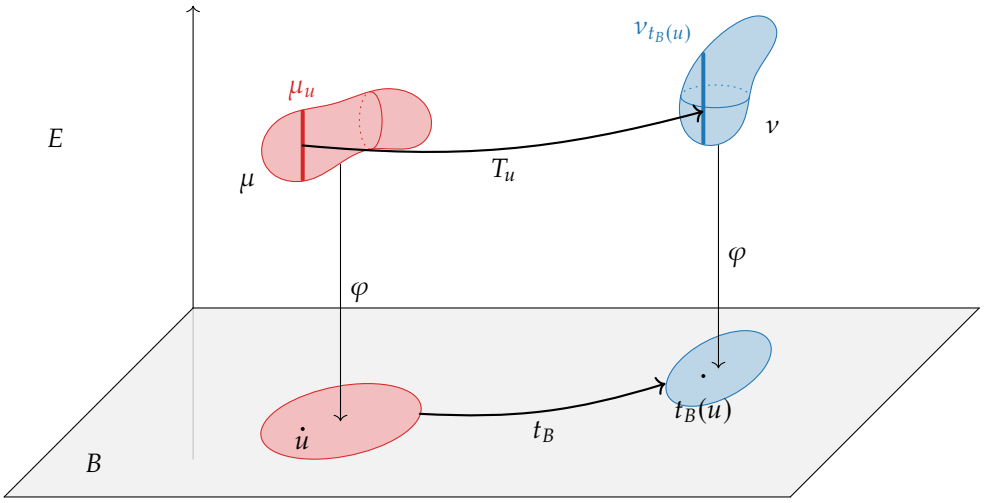


Figure 2.1: Illustration of the construction of a Monge map between μ and ν: we optimally transport the projections of the measures in B and then "lift" the resulting map t_B to E by sending each fiber μ_u onto the fiber $\nu_{t_B(u)}$, resulting respectively from the disintegrations of μ and ν by φ.

$(\varphi^{-1}(u))_{u\in B}$. We stress that ensuring the measurability of the map T is non-trivial and crucial from a theoretical standpoint.

We formalize this idea by the mean of two theorems: the first one guarantees in a fairly general setting the existence of a Monge map for the (GW) problem, but its construction is quite convoluted and there is little to no hope that it can be leveraged in practice, either from a theoretical or computational perspective. Assuming more structure, in particular on the fibers of φ, enables the construction of a Monge map for (GW) with a structure akin to Proposition 1.4. As detailed in Section 2.3, both (GW-Q) and (GW-IP) fall in the latter setting.

Theorem 2.4. Let $\mathcal{X}$ and $\mathcal{Y}$ be two measurable spaces for which there exists two measurable maps $\Phi_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}^d$ and $\Phi_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}^d$ that are injective, and whose inverses are measurable. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be two probability measures. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a cost function, and B_+, B_- be two measurable spaces along with measurable maps $\varphi : \mathcal{X} \to B_+$ and $\psi : \mathcal{Y} \to B_-$. Assume that there exists a cost $\tilde{c} : B_+ \times B_- \to \mathbb{R}$ such that

$$c(x, y) = \tilde{c}(\varphi(x), \psi(y)) \quad \text{for all } x, y \in \mathcal{X} \times \mathcal{Y}.$$

and that there exists a Monge map $t_B : B_+ \to B_-$ that transports $\varphi_*\mu$ onto $\psi_*\nu$ for the cost $\tilde{c}$. Assume that there exists a disintegration $(\mu_u)_{u\in B_+}$ of μ with respect to φ such that $\varphi_*\mu$-a.e., μ_u is atomless.

Then there exists a Monge map between μ and ν for the cost c. Furthermore, it projects onto t_B through (φ, ψ), in the sense that $(\varphi, \psi)_*(\mathrm{id}, T)_*\mu = (\mathrm{id}, t_B)_*(\varphi_*\mu)$.

The proof of this theorem is provided in Section 2.2.2.

Remark 2.1. The atomless assumption on the disintegration $(\mu_u)_u$ is a natural minimal requirement to expect the existence of a map (without specific assumption on the target measure ν) and implies in particular that the fibers $(\varphi^{-1}(u))_{u\in B_+}$ should not be discrete (at least $\varphi_*\mu$-a.e.). Indeed, if for instance $\mathcal{X} = \mathcal{Y} = B_+ = B_- = \mathbb{R}$ and $\varphi : x \mapsto |x|$, the fibers of φ are of the form $\{-u, u\}$ (for $u \geq 0$), hence the disintegrations $(\mu_u)_{u\geq 0}$ and $(\nu_u)_{u\geq 0}$ are discrete and given by $\mu_u(u)\delta_u + (1 - \mu_u(u))\delta_{-u}$ and $\nu_u(u)\delta_u + (1-\nu_u(u))\delta_{-u}$, and there is in general no map T_u between two such discrete measures, unless we assume that $\mu_u(u) = \nu_u(u)$ or $1 - \nu_u(u)$, $\varphi_*\mu$-a.e.

Observe also that $\varphi_*\mu$ may have atoms: as we assume the existence of the Monge map t_B, it would

imply in that case that $\psi_*\nu$ also has atoms.

Remark 2.2. The "projection" property $(\varphi, \psi)_*(\mathrm{id}, T)_*\mu = (\mathrm{id}, t_B)_*(\varphi_*\mu)$ can also be written $\psi \circ T(x) = t_B \circ \varphi(x)$, for μ-a.e. x. A converse implication, that is "every Monge map between μ and ν projects onto a Monge map between $\varphi_*\mu$ and $\varphi_*\nu$" may not hold in general. This is however true if we can guarantee that there is a unique optimal transport plan between $\varphi_*\mu$ and $\psi_*\nu$ and that it is of the form $(\mathrm{id}, t_B)_*\mu$ (*e.g.* if we can apply Theorem 1.3)—in that case, T necessary projects onto t_B in the aforementioned sense.

Under additional assumptions, we can build a more structured Monge map. Namely, as our goal is to apply Proposition 1.4, we will assume that the (common) basis $B = B_+ = B_-$ is a manifold, that *almost all* the fibers of $\varphi : E \to B$ are homeomorphic to the *same* manifold F, and that every source measure of interest $(\mu, \mu_u, \varphi_*\mu)$ has a density. We also introduce the following convention: if $\mu \in \mathcal{P}(E)$ for some measurable space E, $E' \subset E$, and $\varphi : E' \to B$, we let $\varphi_*\mu$ be the (non-negative) measure supported on B defined by $\varphi_*\mu(A) = \mu(\varphi^{-1}(A))$ for $A \subset B$ measurable. If $\mu(E') = 1$, note that $\varphi_*\mu$ defines a probability measure on B (*i.e.* it has mass one). This formalism allows us to state our theorem even when some assumptions only hold λ-a.e.

Theorem 2.5. Let E_0 be a measurable space and B_0 and F be complete Riemannian manifolds. Let $\mu, \nu \in \mathcal{P}(E_0)$ be two probability measures with compact support. Assume that there exists a set $E \subset E_0$ such that $\mu(E) = 1$ and that there exists a measurable map $\Phi : E \to B_0 \times F$ that is injective and whose inverse on its image is measurable as well. Let p_B, p_F denote the projections of $B_0 \times F$ on B_0 and F respectively, and let $\varphi \triangleq p_B \circ \Phi : E \to B_0$. Let $c : E_0 \times E_0 \to \mathbb{R}$ and suppose that there exists a twisted $\tilde{c} : B_0 \times B_0 \to \mathbb{R}$ such that

$$c(x, y) = \tilde{c}(\varphi(x), \varphi(y)) \quad \text{for all } x, y \in E_0 .$$

Assume that $\varphi_*\mu$ is absolutely continuous w.r.t. the Lebesgue measure on B_0 and let thus t_B denote the unique Monge map between $\varphi_*\mu$ and $\varphi_*\nu$ for this cost. Suppose that there exists a disintegration $((\Phi_*\mu)_u)_{u \in B_0}$ of $\Phi_*\mu$ by p_B such that for $\varphi_*\mu$-a.e. u, $(\Phi_*\mu)_u$ is absolutely continuous w.r.t. the volume measure on F.
Then there exists an optimal map T between μ and ν for the cost c that can be decomposed as

$$\Phi \circ T \circ \Phi^{-1}(u, v) = (t_B(u), t_F(u, v)) = (\tilde{c}\text{-}\exp_u(\nabla f(u)), \exp_v(\nabla g_u(v))) , \tag{2.1}$$

with $f : B_0 \to \mathbb{R}$ $\tilde{c}$-convex and $g_u : F \to \mathbb{R}$ $d_F^2/2$-convex for $\varphi_*\mu$-a.e. u. Note that t_F could actually be any measurable function that sends each fiber $(\Phi_*\mu)_u$ onto $(\Phi_*\nu)_{t_B(u)}$.

The proof of this theorem is provided in Section 2.2.3. Let us give a simple example that illustrates the role played by our assumptions. This example has connections with (GW-Q) as detailed in Section 2.3.2.

Example 2.1. Let $E_0 = \mathbb{R}^d$ and $E = E_0 \backslash \{0\}$, let $B_0 = \mathbb{R}$ and $F = S^{d-1} = \{x \in E_0 \mid |x| = 1\}$. For convenience, we also introduce the space $B = \mathbb{R}_+^*$. Consider the cost function $c(x, y) = (|x| - |y|)^2$, so that c only depends on the norm of its entries. The fibers of the map $x \mapsto |x|$ are spheres, with the exception of $x = 0$, which invites us to consider the diffeomorphism

$$\begin{aligned} \Phi : E &\to \mathbb{R}_+^* \times S^{d-1} = B \times F \subset B_0 \times F \\ x &\mapsto \left(|x|, \frac{x}{|x|} \right) . \end{aligned}$$

From this, we can write $c(x, y) = \tilde{c}(\varphi(x), \varphi(y))$ where $\varphi(x) = |x|$ and $\tilde{c}(u, u') = (u - u')^2$ (which is twisted).

Now, if μ has a density on $\mathbb{R}^d$, so does $\Phi_*\mu$ on $B_0 \times F$ as Φ is a diffeomorphism. The coarea formula gives the existence of a disintegration $(\mu_u)_{u\in B}$ of $\Phi_*\mu$ by $p_B : (u, v) \mapsto u$ (note that $p_{B*}(\Phi_*\mu) = \varphi_*\mu$ also has a density) such that all the μ_u admit a density on S^{d-1}.

Our theorem thus applies, ensuring the existence of a structured Monge map between μ and (any) ν for the cost c: it decomposes for almost all $x \in \mathbb{R}^d$ as a Monge map on the basis $B_0 = \mathbb{R}$ (although it is actually only characterized on the image of φ, that is $B = \mathbb{R}_+^*$) obtained as the gradient of a convex function f (there is no need for the exponential map here and ∇f is the non-decreasing mapping between the quantiles of $\varphi_*\mu$ and $\varphi_*\nu$) and a Monge map on each fiber $F = S^{d-1}$, also built from gradients of convex functions (via the exponential map on the sphere).

Note that our theorem only requires assumptions to hold almost everywhere on $E_0 = \mathbb{R}^d$, which is important since it allows to ignore the singularity of φ at $x = 0$.

2.2.2 Proof of Theorem 2.4

The proof decomposes in three steps.

Step 1: Existence and optimality of lifts. We know by assumption that there exists a Monge map t_B that is optimal between the pushforward measures $\varphi_*\mu$ and $\psi_*\nu$.

As our goal is to build a Monge map between the initial measures μ and ν, we first show that (i) there exists a transport plan $\pi \in \Pi(\mu, \nu)$ such that $(\varphi, \psi)_*\pi = (\mathrm{id}, t_B)_*\mu$ and (ii) any such π is an optimal transport plan between μ and ν for the cost c. This is formalized by the following lemmas.

Lemma 2.6 (Existence of a lift)**.** For any transport plan $\tilde{\pi} \in \Pi(\varphi_*\mu, \psi_*\nu)$, there exists a transport plan $\pi \in \Pi(\mu, \nu)$ such that $(\varphi, \psi)_*\pi = \tilde{\pi}$.

Proof. Let $(\mu_u)_{u\in B_+}$ and $(\nu_v)_{v\in B_-}$ be disintegrations of μ and ν by φ and ψ respectively. Given $\tilde{\pi} \in \Pi(\varphi_*\mu, \psi_*\nu)$, we define

$$\pi \triangleq \iint_{B_+\times B_-} (\mu_u \otimes \nu_v)\, \mathrm{d}\tilde{\pi}(u, v)\,,$$

i.e. trivially sending every fiber μ_u onto every fiber ν_v, while weighting by $\tilde{\pi}$. See [AGS05, Sec. 5.3] for the notation. Then, for any Borel set $A \subset \mathcal{X}$,

$$\begin{aligned}
\pi(A \times \mathcal{Y}) &= \iint_{B_+\times B_-} \mu_u(A)\nu_v(\mathcal{Y})\, \mathrm{d}\tilde{\pi}(u, v) \\
&= \iint_{B_+\times B_-} \mu_u(A)\, \mathrm{d}\tilde{\pi}(u, v) \\
&= \int_{B_+} \mu_u(A)\, \mathrm{d}(\varphi_*\mu)(u) && \text{since the first marginal of } \tilde{\pi} \text{ is } \varphi_*\mu \\
&= \mu(A) && \text{by the disintegration theorem,}
\end{aligned}$$

and similarly for ν; hence $\pi \in \Pi(\mu, \nu)$. Now, let us show that $(\varphi, \psi)_*\pi = \tilde{\pi}$. For U and V Borel sets of

B_+ and B_- respectively,

$$\begin{aligned}
((\varphi,\psi)_*\pi)(U\times V) &= \iint_{U\times V} \mathrm{d}((\varphi,\psi)_*\pi)(u,v)\\
&= \iint_{\varphi^{-1}(U)\times\psi^{-1}(V)} \mathrm{d}\pi(x,y)\\
&= \iint_{\varphi^{-1}(U)\times\psi^{-1}(V)} \iint_{B_+\times B_-} \mathrm{d}(\mu_u\otimes\nu_v)(x,y)\,\mathrm{d}\tilde{\pi}(u,v)\\
&= \iint_{B_+\times B_-} \left(\int_{\varphi^{-1}(U)} \mathrm{d}\mu_u(x) \int_{\psi^{-1}(V)} \mathrm{d}\nu_v(y)\right) \mathrm{d}\tilde{\pi}(u,v) \qquad \text{by Fubini's theorem}\\
&= \iint_{B_+\times B_-} \mu_u(\varphi^{-1}(U))\nu_v(\psi^{-1}(V))\,\mathrm{d}\tilde{\pi}(u,v)\\
&= \iint_{B_+\times B_-} \delta_U(u)\delta_V(v)\,\mathrm{d}\tilde{\pi}(u,v)\\
&= \iint_{U\times V} \mathrm{d}\tilde{\pi}(u,v)\\
&= \tilde{\pi}(U\times V)\,.
\end{aligned}$$

□

Lemma 2.7 (Decomposition of optimal plans for the base space cost)**.** Let $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ and $\tilde{c} : B_+\times B_- \to \mathbb{R}$ such that

$$c(x,y) = \tilde{c}(\varphi(x),\psi(y)) \quad \text{for all } x,y \in \mathcal{X}\times\mathcal{Y}.$$

Then

$$\Pi_{\tilde{c}}^\star(\varphi_*\mu,\psi_*\nu) = (\varphi,\psi)_*\Pi_c^\star(\mu,\nu)\,,$$

where $\Pi_c^\star(\mu,\nu)$ denotes the set of optimal transport plan between μ and ν for the cost c, and similarly for $\Pi_{\tilde{c}}^\star(\varphi_*\mu,\psi_*\nu)$.

Proof. Let us first remark that for every $\tilde{\pi}\in\Pi(\varphi_*\mu,\psi_*\nu)$ and $\pi\in\Pi(\mu,\nu)$,

$$\text{if } \tilde{\pi} = (\varphi,\psi)_*\pi, \text{ then } \langle c,\,\pi\rangle = \langle\tilde{c},\,\tilde{\pi}\rangle. \tag{2.2}$$

Indeed, for such a $\tilde{\pi}$

$$\begin{aligned}
\iint_{B_+\times B_-} \tilde{c}(u,v)\,\mathrm{d}\tilde{\pi}(u,v) &= \iint_{\mathcal{X}\times\mathcal{Y}} \tilde{c}(\varphi(x),\psi(y))\,\mathrm{d}\pi(x,y) \qquad \text{by definition of the pushforward}\\
&= \iint_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\pi(x,y)\,.
\end{aligned}$$

$\subset$. Let $\tilde{\pi}^\star\in\Pi_{\tilde{c}}^\star(\varphi_*\mu,\psi_*\nu)$. By Lemma 2.6, there exists a $\pi\in\Pi(\mu,\nu)$ such that $(\varphi,\psi)_*\pi = \tilde{\pi}^\star$. Then for any $\gamma\in\Pi(\mu,\nu)$,

$$\langle c,\,\pi\rangle \overset{(2.2)}{=} \langle\tilde{c},\,\tilde{\pi}^\star\rangle \overset{(*)}{\leq} \langle\tilde{c},\,(\varphi,\psi)_*\gamma\rangle \overset{(2.2)}{=} \langle c,\,\gamma\rangle\,,$$

where $(*)$ follows from the optimality of $\tilde{\pi}^\star$. Hence the optimality of π.
$\supset$. Let $\pi^\star\in\Pi_c^\star(\mu,\nu)$. By Lemma 2.6, for any $\tilde{\gamma}\in\Pi(\varphi_*\mu,\psi_*\nu)$ there exists a $\gamma\in\Pi(\mu,\nu)$ such that $(\varphi,\psi)_*\gamma = \tilde{\gamma}$. We then have

$$\langle\tilde{c},\,(\varphi,\psi)_*\pi^\star\rangle \overset{(2.2)}{=} \langle c,\,\pi^\star\rangle \overset{(*)}{\leq} \langle c,\,\gamma\rangle \overset{(2.2)}{=} \langle\tilde{c},\,\tilde{\gamma}\rangle\,,$$

where $(*)$ follows from the optimality of $\pi^\star$. Hence the optimality of $(\varphi,\psi)_*\pi^\star$. □

Step 2: Existence of Monge maps between the fibers. Using Lemma 2.6 with $\tilde{\pi} = (\mathrm{id}, t_B)_*(\varphi_*\mu)$, we know that we can build an optimal transportation plan $\pi \in \Pi(\mu,\nu)$ that essentially coincides with t_B on $B_+ \times B_-$ and transports each fiber μ_u onto $\nu_{t_B(u)}$ for μ-a.e. $u \in B_+$. In order to build a Monge map between μ and ν, we must show that one can actually transport almost all μ_u onto $\nu_{t_B(u)}$ using a map rather than a plan. For this, we use the following result, see [San15, Rem. 1.23, Lemma 1.28, Cor. 1.29].

Proposition 2.8. Let α, β be two measures supported on $\mathbb{R}^d$ with α atomless. Then:
(i) if $d = 1$, there exists a transport map $\tilde{T}$ that pushes α onto β. Furthermore, it is the *unique* optimal map between these measures for the quadratic cost $(x,y) \mapsto |x-y|^2$;
(ii) there exists a map $\sigma_d : \mathbb{R}^d \to \mathbb{R}$ (that does not depend on α, β) that is (Borel) measurable, injective, and its inverse is measurable as well.

As we assumed that the ground spaces $\mathcal{X}$ and $\mathcal{Y}$ can be embedded in $\mathbb{R}^d$ using the injective, measurable maps $\Phi_\mathcal{X}$ and $\Phi_\mathcal{Y}$, we can apply Proposition 2.8 using $\sigma_\mathcal{X} = \sigma_d \circ \Phi_\mathcal{X}$ and $\sigma_\mathcal{Y} = \sigma_d \circ \Phi_\mathcal{Y}$. As $\sigma_\mathcal{X}$ is injective, $\sigma_{\mathcal{X}*}\mu_u$ is atomless on $\mathbb{R}$, and we can thus consider the *unique* Monge map $\tilde{T}_u$ between $\sigma_{\mathcal{X}*}\mu_u$ and $\sigma_{\mathcal{Y}*}\nu_{t_B(u)}$ for the quadratic cost on $\mathbb{R}$.

From this, as the maps $\sigma_\mathcal{X}$ and $\sigma_\mathcal{Y}$ are measurable and injective (thus invertible on their image) we can define $T_u = \sigma_\mathcal{Y}^{-1} \circ \tilde{T}_u \circ \sigma_\mathcal{X} : \mathcal{X} \to \mathcal{Y}$, that defines a (measurable) transport map between μ_u and $\nu_{t_B(u)}$.

Step 3: building a measurable global map. Now that we have maps $(T_u)_u$ between each μ_u and $\nu_{t_B(u)}$, it may be tempting to simply define a map $T : \mathcal{X} \to \mathcal{Y}$ by $T(x) = T_{\varphi(x)}(x)$ when $\mu_{\varphi(x)}$ is atomless (which, by assumption, holds μ-a.e.). Intuitively, this map induces a transport plan $(\mathrm{id}, T)_*\mu$ that satisfes $(\varphi,\psi)_*(\mathrm{id}, T)_*\mu = (\mathrm{id}, t_B)_*(\varphi_*\mu)$ on $B_+ \times B_-$ and thus must be optimal according to Lemma 2.7.

One remaining step, though, is to prove that this map T can be defined in a measurable way. For this, we use the following *measurable selection theorem* due to [FGM10, Thm. 1.1], that reads:

Proposition 2.9. Let (B, Σ, m) be a σ-finite measure space and consider a measurable function $B \ni u \mapsto (\mu_u, \nu_u) \in \mathcal{P}(\mathbb{R}^d)^2$. Let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a cost function, and assume that for m-a.e. $u \in B$, there is a (unique) Monge map T_u between μ_u and ν_u for the cost c.
Then there exists a measurable function $(u,x) \mapsto T(u,x)$ such that m-a.e., $T(u,x) = T_u(x)$, μ_u-a.e.

We can apply this result in the case $d = 1$ to the family of measures $(\sigma_{\mathcal{X}*}\mu_u, \sigma_{\mathcal{Y}*}\nu_{t_B(u)})_{u\in B_+}$, where the reference measure on B_+ is $\varphi_*\mu$.[1] We first need to show the measurability of this family of measures. By definition of the disintegration of measures (see for instance [AGS05, Thm. 5.3.1]), the map $v \in B_- \mapsto \nu_v$ is measurable; and as the Monge map t_B is measurable as well, so is the map $B_+ \ni u \mapsto \sigma_{\mathcal{Y}*}\nu_{t_B(u)}$ by composition of measurable maps, and thus so is the map $u \mapsto (\mu_u, \nu_{t_B(u)})$. Proposition 2.9 therefore applies and guarantees the existence of a measurable map $\tilde{T} : B_+ \times \mathbb{R} \to \mathbb{R}$ such that $\tilde{T}(u,x) = \tilde{T}_u(x)$ for $\varphi_*\mu$ almost all u and $\sigma_{\mathcal{X}*}\mu$ almost all x. Now, we can define

$$\begin{aligned} T : \mathcal{X} &\to \mathcal{Y} \\ x &\mapsto \sigma_\mathcal{Y}^{-1} \circ \tilde{T}(\varphi(x), \sigma_\mathcal{X}(x)) \,. \end{aligned}$$

This map is measurable as composition of measurable maps. Let us prove that this defines a transport map between μ and ν. For any function $g : \mathcal{Y} \to \mathbb{R}$ continuous with compact support, we can write

$$\int_\mathcal{Y} g(y)\,\mathrm{d}T_*\mu(y) = \int_\mathcal{X} g(T(x))\,\mathrm{d}\mu(x) = \int_{u\in B_+} \int_{x\in\varphi^{-1}(\{u\})} g\left(\sigma_\mathcal{Y}^{-1}\left(\tilde{T}_u(\sigma_\mathcal{X}(x))\right)\right)\,\mathrm{d}\mu_u(x)\,\mathrm{d}\varphi_*\mu(u)\,,$$

where we use the disintegration of μ w.r.t. φ and the fact that the μ_u are supported on $\varphi^{-1}(\{u\})$, allowing us to write $\tilde{T}(\varphi(x), \sigma_\mathcal{X}(x)) = \tilde{T}_u(\sigma_\mathcal{X}(x))$ on that fiber ($\varphi_*\mu$-a.e.).

[1] Note that we cannot apply Proposition 2.9 to the measures $(\mu_u, \nu_{t_B(u)})_u$ and the maps $(T_u)_u$ directly, as T_u may not be the unique Monge map between the measures, a required assumption of the proposition.

Now, recall that $T_u : x \mapsto \sigma_{\mathcal{Y}}^{-1}\left(\tilde{T}_u(\sigma_{\mathcal{X}}(x))\right)$ defines a transport map between μ_u and $\nu_{t_B(u)}$. In particular, the image of the fiber $\varphi^{-1}(\{u\})$ by this map is $\psi^{-1}(\{t_B(u)\}) \subset \mathcal{Y}$. Therefore, we get

$$\begin{aligned}
\int_{\mathcal{Y}} g(y)\,\mathrm{d}T_*\mu(y) &= \int_{u\in B_+}\int_{y\in\psi^{-1}(\{t_B(u)\})} g(y)\,\mathrm{d}\nu_{t_B(u)}\,\mathrm{d}\varphi_*\mu(u)\\
&= \int_{u\in B_+}\int_{y\in\mathcal{Y}} g(y)\,\mathrm{d}\nu_{t_B(u)}\,\mathrm{d}\varphi_*\mu(u) && \text{as } \nu_{t_B(u)} \text{ is supported on } \psi^{-1}(\{t_B(u)\})\\
&= \int_{v\in B_-}\int_{y\in\mathcal{Y}} g(y)\,\mathrm{d}\nu_v(y)\,\mathrm{d}t_{B*}(\varphi_*\mu)(v) && \text{by change of variable } v = t_B(u)\\
&= \int_{v\in B_-}\int_{y\in\mathcal{Y}} g(y)\,\mathrm{d}\nu_v(y)\,\mathrm{d}\psi_*\nu(v) && \text{as } t_B \text{ pushes } \varphi_*\mu \text{ to } \psi_*\nu\\
&= \int_{y\in\mathcal{Y}} g(y)\,\mathrm{d}\nu(y) && \text{as } (\nu_v)_v \text{ is a disintegration of } \nu \text{ by } \psi\,,
\end{aligned}$$

proving that $T_*\mu = \nu$.

By Lemma 2.7, this map is optimal if and only if it satisfies $(\varphi,\psi)_*(\mathrm{id},T)_*\mu = (\mathrm{id},t_B)_*(\varphi_*\mu)$, as t_B is an optimal transportation plan between $\varphi_*\mu$ and $\psi_*\nu$, making $(\mathrm{id},T)_*\mu$ optimal between μ and ν (hence T a Monge map).

For this, let $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ be a continuous function with compact support. We have

$$\begin{aligned}
\iint_{B_+\times B_-} g(u,v)\,\mathrm{d}(\varphi,\psi)_*(\mathrm{id},T)_*\mu(u,v) &= \iint_{\mathcal{X}\times\mathcal{Y}} g(\varphi(x),\psi(y))\,\mathrm{d}(\mathrm{id},T)_*\mu(x,y)\\
&= \int_{\mathcal{X}} g(\varphi(x),\psi(T(x)))\,\mathrm{d}\mu(x)\\
&= \int_{u\in B_+}\int_{x\in\varphi^{-1}(\{u\})} g(u,\psi(\sigma_{\mathcal{Y}}^{-1}(T_u(\sigma_{\mathcal{X}}(x)))))\,\mathrm{d}\mu_u(x)\,\mathrm{d}\varphi_*\mu(u)\\
&= \int_{u\in B_+}\int_{y\in\psi^{-1}(\{t_B(u)\})} g(u,t_B(u))\,\mathrm{d}\nu_{t_B(u)}(y)\,\mathrm{d}\varphi_*\mu(u)\\
&= \int_{u\in B_+} g(u,t_B(u))\,\mathrm{d}\varphi_*\mu(u)\\
&= \int_{B_+\times B_-} g(u,v)\,\mathrm{d}(\mathrm{id},t_B)_*\varphi_*\mu(u,v)\,,
\end{aligned}$$

proving the required equality and thus that T is a Monge map between μ and ν.

2.2.3 Proof of Theorem 2.5

To alleviate notations, we let $\mu' \triangleq \Phi_*\mu$ and $\nu' \triangleq \Phi_*\nu$ in the following. We also denote by B the image of $\varphi = p_B \circ \Phi$, so that μ', ν' are supported on $B\times F \subset B_0\times F$.

Step 1: Construction of the structured Monge map. Given that $\varphi_*\mu$ is absolutely continuous w.r.t. the Lebesgue measure on the complete (separable) Riemannian manifold B_0, by Theorem 1.3 there exists a unique optimal transport plan $\pi_B^\star$ between $\varphi_*\mu$ and $\varphi_*\nu$ for the cost $\tilde{c}$ and it is induced by a map $t_B : B_0 \to B_0$ of the form $t_B = \exp_u(\nabla f)$, with f $\tilde{c}$-convex.

By Lemma 2.7, we know that any transport plan in $\pi \in \Pi(\mu,\nu)$ that satisfy $(\varphi,\varphi)_*\pi = (\mathrm{id},t_B)_*\mu$ must be optimal. Therefore, if π happens to be induced by a map T, that is $\pi = (\mathrm{id},T)_*\mu$, we would obtain a Monge map between μ and ν. To build such a T, we proceed as in Section 2.2.2: we define a Monge map T_u between $(\mu'_u)_u$ and $(\nu'_{t_B(u)})_u$ for $\varphi_*\mu$-a.e. u (recall that those are the disintegration of

$\Phi_*\mu = \mu'$ and $\Phi_*\nu = \nu'$ with respect to p_B) and build a global map between μ' and ν' by (roughly) setting $T(u,x) = T_u(x)$. As in Section 2.2.2, proving the measurability of such T requires care.

Step 2: Transport between the fibers. For $\varphi_*\mu$-a.e. u, μ'_u has a density w.r.t. the volume measure on F and the optimal cost between μ'_u and $\nu'_{t_B(u)}$ is finite by assumption. Whenever μ'_u has a density, we can therefore apply Proposition 1.4 between μ'_u and $\nu'_{t_B(u)}$ with the cost d_F^2 to obtain that there exists a plan π_u between these fibers that is induced by a map $T_u : F \to F$ that can be expressed as $T_u(v) = \exp_v(\tilde{\nabla} g_u(v))$ with g_u being $d_F^2/2$-convex on F.

Step 3: Measurability of the global map. Now that we have built structured maps T_u between corresponding fibers (through t_B), it remains to prove the existence of a measurable map $T : B_0 \times F \to B_0 \times F$ transporting μ' onto ν' satisfying $T(u,x) = (t_B(u), T_u(x))$ for $\varphi_*\mu$-almost every u and μ'_u-almost every x.

For this, we need an adaptation of Proposition 2.9 to the manifold setting. Namely, we have the following:

Proposition 2.10 (Measurable selection of maps, manifold case)**.** Let M be a complete Riemannian manifold and (B, Σ, m) a measure space. Consider a measurable function $B \ni u \mapsto (\mu_u, \nu_u) \in \mathcal{P}(M)^2$. Assume that for m-almost every $u \in B$, $\mu_u \ll \mathrm{vol}_M$ and μ_u and ν_u have a finite transport cost. Let T_u denote the (unique by Proposition 1.4) optimal transport map induced by the quadratic cost d_M^2 on M between μ_u and ν_u.
Then there exists a function $(u,x) \mapsto T(u,x)$, measurable w.r.t. $\Sigma \otimes \mathcal{B}(\mathbb{R}^d)$, such that m-a.e.,

$$T(u,x) = T_u(x) \quad \mu_u\text{-a.e.}$$

This proposition can essentially be proved by adapting the proof of [FGM10] to the manifold setting, and most steps adapt seamlessly. A complete proof, where we stress the points that need specific care in adaptation, is deferred to Appendix A.3.

We can apply this proposition with the manifold being the (common) fiber F on which the $\mu'_u, \nu'_{t_B(u)}$ are supported for $\varphi_*\mu$-a.e. u, and for which we have access to the (unique) Monge map T_u. It gives the existence of a global map t_F satisfying $t_F(u,v) = T_u(v)$ for $\varphi_*\mu$-a.e. u, and μ'_u-a.e. v, and we can thus define the (measurable) map $T(u,x) = (t_B(u), t_F(u,x))$.

One then has for any continuous function z with compact support:

$$\begin{aligned}
\int_{B_0\times F} z(u',v')\,\mathrm{d}(T_*\Phi_*\mu)(u',v') &= \int_{B_0\times F} z(t_B(u), T_u(v))\,\mathrm{d}(\Phi_*\mu)(u,v) && \text{(pushforward } T \text{ on } \Phi_*\mu)\\
&= \iint_{B_0\times F} z(t_B(u), T_u(v))\,\mathrm{d}(\Phi_*\mu)_u(v)\,\mathrm{d}(\varphi_*\mu)(u) && \text{(disintegration theorem)}\\
&= \iint_{B_0\times F} z(t_B(u), v')\,\mathrm{d}(g_{u*}(\Phi_*\mu)_u)(v')\,\mathrm{d}(\varphi_*\mu)(u) && \text{(pushforward } T_u \text{ on } (\Phi_*\mu)_u)\\
&= \iint_{B_0\times F} z(t_B(u), v')\,\mathrm{d}((\Phi_*\nu)_{t_B(u)})(v')\,\mathrm{d}(\varphi_*\mu)(u) && (g_{u*}(\Phi_*\mu)_u = (\Phi_*\nu)_{t_B(u)})\\
&= \iint_{B_0\times F} z(u',v')\,\mathrm{d}((\Phi_*\nu)_{u'})(v')\,\mathrm{d}(t_B\varphi_*\mu)(u') && \text{(pushforward } t_B \text{ on } \varphi_*(\Phi_*\mu)_u)\\
&= \iint_{B_0\times F} z(u',v')\,\mathrm{d}((\Phi_*\nu)_{u'})(v')\,\mathrm{d}(\varphi_*\nu)(u') && (t_{B*}(\varphi_*\mu) = \varphi_*\nu)\\
&= \int_{B_0\times F} z(u',v')\,\mathrm{d}(\Phi_*\nu)(u',v')\,, && \text{(disintegration theorem)}
\end{aligned}$$

hence T sends $\Phi_*\mu$ to $\Phi_*\nu$ and $T_E \triangleq \Phi^{-1} \circ T \circ \Phi$ therefore sends μ to ν; and since

$$(\varphi,\varphi)_*(\mathrm{id}, T_E)_*\mu = (\varphi, \varphi\circ T_E)_*\mu = (\varphi, t_B\circ\varphi)_*\mu = (\mathrm{id}, t_B)_*\varphi_*\mu = \pi_B^\star\,,$$

we have that T_E is an optimal map between μ and ν.

2.3 Applications to the quadratic and inner-product GW problems

2.3.1 The inner-product cost

We recall the (GW-IP) problem:

$$\min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \left|\langle x,\, x'\rangle - \langle y,\, y'\rangle\right|^2 \,\mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y')\,, \tag{GW-IP}$$

Expanding the integrand and using the fact that $\iint \langle x,\, x'\rangle^2 \,\mathrm{d}\pi\,\mathrm{d}\pi = \iint \langle x,\, x'\rangle^2 \,\mathrm{d}\mu\,\mathrm{d}\mu$ is constant (the same goes for the terms that depend on ν), (GW-IP) is equivalent to

$$\min_{\pi\in\Pi(\mu,\nu)} \iint -\langle x,\, x'\rangle\langle y,\, y'\rangle \,\mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y')\,.$$

This problem is not invariant to translations but it is to the action of $O_n(\mathbb{R}) \times O_d(\mathbb{R})$. Assuming an optimal correspondence plan $\pi^\star$, this plan is also an optimal transport plan for the linearized problem (1.11) with cost

$$C_{\pi^\star}(x,y) = -\int \langle x,\, x'\rangle\langle y,\, y'\rangle \,\mathrm{d}\pi^\star(x',y') = \left\langle -\int (y'\otimes x')x \,\mathrm{d}\pi^\star(x',y'),\, y\right\rangle = -\langle M^\star x,\, y\rangle\,,$$

where $M^\star \triangleq \int y'\otimes x' \,\mathrm{d}\pi^\star(x',y') \in \mathbb{R}^{d\times n}$.

One can already check if this linearized cost satisfies the twist conditions for admitting an optimal transport map defined in Sec. 1.2.1:

Table 2.1: Twist conditions for the cost $c(x,y) = -\langle M^\star x,\, y\rangle$. See Appendix A.2 for a proof.

$\operatorname{rk} M^\star$	$= d$	$\leq d-1$
twist	✓	·
subtwist	✓	·
m-twist, $m \geq 2$	✓	·
non-degeneracy	✓	·

This linearized cost satisfies the (Twist) condition if and only if $M^\star$ is of full rank, hence in this case the solution $\pi^\star$ of (GW-IP) is unique and induced by a map, and Theorem 4.2.3 from [Vay20] gives a result on the structure of this map. We can actually generalize this to the case where $M^\star$ is arbitrary:

Theorem 2.11 (Existence of an optimal map for the inner product cost)**.** Let $n \geq d$, $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)\times\mathcal{P}(\mathbb{R}^d)$ two measures with compact supports. Suppose that $\mu \ll \mathcal{L}^n$. Then there exists an optimal map for (GW-IP) that can be written as

$$T = O_{\mathcal{Y}}^\top \circ (T_0 \circ p_{\mathbb{R}^d}) \circ O_{\mathcal{X}}\,, \tag{2.3}$$

where $O_{\mathcal{X}}$ and $O_{\mathcal{Y}}$ are change-of-basis matrices of $\mathbb{R}^n$ and $\mathbb{R}^d$, $p_{\mathbb{R}^d} : \mathbb{R}^n \to \mathbb{R}^d$ is defined by $p_{\mathbb{R}^d}(x_1,\dots,x_n) = (x_1,\dots,x_d)$, and

$$T_0(x_1,\dots,x_d) = (\nabla f \circ \Sigma(x_1,\dots,x_h), \nabla g_{x_1,\dots,x_h}(x_{h+1},\dots,x_d))\,, \tag{2.4}$$

with $h \leq d$, $\Sigma \in \mathbb{R}^{h\times h}$ diagonal with positive entries, $f : \mathbb{R}^h \to \mathbb{R}$ convex and all $g_{x_1,\dots,x_h} : \mathbb{R}^{d-h} \to \mathbb{R}$ convex.

In order to show this, we will need two simple lemmas that we state now and prove in Appendix A.1, the second one being a simple corollary of the first:

Lemma 2.12. Let $\mu, \nu \in \mathcal{P}(E)$ and let $\psi_1, \psi_2 : E \to F$ be homeomorphisms. Let $\tilde{c} : F \times F \to \mathbb{R}$ and consider the cost $c(x, y) = c(\psi_1(x), \psi_2(y))$. Then a map is optimal for the cost c between μ and ν if and only if it is of the form $\psi_2^{-1} \circ T \circ \psi_1$ with T optimal for the cost $\tilde{c}$ between $\psi_{1*}\mu$ and $\psi_{2*}\nu$.

Lemma 2.13 (Brenier with scaled inner product)**.** Let $h \geq 1$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^h)$ with $\mu \ll \mathcal{L}^h$ with compact supports. Consider the cost $c(x, y) = -\langle \psi_1(x), \psi_2(y) \rangle$ where $\psi_1, \psi_2 : \mathbb{R}^h \to \mathbb{R}^h$ are diffeomorphisms. Then, there exists a unique optimal transport plan between μ and ν for the cost c, and it is induced by a map $t : \mathbb{R}^h \to \mathbb{R}^h$ of the form $t = \psi_2^{-1} \circ \nabla f \circ \psi_1$, with f convex.

We are now ready to prove Theorem 2.11:

Proof of Theorem 2.11. Using a singular value decomposition, we have $M^\star = O_{\mathcal{Y}}^\top \Sigma O_{\mathcal{X}} \in \mathbb{R}^{d \times n}$ with $O_{\mathcal{X}}, O_{\mathcal{Y}} \in O_n(\mathbb{R}) \times O_d(\mathbb{R})$ orthogonal matrices of each Euclidean space and $\Sigma \in \mathbb{R}^{d \times n}$ diagonal with non-negative coefficients. The cost then becomes $C_{\pi^\star}(x, y) = -\langle O_{\mathcal{Y}}^\top \Sigma O_{\mathcal{X}} x,\, y \rangle = -\langle \Sigma(O_{\mathcal{X}} x),\, O_{\mathcal{Y}} y \rangle$. Using Lemma 2.12, the problem transforms into an optimal transportation problem between $\mu' \triangleq O_{\mathcal{X}}\mu$ and $\nu' \triangleq O_{\mathcal{Y}}\nu$; and choosing $O_{\mathcal{Y}}$ and $O_{\mathcal{X}}$ that sort the singular values in decreasing order, *i.e.* assuming $\sigma_1 \geq \cdots \geq \sigma_h > 0$ with $h \triangleq \mathrm{rk}(M^\star) \leq d$, the problem therefore transforms into $\min_{\tilde{\pi}} \langle c_\Sigma, \tilde{\pi} \rangle$ for $\tilde{\pi} \in \Pi(\mu', \nu')$, where $c_\Sigma(\tilde{x}, \tilde{y}) = -\sum_{i=1}^h \sigma_i \tilde{x}_i \tilde{y}_i \triangleq \tilde{c}(p(\tilde{x}), p(\tilde{y}))$, p being the orthogonal projection on $\mathbb{R}^h$. We reduce to the case where both measures live in the same space by noting that since $c_\Sigma(\tilde{x}, \tilde{y}) = c_\Sigma(p_{\mathbb{R}^d}(\tilde{x}), \tilde{y})$ for all $\tilde{x}$ and $\tilde{y}$, any map T_0 optimal between $\mu'' \triangleq p_{\mathbb{R}^d *}\mu'$ and ν' will induce a map $T = T_0 \circ p_{\mathbb{R}^d}$ optimal between μ' and ν'[2]. One can then recover an optimal map between μ and ν by composing with $O_{\mathcal{X}}$ and $O_{\mathcal{Y}}^\top$ (Lemma 2.12), hence Eq. (2.3).

The existence of such a map T_0 optimal between μ'' and ν' satisfying (2.4) follows from the application of Theorem 2.5 for $E = E_0 = \mathbb{R}^d = \mathbb{R}^h \times \mathbb{R}^{d-h} = B_0 \times F$ and $\varphi = p$. Indeed, B_0 and F are complete Riemannian manifolds; $\tilde{c}$ is twisted on $B_0 \times B_0$; $p_*\mu''$ has a density on B_0 and every $(\mu'')_u$ has a density w.r.t. the Lebesgue measure on F as a conditional probability. We then make t_B explicit. One has that $c_\Sigma(x, y) = -\langle \tilde{\Sigma} x,\, y \rangle$, where $\tilde{\Sigma} = \mathrm{diag}(\sigma_i)_{1 \leq i \leq h}$. As $p_*\mu''$ has a density, we can apply Lemma 2.13 stated above with $(\psi_1, \psi_2) = (\tilde{\Sigma}, \mathrm{id})$ to obtain that there exists a unique optimal transport plan $\pi_B^\star$ between $p_*\mu''$ and $p_*\nu'$ for the cost c_Σ and that it is induced by a map $t_B : B \to B$ of the form $t_B = \nabla f \circ \tilde{\Sigma}$, with f convex. □

Remark 2.3. A special case of our theorem is Theorem 4.2.3 from [Vay20] (Proposition 2.1 in this work): when $h = d$, the optimal map between $O_{\mathcal{X}}\mu$ and $O_{\mathcal{Y}}\nu$ writes $T_0 \circ p_{\mathbb{R}^d}$ with $T_0 = \nabla f \circ \tilde{\Sigma}$. The induced optimal map between μ and ν is then:

$$
\begin{aligned}
T &= O_{\mathcal{Y}}^\top \circ (\tilde{T}_0 \circ p_{\mathbb{R}^d}) \circ O_{\mathcal{X}} \\
&= O_{\mathcal{Y}}^\top \circ (\nabla f \circ \tilde{\Sigma} \circ p_{\mathbb{R}^d}) \circ O_{\mathcal{X}} \\
&= O_{\mathcal{Y}}^\top \circ (\nabla f \circ \Sigma) \circ O_{\mathcal{X}} \\
&= \nabla(f \circ O_{\mathcal{Y}}) \circ O_{\mathcal{Y}}^\top \circ \Sigma \circ O_{\mathcal{X}} && \text{since } \nabla(f \circ A) = A^\top \circ \nabla f \circ A \\
&= \nabla \tilde{f} \circ M^\star\,,
\end{aligned}
$$

where $\tilde{f} \triangleq f \circ O_{\mathcal{Y}}$ is convex.

[2] by Lemma 2.7 it suffices to check that $(p_{\mathbb{R}^d}, \mathrm{id})_*(\mathrm{id}, T)_*\mu'$ is in $\Pi^\star(p_{\mathbb{R}^d *}\mu', \nu')$:

$$(p_{\mathbb{R}^d}, \mathrm{id})_*(\mathrm{id}, T)_*\mu' = (p_{\mathbb{R}^d}, T_0 \circ p_{\mathbb{R}^d})_*\mu' = (\mathrm{id}, T_0)p_{\mathbb{R}^d *}\mu'\,.$$

2.3.2 The quadratic cost

We recall the (GW-Q) problem:

$$\min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} \left||x-x'|^2 - |y-y'|^2\right|^2 \mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y')\,, \tag{GW-Q}$$

which is invariant by translation of μ and ν. With no loss of generality, we suppose both measures centered. Expanding the integrand provides

$$\left||x-x'|^2 - |y-y'|^2\right|^2 = |x-x'|^4 + |y-y'|^4 - 2|x-x'|^2|y-y'|^2\,,$$

and the two first terms only depend on μ and ν, not on π. Expanding the remaining term yields nine terms. Two of them also lead to a constant contribution: $-|x|^2|y'|^2$ and $-|x'|^2|y|^2$; four lead to vanishing integrals since μ and ν are centered: $2|x|^2\langle y,\, y'\rangle, 2|x'|^2\langle y,\, y'\rangle, 2|y|^2\langle x,\, x'\rangle$ and $2|y'|^2\langle x,\, x'\rangle$. The remaining three terms then yield the following equivalent problem:

$$\min_{\pi\in\Pi(\mu,\nu)} \int -|x|^2|y|^2\,\mathrm{d}\pi(x,y) + 2\iint -\langle x,\, x'\rangle\langle y,\, y'\rangle\,\mathrm{d}\pi(x,y)\,\mathrm{d}\pi(x',y')\,.$$

Assuming an optimal correspondence plan $\pi^\star$, this plan is also an optimal transport plan for the linearized problem (1.11) with cost

$$C_{\pi^\star}(x,y) = -|x|^2|y|^2 - 4\int \langle x,\, x'\rangle\langle y,\, y'\rangle\,\mathrm{d}\pi^\star(x',y') = -|x|^2|y|^2 - 4\langle M^\star x,\, y\rangle\,,$$

where $M^\star \triangleq \int y'\otimes x'\,\mathrm{d}\pi^\star(x',y') \in \mathbb{R}^{d\times n}$.

One can already check if this linearized cost satisfies the twist conditions for admitting an optimal transport map defined in Sec. 1.2.1:

Table 2.2: Twist conditions for the cost $c(x,y) = -|x|^2|y|^2 - 4\langle M^\star x,\, y\rangle$. See Appendix A.2 for a proof.

$\mathrm{rk}\, M^\star$	$= d$	$= d-1$	$\leq d-2$
twist	·	·	·
subtwist	✓	·	·
2-twist	·	✓	·
m-twist, $m\geq 3$	·	·	·
non-degeneracy	∼	·	·

In the cases where the rank of $M^\star$ is d (resp. $d-1$), this linearized cost satisfies the subtwist (resp. 2-twist) condition, yielding an optimal map/anti-map (resp. bimap) structure by compactness of the support of μ and ν when μ has a density. In the case where $\mathrm{rk}\, M^\star \leq d-2$, nothing can be said and there is *a priori* little hope for the existence of an optimal correspondence map; but quite not unsurprisingly, we can actually prove it.

Theorem 2.14 (Existence of an optimal map, bimap or map/anti-map for the quadratic cost)**.** Let $n\geq d$, $\mu,\nu\in\mathcal{P}(\mathbb{R}^n)\times\mathcal{P}(\mathbb{R}^d)$ two measures with compact supports. Suppose that $\mu \ll \mathcal{L}^n$. Let $\pi^\star$ be a solution of (GW-Q) and $M^\star \triangleq \int y'\otimes x'\,\mathrm{d}\pi^\star(x',y')$. Then:

(i) if $\mathrm{rk}\, M^\star = d$, there exists an optimal plan that is induced by a map/anti-map;
(ii) if $\mathrm{rk}\, M^\star = d-1$, there exists an optimal plan that is induced by a bimap;
(iii) if $\mathrm{rk}\, M^\star \leq d-2$, there exists an optimal plan that is induced by a map that can be written as

$$T = O_{\mathcal{Y}}^\top \circ T_0 \circ O_{\mathcal{X}}\,,$$

where $O_{\mathcal{X}}$ and $O_{\mathcal{Y}}$ are change-of-basis matrices of $\mathbb{R}^n$ and, writing any $x\in\mathbb{R}^n$ as $x = (x_H, x_\perp)\in$

$\mathbb{R}^h \times \mathbb{R}^{n-h}$ and $\Phi(x) \triangleq (x_B, x_F) \triangleq ((x_H, |x_\perp|^2), x_\perp/|x_\perp|)$,

$$\Phi \circ T_0(x) = \Big(\tilde{c}\text{-}\exp_{x_B}(\nabla f(x_B)), \exp_{x_F}(\nabla g_{x_B}(x_F))\Big)$$

with $h = \operatorname{rk} M^\star \leq d-2$, $f : \mathbb{R}^{h+1} \to \mathbb{R}$ being $\tilde{c}$-convex and all $g_{x_B} : \mathbb{R}^{n-h} \to \mathbb{R}$ being $d^2_{S^{n-h-1}}/2$-convex.

Proof.

(i) We show that in this case the subtwist condition is satisfied. Consider $y_1 \neq y_2 \in \mathcal{Y}$. Any $x \in \mathcal{X}$ is a zero of $\nabla_x c(x, y_1) - \nabla_x c(x, y_2)$ if and only if

$$(|y_1|^2 - |y_2|^2)x = -(M^\star)^\top (y_1 - y_2)\,. \tag{2.5}$$

Suppose that M is of full rank. If $|y_1| = |y_2|$, then (2.5) has no solution since $y_1 - y_2$ cannot be in $\operatorname{Ker}(M^\star)^\top$, and if $|y_1| \neq |y_2|$, then (2.5) has a unique solution $x^\star = -(|y_1|^2 - |y_2|^2)^{-1}(M^\star)^\top(y_1 - y_2)$; hence the result.

(ii) We show that in this case the 2-twist condition is satisfied. Consider $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$. Let $y \in \mathcal{Y}$ such that $|y|^2 x_0 + (M^\star)^\top y = |y_0|^2 x_0 + (M^\star)^\top y_0$. Similarly to the inner product case, up to singular value decomposition suppose $M^\star$ rectangular diagonal in $\mathbb{R}^{d\times n}$ with sorted singular values and write $\tilde{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_h)$ with $h \triangleq \operatorname{rk}(M^\star)$. Denote by $v \in \mathbb{R}^d$ the right-hand side of the equation and decompose each vector z of $\mathbb{R}^n$ or $\mathbb{R}^d$ as $z = (z_H, z_\perp)$, where $z_H \in \mathbb{R}^h$ and $z_\perp$ contains the remaining coordinates. The equation becomes:

$$\begin{cases} |y|^2 x_H + \tilde{\Sigma} y_H &= v_H \\ |y|^2 x_\perp &= v_\perp\,. \end{cases}$$

If $x_\perp$ and $v_\perp$ are not colinear then it is absurd and there is no such y; else, since $x_\perp$ and $v_\perp$ are fixed, this means that $|y|^2$ is fixed and y lives on the $(d-1)$-dimensional sphere S^{d-1}. The first equation of the system above then gives $y_H = \tilde{\Sigma}^{-1}(v_H - |y|^2 x_H)$; hence y lives in the intersection of S^{d-1} and of a $(d-r)$-dimensional affine subspace of vectors $z \in \mathbb{R}^d$ with fixed z_H, and this intersection is either empty or a $(d-r-1)$-dimensional affine sphere. As $r = d-1$, y belongs to a set of at most 2 points and the 2-twist condition is satisfied.

(iii) The case $\operatorname{rk} M^\star \leq d-2$ is a consequence of Theorem 2.5 and the proof is as follows. We consider the measure ν as a measure of $\mathbb{R}^n$ of d-dimensional support. Similarly to the inner product cost, by SVD the cost becomes $c(x, y) = -|x|^2|y|^2 - \langle O_{\mathcal{Y}}^\top \Sigma O_{\mathcal{X}} x,\, y\rangle = -|O_{\mathcal{X}} x|^2 |O_{\mathcal{Y}} y|^2 - \langle \Sigma(O_{\mathcal{X}} x),\, O_{\mathcal{Y}} y\rangle$, and using Lemma 2.12 the problem transforms into $\min_{\tilde{\pi}} \langle c_\Sigma,\, \tilde{\pi}\rangle$ for $\tilde{\pi} \in \Pi(O_{\mathcal{X}}\mu, O_{\mathcal{Y}}\nu)$, where $c_\Sigma(x, y) \triangleq -|x|^2|y|^2 - \langle \Sigma x,\, y\rangle$. Further assuming $\sigma_1 \geq \cdots \geq \sigma_h > 0$ and writing any $z \in \mathbb{R}^n$ as $z = (z_u, z_v) \in \mathbb{R}^h \times \mathbb{R}^{n-h}$,

$$\begin{aligned} c_\Sigma(x, y) &= -|x_H|^2|y_H|^2 - |x_H|^2|y_\perp|^2 - |x_\perp|^2|y_H|^2 - |x_\perp|^2|y_\perp|^2 - \langle \tilde{\Sigma} x_H,\, y_H\rangle \\ &= -|x_H|^2|y_H|^2 - |x_H|^2 n_y - n_x|y_H|^2 - n_x n_y - \langle \tilde{\Sigma} x_H,\, y_H\rangle \quad \text{with } n_x = |x_\perp|^2 \text{ and } n_y = |y_\perp|^2 \\ &\triangleq \tilde{c}(\varphi(x), \varphi(y))\,, \end{aligned}$$

where $\varphi : x \mapsto (x_H, |x_\perp|^2)$, and the cost $c_\Sigma(x, y)$ only depends of the values of $\varphi(x)$ and $\varphi(y)$. Let us now examine the injectivity of $\nabla_x \tilde{c}(\tilde{x}, \cdot)$ for a fixed $\tilde{x} = \binom{x_H}{n_x}$. For any $\tilde{y} = \binom{y_H}{n_y}$:

$$\nabla_x \tilde{c}(\tilde{x}, \tilde{y}) = (w, t) \iff \begin{cases} w = 2(|y_H|^2 + n_y)x_H + \tilde{\Sigma} y_H \\ t = |y_H|^2 + n_y \end{cases} \iff \begin{cases} y_H = \tilde{\Sigma}^{-1}(w - 2t x_H) \\ n_y = t - |y_H|^2 \end{cases}$$

hence $\tilde{c}$ satisfies the twist condition.

Now, the same as in Example 2.1 applies, but this time with $E_0 = \mathbb{R}^h \times \mathbb{R}^{n-h}$, $E = E_0 \backslash (\mathbb{R}^h \times \{0\})$, $B_0 = \mathbb{R}^h \times \mathbb{R}$ and $F = S^{n-h-1} = \{x \in E_0 \mid |x_\perp| = 1\}$. Is then ensured the existence of a structured

Monge map between μ and ν for the cost c: it decomposes for almost all $x \in \mathbb{R}^n$ as a Monge map on the basis $B_0 = \mathbb{R}^{h+1}$ obtained as the gradient of a $\tilde{c}$-convex function $f : \mathbb{R}^{h+1} \to \mathbb{R}$ (via the $\tilde{c}$-exponential map on $\mathbb{R}^{h+1}$) and a Monge map on each fiber $F = S^{n-h-1}$, also built from gradients of convex functions $h_{(x_H, |x_\perp|^2)} : S^{n-h-1} \to \mathbb{R}$ (via the exponential map on the sphere); hence the result. Last, note that the case where $M^\star = 0$ has not been explictly treated. In this case, the cost is simply $c(x, y) = -|x|^2|y|^2 = \tilde{c}(n_x, n_y)$ and the strategy above directly applies. □

2.4 Complementary study of the quadratic cost in dimension 1

Recalling that the (GW-IP) problem is invariant by translation, we assume that measures μ and ν below are centered. In the one-dimensional case $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$, the linearized GW problem (1.11) reads, with $\pi^\star$ an optimal correspondence plan:

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} (-x^2y^2 - 4mxy)\, \mathrm{d}\pi(x, y)\,, \quad \text{where } m = \int_{\mathcal{X}\times\mathcal{Y}} x'y'\, \mathrm{d}\pi^\star(x', y')\,, \tag{2.6}$$

and for any plan $\pi \in \Pi(\mu, \nu)$ (not necessarily optimal), we denote by $m(\pi) = \int xy\, \mathrm{d}\pi(x, y)$ what we call the *correlation* of π. The associated OT cost function $c_m(x, y) = -x^2y^2 - 4mxy$ only satisfies the subtwist condition when $m \neq 0$ and the 2-twist condition when $m = 0$, which does not allow to conclude on the deterministic structure of optimal correspondence plans. However, in the one-dimensional case one has at their disposal the useful Proposition 1.9 on (Submod). The linearized quadratic GW cost with parameter $m \geq 0$ is submodular on the region $S = \{(x, y) \mid xy \geq -m\}$ and supermodular elsewhere (see Fig. 2.2 for an illustration); so we cannot directly apply this proposition. Still, it is reasonable to expect that optimal correspondence plans exhibit a monotone non-decreasing structure on S (written ↗ in Fig. 2.2) and a monotone non-increasing one elsewhere (written ↘), and we can actually leverage this type of property to obtain the optimality of the monotone rearrangements in some particular cases (see Sec. 2.4.3).

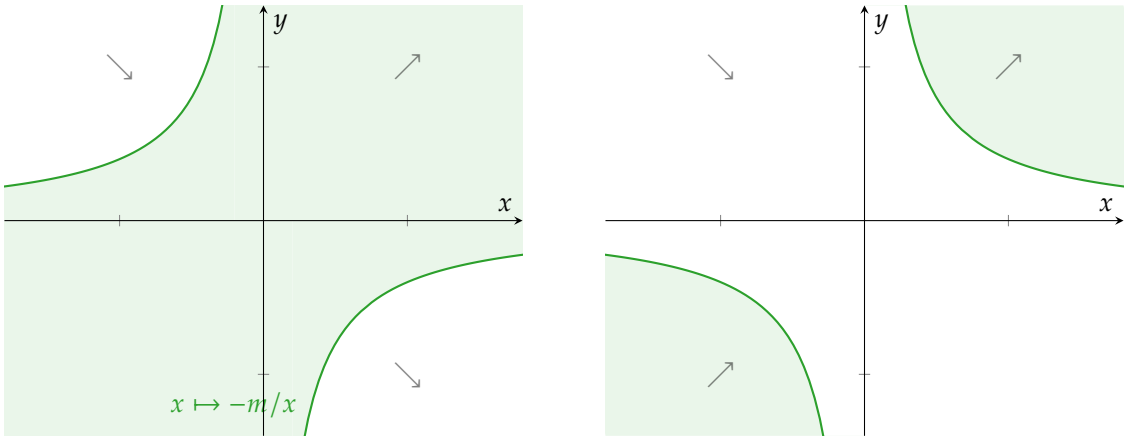


Figure 2.2: Submodularity region S (↗, in light green) and supermodularity region $\bar{S}$ (↘) for the linearized quadratic GW cost with parameter $m > 0$ **(Left)** or $m < 0$ **(Right)**.

In view of the form of the regions of modularity in our particular case, we can state:

Proposition 2.15. Let $m \geq 0$, $S = \{(x, y) \mid xy \geq -m\}$ and denote by $\pi^\star$ an optimal transportation plan for the cost C_m. Then, $[\pi^\star]_{|S}$ (the plan restricted to the submodularity region) is monotone non-decreasing.

Proof. In this proof, we use the fact that if $\pi^\star$ is optimal, then it is also optimal when restricted on a domain between the corresponding marginals. In particular, the plan $[\pi^\star]_{|S}$ is necessarily optimal for the cost c_m between its marginals. Consider $(x_0, y_0), (x_1, y_1) \in \mathrm{supp}([\pi^\star]_{|S})$ such that $x_0 < x_1$ and $y_0 > y_1$; this condition implies that the rectangle R defined by the four coordinates x_0, x_1, y_0, y_1 is contained in S. As a consequence, the submodularity of the cost can be applied on R to prove that $[\pi^\star]_{|R}$ is monotone non-decreasing, contradicting the configuration. □

Before going into further details on our complementary study, we recall the discrete formulation of (KP) in dimension one. Given two sets $\{x_1, \dots, x_N\}$ and $\{y_1, \dots, y_M\}$ of $\mathbb{R}$ and two probability vectors a and b, the (K̂P) problem between the discrete measures $\mu = \sum_{i=1}^N a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^M b_j \delta_{y_j}$ reads

$$\min_{\pi \in U(a,b)} \langle C, \pi \rangle ,$$

where $C = (c(x_i, y_j))_{i,j}$ is the cost matrix and $\langle \cdot, \cdot \rangle$ is the Frobenius inner product. In the case of the linearized problem (2.6), we denote by $C_{\mathrm{GW}(m)}$ the cost matrix, that has coefficients $(C_{\mathrm{GW}(m)})_{i,j} = -x_i^2 y_j^2 - 4m x_i y_j$ with $m = \langle C_{xy}, \pi^\star \rangle$ and $(C_{xy})_{i,j} = x_i y_j$.

In the following sections, we study the optimality of the monotone non-decreasing and non-increasing rearrangements $\pi^\oplus_{\mathrm{mon}}$ and $\pi^\ominus_{\mathrm{mon}}$. It is worth noting that by submodularity of $x, y \mapsto -xy$, these two correspondence plans have respective correlations $m_{\min}$ and $m_{\max}$, where

$$\begin{cases} m_{\min} &= \min_\pi \langle C_{xy}, \pi \rangle \\ m_{\max} &= \max_\pi \langle C_{xy}, \pi \rangle \end{cases}, \quad \text{with } (C_{xy})_{i,j} = x_i y_j , \tag{2.7}$$

and that for any correspondence plan π, the value of its correlation $m(\pi)$ lies in the interval $[m_{\min}, m_{\max}]$. We provide in the following a complementary study of the quadratic cost in dimension one, namely

(i) a procedure to find counter-examples to the optimality of the monotone rearrangements;
(ii) empirical evidence for the tightness of Theorem 2.14;
(iii) a proof of the instability of having a monotone rearrangement as an optimal correspondence plan;
(iv) a new result on the optimality of the monotone rearrangements when the measures are composed of two distant parts.

All experiments are reproducible and the code can be found on GitHub[3].

2.4.1 Adversarial computation of non-monotone optimal correspondence plans

Theorem 4.1.1 of [Vay20] claims that in the discrete case in dimension 1 with $N = M$ and $a = b = \mathbb{1}_N$, the optimal solution of (QAP) is either the monotone non-decreasing rearrangement $\pi^\oplus_{\mathrm{mon}}$ or the monotone non-increasing one $\pi^\ominus_{\mathrm{mon}}$ (or equivalently the identity $\sigma(i) = i$ or the anti-identity $\sigma(i) = N + 1 - i$); which seems to be the case with a high probability empirically when generating random discrete measures. While this claim is true for $N = 1, 2$ and 3, a counter-example for $N \geq 7$ points has recently been exhibited in [BHS22]. We further propose a procedure to automatically obtain additional counter-examples, demonstrating empirically that such adversarial distributions occupy a non-negligible place in the space of empirical measures. We propose to move away from distributions of optimal plans $\pi^\oplus_{\mathrm{mon}}$ and $\pi^\ominus_{\mathrm{mon}}$ by performing a gradient descent over the space of empirical distributions with N points using an objective function that favors the strict sub-optimality of the monotone rearrangements; we now detail this procedure.

For $N \geq 1$, we consider the set of empirical distributions over $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times \mathbb{R}$ with N points and uniform mass, *i.e.* of the form $\pi = \frac{1}{N} \sum_{i=1}^N \delta_{(x_i, y_i)}$. Such plans π can be seen as the identity mapping between vectors $X = (x_1, \dots, x_N)$ and $Y = (y_1, \dots, y_N)$, and we therefore note $\pi = \mathrm{id}(X, Y)$. Denoting

[3] link of the code: https://github.com/theodumont/monge-gromov-wasserstein.

by c_{GW} the functional that takes a correspondence plan and returns its cost on the GW problem, we then define $\mathcal{F}$ on $\mathbb{R}^N \times \mathbb{R}^N$ by

$$\mathcal{F}(X,Y) \triangleq c_{\mathrm{GW}}(\pi) - \min\left\{c_{\mathrm{GW}}(\pi^{\oplus}_{\mathrm{mon}}),\, c_{\mathrm{GW}}(\pi^{\ominus}_{\mathrm{mon}})\right\},$$

$$\text{where} \quad \begin{cases} \pi = \mathrm{id}(X,Y) \\ \pi^{\oplus}_{\mathrm{mon}} \text{ and } \pi^{\ominus}_{\mathrm{mon}} \text{ are the monotone rearrangements between } X \text{ and } Y. \end{cases}$$

This quantifies how well the plan π performs when compared to the best of the two monotone rearrangements. We generate N points at random in $[0,1]^2$ and then perform a simple gradient descent over the positions of the points $(X,Y) = (x_i, y_i)_i$ following the objective

$$\min_{X,Y \in \mathbb{R}^N} \mathcal{F}(X,Y).$$

We include an early-stopping threshold t, since when $\mathcal{F}(\pi)$ becomes negative (*i.e.* we found an slightly adversarial example), the objective function often starts to decrease exponentially fast, exploiting the adversarial behaviour of the plan as much as it can. We found that choosing $t = -2$ gave good results in our experiments. The procedure can be found in Algorithm 1 below. We implemented it using PyTorch's autodiff [PGM+19] and used [BTBD20] to implement a differentiable sorting operator to compute the monotone rearrangements. Adversarial plans $\pi_f = \mathrm{id}(X_f, Y_f)$ obtained by Algorithm 1 are not *a priori* optimal for the GW cost between their marginals; but they have at least a better cost than the monotone rearrangements since $\mathcal{F}(X_f, Y_f) < 0$, proving the sub-optimality of the latter.

Algorithm 1 Simple gradient descent over the positions $(x_i)_i$ and $(y_i)_i$.

Parameters:

- N: number of points of the distributions
- N_{iter}: maximum number of iterations
- η: step size
- t: early stopping threshold

Algorithm:

$X \leftarrow N$ random values in $[0,1]$, then centered
$Y \leftarrow N$ random values in $[0,1]$, then centered
for $i \in \{1, \dots, N_{\mathrm{iter}}\}$ **do**
  $\pi^{\oplus}_{\mathrm{mon}} \leftarrow$ `id(sort(`X`),sort(`Y`))` ▷ `id` is the identity mapping
  $\pi^{\ominus}_{\mathrm{mon}} \leftarrow$ `id(sort(`X`),reverse(sort(`Y`)))`
  $\pi \leftarrow$ `id(`X`,`Y`)`
  $\mathcal{F}(X,Y) \leftarrow$ `GW(`π`) - min(GW(`$\pi^{\oplus}_{\mathrm{mon}}$`),GW(`$\pi^{\ominus}_{\mathrm{mon}}$`))` ▷ `GW` computes c_{GW}
  if $\mathcal{F}(X,Y) < t$ **then** stop ▷ early stopping
  $(X,Y) \leftarrow (X,Y) - \eta \nabla \mathcal{F}(X,Y)$ ▷ step of gradient descent
end for
return $\pi_f =$ `id(`X`,`Y`)`

Output: a plan π_f with better GW cost than $\pi^{\oplus}_{\mathrm{mon}}$ and $\pi^{\ominus}_{\mathrm{mon}}$

On Figure 2.3 is displayed an example of adversarial plans obtained following this procedure. It can be observed that during the descent, the plan π has difficulties getting out of what seems to be a saddle point consisting in being the monotone rearrangements between its marginals. Moreover, it is worth noting that the marginals of our typical adversarial plans, such as the one of Fig. 2.3, are often similar to the counter-example proposed in [BHS22], where both measures have their mass concentrated near zero, except for one outlier for μ and two for ν, one on each tail.

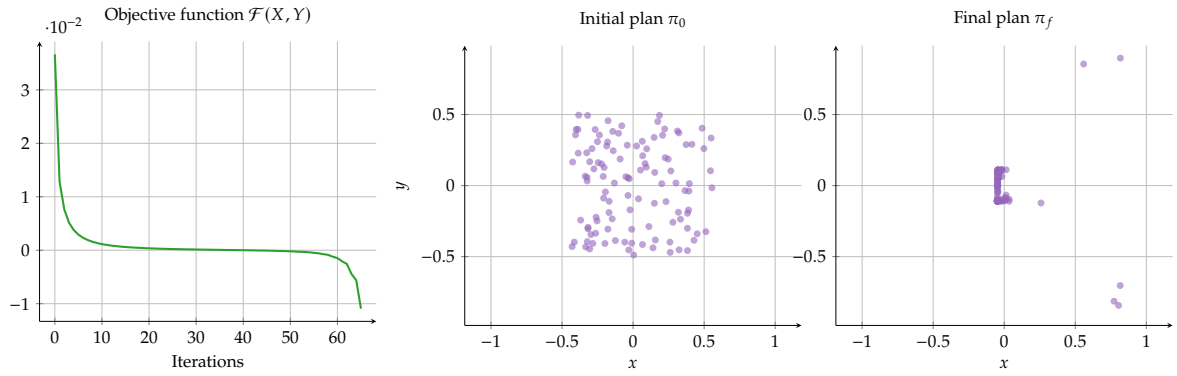


Figure 2.3: Gradient descent results with parameters $N = 122$, $\eta = 26$, $t = -2$. **(Left)** Evolution of the objective function $\mathcal{F}$. **(Center)** Initial plan π_0, generated at random. **(Right)** Final plan π_f (iter. 66).

Furthermore, examining the optimal correspondence plan for these adversarial examples allows to exhibit cases where it is not a map, providing empirical evidence for the following conjecture:

Conjecture 2.16. Theorem 2.14 is tight, *i.e.* there exists μ and ν for which optimal correspondence plans for (GW-Q) are not maps but rather a union of two graphs (either that of two maps or that of a map and an anti-map); and this even if μ has a density, classical OT assumption for the existence of an optimal transport map.

In order to approximate numerically the case of a measure which has density w.r.t. the Lebesgue measure, we convolve our distributions $\mu = (X_f, \mathbb{1}_N)$ and $\nu = (Y_f, \mathbb{1}_N)$ with a Gaussian of standard deviation σ and represent it in Eulerian coordinates; that is we evaluate the closed form density on a fine enough grid. When σ is large, the optimal correspondence plan for GW is probably induced by a monotone map, as it is the case very frequently empirically; on the contrary, if σ is sufficiently small, *i.e.* when the distributions are very close to their sum-of-Diracs discrete analogous, the optimal correspondence plan should not be a monotone map, by construction of μ and ν.

Remark 2.4. Because of the adversarial nature of π_f for the sub-optimality of $\pi^{\oplus}_{\text{mon}}$ and $\pi^{\ominus}_{\text{mon}}$, we know that when σ is sufficiently small, the optimal correspondence plan is not a monotone rearrangement. Still, it could be the case that this optimal plan is a map, but not a monotone one, and there is *a priori* no reason to believe that π_f will agree with Conjecture 2.16. Surprisingly, it sometimes does, as numerical experiments below suggest.

In order to find the optimal correspondence plan $\pi^\star$ between μ and ν, we leverage the fact that $\pi^\star$ is a solution of its associated linearized problem. Therefore, a minimizer of the GW functional is given by

$$\arg\min \left\{ \mathrm{GW}(\pi^\star_m) \,\middle|\, \pi^\star_m \in \underset{\pi \in U(a,b)}{\arg\min} \langle C_{\mathrm{GW}(m)}, \pi \rangle \,,\, m \in [m_{\min}, m_{\max}] \right\}, \tag{2.8}$$

where $(C_{\mathrm{GW}(m)})_{i,j} = -x_i^2 y_j^2 - 4m x_i y_j$. We therefore compute both $m_{\min}$ and $m_{\max}$ by solving the linear programs in (2.7), discretize the interval $[m_{\min}, m_{\max}]$ with $N_{\Delta m}$ points, and solve the corresponding linear optimization problem for every value of the parameter m and evaluate the GW cost on each optimal plan for the given parameter m. We then check if the optimal plan exhibits a bimap or a map/anti-map structure. The procedure is described in Algorithm 2.

We display the results on Fig. 2.4, where we plot the optimal correspondence plan $\pi^\star$ in two cases:

(a) starting from an adversarial plan with both marginals convolved as to simulate densities;
(b) starting from an adversarial plan with only the first marginal convolved and the second marginal being a sum-of-Diracs measure.

Algorithm 2 Generating bimaps from adversarial examples.

Input: an adversarial plan $\pi_f = \mathrm{id}(X_f, Y_f)$ obtained from Algorithm 1

Parameters:
- σ: standard deviation of convolution
- $N_{\Delta x}$: discretization precision
- $N_{\Delta m}$: discretization precision of the interval $[m_{\min}, m_{\max}]$

Algorithm:

$a \leftarrow \texttt{convolution}(X_f, \sigma, N_{\Delta x})$
$b \leftarrow \texttt{convolution}(Y_f, \sigma, N_{\Delta x})$ ▷ optional (see below)
$m_{\min} \leftarrow \min_{\pi \in U(a,b)} \langle C_{xy}, \pi \rangle$ ▷ solve linear programs
$m_{\max} \leftarrow \max_{\pi \in U(a,b)} \langle C_{xy}, \pi \rangle$
$\texttt{scores} \leftarrow \texttt{[]}$
for $m \in \{m_{\min}, \dots, m_{\max}\}$ **do** ▷ with $N_{\Delta m}$ points
  $\pi^\star_m \leftarrow \arg\min_{\pi \in U(a,b)} \langle C_{\mathrm{GW}(m)}, \pi \rangle$ ▷ solve linear program
  append $\mathrm{GW}(\pi^\star_m)$ to $\texttt{scores}$
end for
$\pi^\star \leftarrow \arg\max_\pi \texttt{scores}$ ▷ take best plan for GW
$b \leftarrow$ "$\pi^\star$ is a bimap"
return $\pi^\star, b$

Outputs:
- $\pi^\star$: optimal plan for GW
- b: boolean asserting if $\pi^\star$ is a bimap

To facilitate the reading, we draw a blue pixel at a location x on the discretized x-axis (resp. y on the y-axis) each time x (resp. y) has two (disjoint) images (resp. antecedents), making $\pi^\star$ a bimap (resp. a bi-anti-map), or the union of a graph and an anti-graph. In both cases, we observe that $\pi^\star$ is not a map but a bimap instead, similarly to [CMN10, Sec. 4.5]. Note that in case (b), ν being atomic, there cannot be a map from ν to μ, so in both (a) and (b) we numerically exhibit an instance where there is *a priori* no map from neither μ to ν nor ν to μ. We also plot the submodularity regions of the linearized GW cost function with parameter $m(\pi^\star)$ as an overlay and we observe that when the optimal plan gives mass to a region where the cost is submodular (resp. supermodular), is has a monotone non-decreasing (resp. non-increasing) behaviour in this region.

Remark 2.5. Although the region on which the optimal plan $\pi^\star$ is a bimap is of small size on Fig. 2.4 right, we cannot expect better due to the form of the adversarial example π_f. Indeed, the bimap behaviour is governed by the outliers of the distributions (see Fig. 2.3), as points in the right tail of μ are encouraged to split in half between points in the right and left tails of ν. As the bimap region only spans the outlier region, it stays of small size when μ and ν have only few outliers.

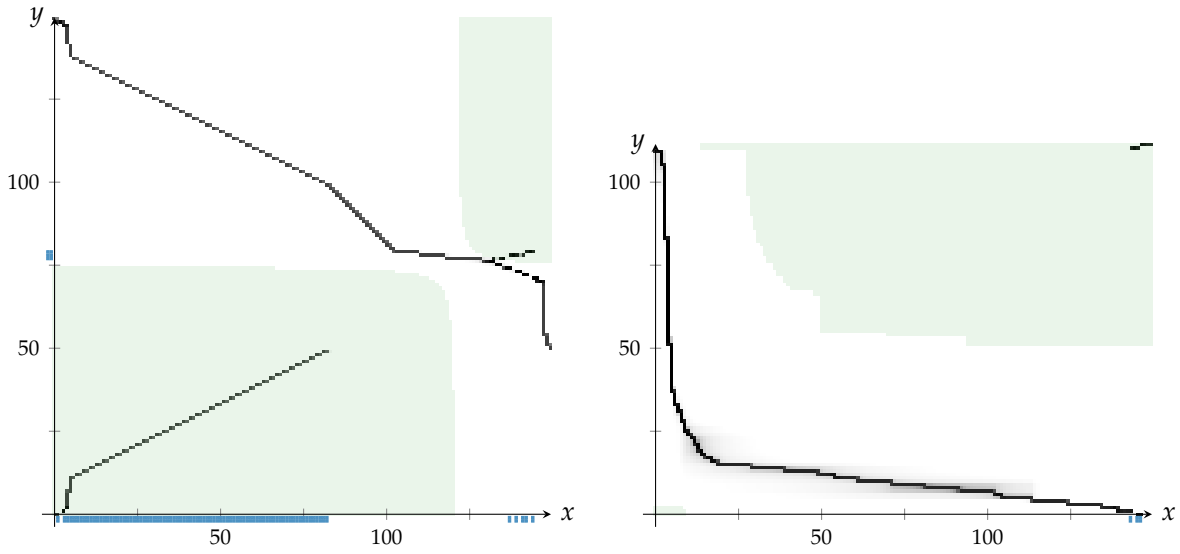


Figure 2.4: Optimal correspondence plan (in log scale) obtained with our procedure, starting either from a plan with both marginals convolved **(Left)** or with only the first marginal convolved **(Right)** ; bimap and anti-bimap coordinates (blue); submodularity regions (light green). Parameters: $\sigma = 5.10^{-3}$, $N_{\Delta x} = 150$, $N_{\Delta m} = 2000$.

2.4.2 Empirical instability of the optimality of monotone rearrangements

The above study demonstrates that there exist probability measures μ and ν for which property

$$P(\mu,\nu): \quad \pi^{\oplus}_{\mathrm{mon}} \text{ or } \pi^{\ominus}_{\mathrm{mon}} \text{ is an optimal correspondence plan between } \mu \text{ and } \nu$$

does not hold. However, as it is very likely in practice when generating empirical distributions at random, one could ask if property P is at least *stable*, *i.e.* if when we have μ_0 and ν_0 satisfying $P(\mu_0,\nu_0)$ there is a small ball around μ_0 and ν_0 (for a given distance, say Wasserstein 2) inside which property P remains valid. A negative answer to this—besides, in the symmetric case—is given by the counter-example by [BHS22] with an increasing number of points:

Proposition 2.17. There exists two *symmetric* measures μ,ν on $\mathbb{R}$ and sequences $(\mu_n)_n$, $(\nu_n)_n$ that weakly* converge to μ,ν such that optimal plans π_n between μ_n and ν_n are never supported by a monotone map.

Proof. We consider $\mu=\nu=\delta_0$ and the discrete measures $\mu_n=\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu_n=\frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ defined as follows for $n\geq 7$:

$$x_i \triangleq \begin{cases} -1 & \text{for } i=1 \\ (i-\frac{n+1}{2})\frac{1}{n^2} & \text{for } i=2,\dots,n-1 \\ 1 & \text{for } i=n \end{cases} \quad \text{and} \quad y_i \triangleq \begin{cases} -1 & \text{for } i=1 \\ -1+\frac{1}{n^2} & \text{for } i=2 \\ (i-2)\frac{1}{n^2} & \text{for } i=3,\dots,n \end{cases}$$

which is simply the counter-example from [BHS22] with n points and $\varepsilon_n=1/n^2$. Since $n\geq 7$, $\varepsilon_n<2/(n-3)$ and the identity or anti-identity mappings are not optimal between μ_n and ν_n. By direct computation,

$$\mathrm{W}_2^2(\delta_0,\mu_n)=O(2/n+\varepsilon_n^2 n^2)\xrightarrow{n\to\infty}0\,,$$

and the exact same goes for $\nu=\delta_0$ and ν_n. □

One can actually obtain non-degenerate (although not symmetric anymore) examples of such measures μ,ν. We start from the counter-example given in [BHS22] with $N=7$ points and $\varepsilon=10^{-2}$, that we convolve with a Gaussian of standard deviation σ as before. We then plot as a function of $m\in[m_{\min},m_{\max}]$ the (true) GW cost of a plan $\pi^\star_m$, optimal for the linearized GW problem $\pi^\star_m\in\arg\min_\pi\langle C_{\mathrm{GW}(m)},\pi\rangle$. The minimum values of this graph are attained by the correlations of optimal correspondence plans, as explained in Section 2.4.1. Hence if σ is small, this optimal plan is not a monotone rearrangement by construction and the minimum are not located on the boundary of the domain. On the contrary, when σ is large, the convolved measures stop being adversarial and the monotone rearrangements start being optimal again. In order to study the phase transition, we plot on Figure 2.5 the landscape of $m\mapsto \mathrm{GW}(\pi^\star_m)$ while gradually increasing the value of σ.

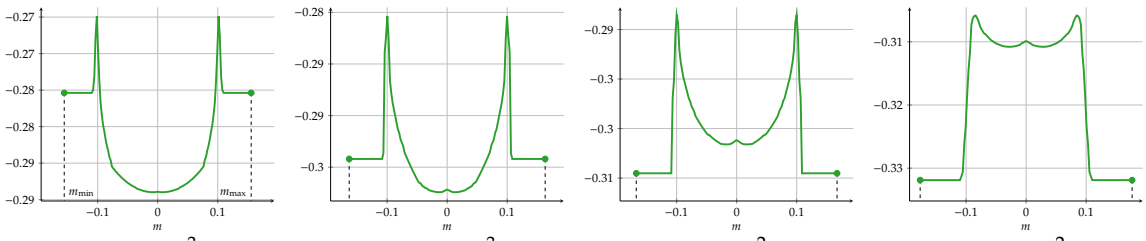


Figure 2.5: Evolution of the graph of $m\mapsto \mathrm{GW}(\pi^\star_m)$ when varying σ on the counter-example of [BHS22] with $N=7$ points and $\varepsilon=10^{-2}$. Parameters: $N_{\Delta x}=100$, $N_{\Delta m}=150$.

Looking at Fig. 2.5, it is worth noting that there is an incentive for plans of correlation close to $m_{\min}$ or $m_{\max}$ to be the monotone rearrangements, as the horizontal portions of the plot suggest. More

importantly, it can be observed that when $\sigma = \sigma_3$ or σ_4, the monotone rearrangements are optimal, as their correlations realize the minimum of $m \mapsto \mathrm{GW}(\pi_m^\star)$; unlike for σ_1 and σ_2, for which the minimum value of the plot is located near zero. Hence there exists a $\sigma_0 \in (\sigma_2, \sigma_3)$ for which the convolved measures have both $\pi^{\oplus}_{\mathrm{mon}}$, $\pi^{\ominus}_{\mathrm{mon}}$ and another π_0 as optimal correspondence plans; it is direct that property P does not hold in the neighbourhood of these specific measures μ_0 and ν_0.

2.4.3 A positive result for measures with two components

Disclaimer: This result is not mine but my supervisors'. It was discovered following discussions during which we tried to construct optimal plans that are double bimaps ($x \to y$ and $y \to x$). I state it here for completeness and because it is a new positive result on the optimality of monotone rearrangement in dimension 1.

In the following, μ_1, μ_2, ν_1 and ν_2 are four probability measures supported on a compact interval $A \subset \mathbb{R}$. Denote $\Delta = \mathrm{diam}(A)$, and fix $t \in (0,1)$ and $K > \Delta$. Let $\tau_K : x \mapsto x + K$ denote the translation by K, and $A + K = \tau_K(A) = \{x + K \mid x \in A\}$. Now, introduce the measures

$$\mu = (1-t)\mu_1 + t\tau_{K*}\mu_2 \quad \text{and} \quad \nu = (1-t)\nu_1 + t\tau_{K*}\nu_2\,. \tag{2.9}$$

Note that μ_1 and $\tau_{K*}\mu_2$ (resp. ν_1 and $\tau_{K*}\nu_2$) have disjoint supports. We want to prove the following:

Proposition 2.18. For K large enough, the unique optimal plan for the quadratic cost between μ and ν is given by one of the two monotone maps (non-decreasing or non-increasing).

Remark 2.6. The hypothesis of the theorem illustrates that monotone maps are favored when μ and ν both contain a single or more outliers. The proof of the theorem actually shows the importance of long range correspondences or global effect over the local correspondences on the plan. In other words, even though locally monotone maps may not be optimal, global correspondences favor them. Moreover, these global correspondences have proportionally more weight in the GW functional since the cost is the squared difference of the squared distances. In conclusion, pair of points which are at long distances tend to be put in correspondence. In turn, this correspondence, as shown in the proof, favors monotone matchings. Although non-quantitative, this argument gives some insight on the fact that a monotone map is often optimal.

We first prove the following lemma:

Lemma 2.19. In the setting described above, there exists $K_0 > 0$ such that if $K \geq K_0$, every π optimal plan for $\mathrm{GW}(\mu, \nu)$ can be decomposed as $\pi = \pi_1 + \pi_2$, where either:

1. π_1 is supported on $A \times A$ and π_2 on $(A+K) \times (A+K)$ (that is, we separately transport μ_1 to ν_1 and $\tau_{K*}\mu_2$ to $\tau_{K*}\nu_2$), or
2. π_1 is supported on $A \times (A+K)$ and π_2 on $A \times (A+K)$ (that is, we transport μ_1 to $\tau_{K*}\nu_2$ and μ_2 to $\tau_{K*}\nu_1$).

Furthermore, whenever $t \neq \frac{1}{2}$, only the first point can occur.

Proof. Consider first the case $t = \frac{1}{2}$. To shorten the notations, we introduce the notations $A_1 = A$ and $A_2 = A + K$. We can now decompose any plan π as $\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22}$ where for instance π_{12} denotes the restriction of the plan π to the product $A_1 \times A_2$. Let us also denote by r the mass of π_{12}, one has $0 \leq r \leq 1/2$ and by symmetry, one can choose that $r \leq 1/4$, otherwise we exchange A_1 and A_2 for the second measure since the cost is invariant to isometries. Remark that, due to marginal constraints, the total mass of π_{11} and π_{22} is $1/2 - r$ and the mass of π_{21} is r. Therefore, it is possible to consider a coupling plan $\tilde{\pi}_{11}$ between the first marginal of π_{12} and the second marginal of π_{21}, and similarly, let $\tilde{\pi}_{22}$ be a coupling plan between the first marginal of π_{21} and the second marginal of π_{12}. We then

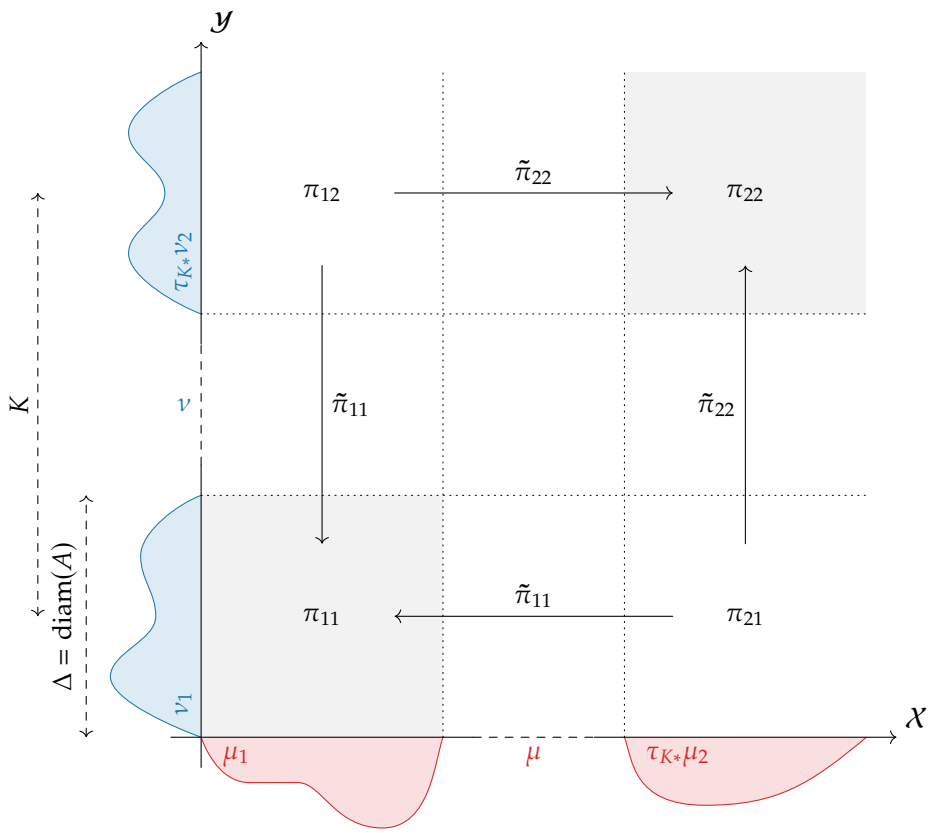


Figure 2.6: Visual sketch of the proof of Lemma 2.19.

define a competitor plan $\tilde{\pi} = \pi_{11} + \tilde{\pi}_{11} + \pi_{22} + \tilde{\pi}_{22}$. The first step is to get a lower bound on the term $\mathrm{GW}(\pi, \pi)$. Slightly overloading the notations, we introduce

$$\mathrm{GW}(\pi, \gamma) = \int c \, \mathrm{d}\pi \otimes \gamma \,. \tag{2.10}$$

We expand GW by bilinearity

$$\mathrm{GW}(\pi, \pi) = \sum_{i,j,i',j'} \mathrm{GW}(\pi_{ij}, \pi_{i'j'}) = \sum_{i,j} \mathrm{GW}(\pi_{ii}, \pi_{jj}) + R \,,$$

where R is the remainder that contains 12 terms from which one can identify two types. 8 terms are of the type $\mathrm{GW}(\pi_{12}, \pi_{11}) \geq r(1/2 - r)(K^2 - \Delta^2)^2$. Indeed, one compares pairs of points (x, x') and (y, y') for $(x, y) \in A_1 \times A_1$ and $(x', y') \in A_1 \times A_2$, therefore $(x - x')^2$ is upper bounded by Δ^2 and $(y - y')^2$ lower bounded by K^2 and the bound above follows after integration against the corresponding measures. The second type is $\mathrm{GW}(\pi_{12}, \pi_{21}) \geq 0$, there are 4 of such terms. We thus have

$$R \geq 8r(1/2 - r)(K^2 - \Delta^2)^2 \,.$$

We now upper-bound the competitor. Similarly, one has

$$\mathrm{GW}(\tilde{\pi}, \tilde{\pi}) = \sum_{i,j} \mathrm{GW}(\pi_{ii}, \pi_{jj}) + \tilde{R}$$

where $\tilde{R} = 2\,\mathrm{GW}(\tilde{\pi}_{11}, \pi_{22} + \tilde{\pi}_{22}) + 2\,\mathrm{GW}(\tilde{\pi}_{22}, \pi_{11} + \tilde{\pi}_{11}) + 2\,\mathrm{GW}(\pi_{11}, \tilde{\pi}_{11}) + 2\,\mathrm{GW}(\pi_{22}, \tilde{\pi}_{22})$. The two last terms can be upper bounded by $2r(1/2 - r)\Delta^2$. Indeed, one compares distance squared of couples of points in A_1 to couple of points in A_1, so it is upper bounded by Δ^2. Again by elementary inequalities (see Fig. 2.6), the two first terms can be upper bounded by $r(2K\Delta + \Delta^2)^2$. Note that the total mass of the plan $\pi_{11} + \tilde{\pi}_{11}$ is $1/2$ which explains why $(1/2 - r)$ does not appear. Therefore, the difference between the two values of GW is

$$\mathrm{GW}(\pi, \pi) - \mathrm{GW}(\tilde{\pi}, \tilde{\pi}) \geq r\left(8(1/2 - r)(K^2 - \Delta^2)^2 - 4(1/2 - r)\Delta^2 - 2(2K\Delta + \Delta^2)^2\right). \tag{2.11}$$

Then, since $1/2 - r \geq 1/4$ the limit in K of the polynomial function on the r.h.s. of Eq. (2.11) is $+\infty$ uniformly in $r \in [0, \frac{1}{4}]$, and the result follows; there exists $K > 0$ such that the polynomial function above is nonnegative, for instance $\max(0, K_0)$ where K_0 is the largest root.

The proof in the case $t > 1/2$ (the other is symmetric) is even simpler since $t - r > t - 1/2$ and consequently, there is no choice in the matching of the two measures; it is determined by the corresponding masses. One can directly apply the argument above. □

We now prove Proposition 2.18.

Proof of Proposition 2.18. Thanks to Lemma 2.19, we know that we can restrict to transportation plans $\pi = \pi_1 + \pi_2$ where, up to flipping ν, we can assume that π_1 is supported on $A \times A$ and π_2 on $(A + K) \times (A + K)$.[4]

Using again the bilinear form $\mathrm{GW}(\pi, \gamma)$ defined in (2.10), the objective values reached by any transport plan $\pi = \pi_1 + \pi_2$ actually decomposes as

$$\mathrm{GW}(\pi, \pi) = \mathrm{GW}(\pi_1, \pi_1) + 2\,\mathrm{GW}(\pi_1, \pi_2) + \mathrm{GW}(\pi_2, \pi_2).$$

Now, assume that we have found $\pi_2^\star$ optimal. Let us minimize in π_1 the resulting quadratic problem:

$$\min_{\pi_1} \mathrm{GW}(\pi_1, \pi_1) + 2\,\mathrm{GW}(\pi_1, \pi_2^\star).$$

We know that if $\pi_1^\star$ is a minimizer of this quantity, it must also be a solution of the *linear* problem

$$\min_{\pi_1} \mathrm{GW}(\pi_1, \pi_1^\star) + \mathrm{GW}(\pi_1, \pi_2^\star).$$

This minimization problem is exactly the optimal transportation problem for the cost

$$c(x, y) = \int_{A\times A} ((x - x')^2 - (y - y')^2)^2 \,\mathrm{d}\pi_1^\star(x', y') + 2\int_{(A+K)^2} ((x - x'')^2 - (y - y'')^2)^2 \,\mathrm{d}\pi_2^\star(x'', y'').$$

Now, using the relation $((x - x'')^2 - (y - y'')^2)^2 = ((x - y) - (x'' - y''))^2((x + y) - (x'' + y''))^2$, and that $\pi_2^\star$ is a transportation plan between $\tau_{K*}\mu_2$ and $\tau_{K*}\nu_2$ so that we can make a change of variable, observe that

$$c(x, y) = \int_{A\times A} ((x - x')^2 - (y - y')^2)^2 \,\mathrm{d}\pi_1^\star(x', y') + \int_{A\times A} ((x - y) - (x'' - y''))^2((x + y) - (x'' + y'' + 2K))^2 \,\mathrm{d}(\tau_{-K}, \tau_{-K})_*\pi_2^\star(x'', y'').$$

[4] Note: this is where the choice is made, as in the proof of Lemma 2.19, between the increasing and the non-increasing matchings. Using this convention, the non-decreasing monotone map is shown to be optimal.

Now, observe that $\partial_{xy} c(x, y)$ is a polynomial function in K, x, y whose dominant term in K is simply $-2K^2$, and recall that A is compact, so that this polynomial function is bounded in x, y. We conclude

$$\partial_{xy} c(x, y) = -2K^2 + O(K) < 0$$

for K large enough, for all $x, y \in A$.

The plan $\pi_1^\star$ is optimal for a submodular cost, and by Proposition 1.9 must be the non-decreasing matching between μ_1 and ν_1. By symmetry, so is $\pi_2^\star$. □

Conclusion

Summary

In this work, we showed the existence of deterministic optimal correspondence plans for the Gromov–Wasserstein problem in two settings: (i) with the inner product cost, where we showed that there always exists a Monge map, and (ii) with the quadratic cost, where we derived a condition on the rank of a certain matrix under which there exists either a map, a bimap, or a map/anti-map that is optimal for the GW problem. We also illustrated computationally in dimension 1 that the latter condition is tight, *i.e.* that there exists cases where the optimal plan is not a map. On a different note, we studied the optimality of the monotone non-decreasing and non-increasing plans for GW with quadratic cost in dimension 1, illustrating empirically that they are not always optimal (following [BHS22]) and that having these plans as optimal correspondence plans is not stable by small perturbations of μ and ν. We also provide a positive result for the optimality of the monotone plans when the measures are composed of two distant parts.

Discussion and future work

Existence of deterministic transport plans. Our general existence theorem (Theorem 2.5) naturally applies to both quadratic and inner product costs. It could be the case that it is sufficiently general to be applied to other costs functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$, for which the method that we used in this work could then prove the existence of deterministic transport plans.

Optimality of the monotone rearrangements. In dimension 1, [BHS22] gave a counter-example to the fact that the quadratic assignment problem is solved by a monotone rearrangement that needs at least $N = 7$ points. One could wonder if this property holds for $N \leq 6$: it is actually trivial for $N = 1, 2$, true for $N = 3$ by combinatorial arguments. This leaves the cases $N = 4, 5, 6$ open. Furthermore, the optimality of these plans can be observed empirically with a very high probability when generating distributions at random, and yet we dispose of no claim that could explain this. Ideally, we would like to give a condition on the distributions μ and ν under which one of $\pi^{\oplus}_{\text{mon}}$ and $\pi^{\ominus}_{\text{mon}}$ is optimal; so far, only the symmetric case from [Stu12] and our case of two-parts measures have been treated, and we believe that this result is much more general.

The non-optimality of $\pi^{\oplus}_{\text{mon}}$ and $\pi^{\ominus}_{\text{mon}}$ in the general case jeopardizes the well-posedness of the so-called *sliced Gromov–Wasserstein distance* [VFC+19, Vay20], that relies on the assumption that the GW distance can be efficiently computed in dimension 1. Designing a cost function $c_{\mathcal{X}} = c_{\mathcal{Y}} = c$ for which $\pi^{\oplus}_{\text{mon}}$ and $\pi^{\ominus}_{\text{mon}}$ are always optimal would allow this theory to be made valid with (supposedly) only a few tweaks.

Computational aspects. This work does not contain any computational concern, and our existence result does not give any insight on the complexity of the GW problems since it does not provide a close form for the map. In dimension 1 with the quadratic cost, we provide a method to exhibit an approximation of an optimal GW plan.

Acknowledgements

Sources of the figures

I would like to give the source (direct or only inspiration) of the figures in this work. The numbered limb system at the right of Fig. 1.3 (and therefore Fig. B.4) is inspired from [AKM11]; Fig. B.5 is taken from and Fig. 1.5 is inspired from T. Vayer's PhD thesis [Vay20]; the illustration on Fig. 1.7 of the GW distance in the discrete case is inspired from [PC19]; the illustration of c-convexity of Fig. B.2 is from [Vil08]; finally, the tikz code that I adapted to produce Fig. 1.2 is a courtesy of L. Chizat. All other figures are my own.

Thanks

I would like to thank my two advisors François-Xavier Vialard and Théo Lacombe with whom I learned a lot about the beautiful field of optimal transport and more generally about mathematics, both on a theoretical and personal level, helping me to identify the areas of mathematics that I am most interested about and confirming my wish to continue in academic research. I also thank Gabriel Peyré for making me discover the computational aspects of this fascinating field of mathematics through his MVA course. Special thanks to Kayané, Julien and Siwan, who, through our mathematical discussions, made the temperature in the intern's room a little more bearable.

Appendix

Here in the appendix can be found additional definitions (Appendix B) as well as the proofs of some of our claims that we did not put in the main corpus for the sake of clarity (Appendix A).

We advise the reader that is not familiar with the fields of measure theory and optimal transport to have a quick look at measure disintegration (Appendix B.1.2) and at c-convexity (Appendix B.2) before digging into Chapter 2, where we present our contributions. More generally, if some notions are unknown to the reader, it may be worthwhile to have a look here, where they may be defined.

Appendix A

Proofs of claims

Contents

A.1 Proofs of Lemmas 2.12 and 2.13: reparametrization of cost

Proof of Lemma 2.12. Remark that the continuity of ψ_1 and ψ_2 and their inverse ensures their measurability. We have the following equalities:

$$\begin{aligned}\operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)} \int \tilde{c}(\psi_1(x),\psi_2(y))\,\mathrm{d}\pi(x,y) &= \operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)} \int \tilde{c}(u,v)\,\mathrm{d}(\psi_1,\psi_2)_*\pi(u,v)\\ &= (\psi_1^{-1},\psi_2^{-1})_* \operatorname*{arg\,min}_{\tilde{\pi}\in\Pi(\psi_{1*}\mu,\psi_{2*}\nu)} \int \tilde{c}(u,v)\,\mathrm{d}\tilde{\pi}(u,v)\end{aligned}$$

since the mapping $(\psi_1^{-1},\psi_2^{-1})_*$ is a one-to-one correspondence from $\Pi(\psi_{1*}\mu,\psi_{2*}\nu)$ to $\Pi(\mu,\nu)$ by bijectivity of ψ_1 and ψ_2. This bijectivity ensures that any optimal deterministic transport plan $\tilde{\pi}^\star$ between $\psi_{1*}\mu$ and $\psi_{2*}\nu$ induces an optimal deterministic transport plan $\pi^\star$ between μ and ν, and *vice versa*. Writing $\tilde{\pi}^\star = (\mathrm{id},T)_*(\psi_{1*}\mu)$, this plan $\pi^\star$ is given by

$$\begin{aligned}\pi^\star &= (\psi_1^{-1},\psi_2^{-1})_*\tilde{\pi}^\star\\ &= (\psi_1^{-1},\psi_2^{-1})_*(\mathrm{id},T)_*\psi_{1*}\mu\\ &= (\mathrm{id},\psi_2^{-1}\circ T\circ\psi_1)_*\mu\,.\end{aligned}$$

□

Proof of Lemma 2.13. As $\psi_{1*}\mu$ has a density w.r.t. the Lebesgue measure since ψ_1 is a diffeomorphism and $\psi_{1*}\mu$ and $\psi_{2*}\nu$ have compact support, Brenier's theorem states that there exists a unique optimal transport plan between $\psi_{1*}\mu$ and $\psi_{2*}\nu$ and that it is induced by a map $\nabla\varphi$, where φ is a convex function. Using Lemma 2.12 then gives the result. □

Remark A.1 (Discussion on the hypothesis of Lemma 2.13)**.** In the proof of Lemma 2.13, we only needed (i) ψ_1, ψ_2 and their inverse to be measurable, (ii) $\psi_{1*}\mu$ to have a density w.r.t. Lebesgue, and (iii) $\psi_{1*}\mu$ and $\psi_{2*}\nu$ to have compact support. Imposing ψ_1 to be a diffeomorphism and ψ_2 to be a homeomorphism ensures both (i) and (ii) and is natural to expect.

A.2 Proofs of Tabs. 2.1 and 2.2: twist conditions

In this section, we prove the results summarized in the table below:

Table A.1: Twist conditions for the linearized GW costs (inner product and quadratic) in $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^d$ with $n \geq d$. Given an optimal plan π, we denote by M the matrix $\int y \otimes x \, \mathrm{d}\pi(x, y)$ of size $d \times n$.

	$\mathrm{rk}\, M$	$= d$	$= d-1$	$\leq d-2$
inner product cost $c(x,y) = \langle Mx,\, y\rangle$	twist	✓	·	·
	subtwist	✓	·	·
	2-twist	✓	·	·
	m-twist, $m \geq 3$	✓	·	·
	non-degeneracy	✓	·	·

	$\mathrm{rk}\, M$	$= d$	$= d-1$	$\leq d-2$
quadratic cost $c(x,y) = \lvert x\rvert^2\lvert y\rvert^2 + \langle Mx,\, y\rangle$	twist	·	·	·
	subtwist	✓	·	·
	2-twist	·	✓	·
	m-twist, $m \geq 3$	·	·	·
	non-degeneracy	$\sim$	·	·

Proof. Let us express the gradient and Hessian of both costs c_{ip} and c_{q}:

$$\begin{cases} c_{\mathrm{ip}}(x,y) = \langle Mx,\, y\rangle \\ \nabla_x c_{\mathrm{ip}}(x,y) = M^\top y \\ \nabla^2_{xy} c_{\mathrm{ip}}(x,y) = M^\top \end{cases} \quad \text{and} \quad \begin{cases} c_{\mathrm{q}}(x,y) = \lvert x\rvert^2\lvert y\rvert^2 + 4\langle Mx,\, y\rangle \\ \nabla_x c_{\mathrm{q}}(x,y) = 2\lvert y\rvert^2 x + 4M^\top y \\ \nabla^2_{xy} c_{\mathrm{q}}(x,y) = 4xy^\top + 4M^\top . \end{cases}$$

(Twist) and (Subtwist) conditions. The inner product case is direct. Since $\nabla_x c_{\mathrm{ip}}(x,y) = M^\top y$, c_{ip} satisfies the twist condition if and only if M is of full rank; in this case, the subtwist condition is also satisfied. If M is not of full rank, taking any $0 \neq y_1 \in \mathrm{Ker}(M^\top)$ and $y_2 = 0$ gives $\nabla_x c_{\mathrm{ip}}(x, y_1) - \nabla_x c_{\mathrm{ip}}(x, 0) = 0$ for all $x \in \mathcal{X}$, hence $x \mapsto c(x, y_1) - c(x, 0)$ has an infinite number of critical points and the subtwist condition is not satisfied.

For the quadratic cost, we fix $y_1 \neq y_2 \in \mathcal{Y}$. Any $x \in \mathcal{X}$ is a zero of $\nabla_x c_{\mathrm{q}}(x, y_1) - \nabla_x c_{\mathrm{q}}(x, y_2)$ if and only if

$$(\lvert y_1\rvert^2 - \lvert y_2\rvert^2)x = -M^\top(y_1 - y_2)\,. \tag{A.1}$$

Let us first suppose first that M is of full rank. If $\lvert y_1\rvert = \lvert y_2\rvert$, then (A.1) has no solution since $y_2 - y_1$ cannot be in $\mathrm{Ker}(M^\top)$. If $\lvert y_1\rvert \neq \lvert y_2\rvert$, then (A.1) has a unique solution $x^\star = -(\lvert y_1\rvert^2 - \lvert y_2\rvert^2)^{-1} M^\top(y_1 - y_2)$. Hence when M is of full rank, the subtwist condition is satisfied but the twist condition is not. Now let's suppose that M is not of full rank. We take any $0 \neq z \in \mathrm{Ker}(M^\top)$ and set $y_1 = \frac{1}{2}z$ and $y_2 = -\frac{1}{2}z$. Equation (A.1) in $x \in \mathcal{X}$ becomes $M^\top z = 0$, which is true for all x and therefore has an infinite number of solutions. Hence neither the twist nor subtwist conditions are satisfied in this case.

(m-twist) condition. Since the 1-twist condition is the twist condition, the inner product cost satisfies the m-twist condition for all $m \geq 2$ when M is of full rank. If M is not of full rank, then for any $y_0 \in \mathcal{Y}$, any $y \in y_0 + \mathrm{Ker}(M^\top)$ satisfies $M^\top y = M^\top y_0$, hence there is an infinite number of elements in the set $\lvert\{y \mid \nabla_x c_{\mathrm{q}}(x_0, y) = \nabla_x c_{\mathrm{q}}(x_0, y_0)\}\rvert$ for any x_0, y_0 and the m-twist condition is therefore not satisfied.

For the quadratic cost, let's consider $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$. Let $y \in \mathcal{Y}$ such that $\lvert y\rvert^2 x_0 + M^\top y = \lvert y_0\rvert^2 x_0 + M^\top y_0$. If M is of full rank, then [...]. Else, denoting by $v \in \mathbb{R}^d$ the right-hand side and by decomposing each vector $z \in \mathbb{R}^d$ into $z = (z_M, z_\perp) \in \mathbb{R}^r \times \mathbb{R}^{d-r}$ where $r \triangleq \mathrm{rk}(M)$,

$$\begin{cases} \lvert y\rvert^2 x_M + \tilde{M}^\top y_M & = v_M \\ \lvert y\rvert^2 x_\perp & = v_\perp\,. \end{cases}$$

If $x_\perp$ and $v_\perp$ are not colinear then it is absurd; else, since $x_\perp$ and $v_\perp$ are fixed, this means that $\lvert y\rvert^2$ is fixed and y lives on the $(d-1)$-dimensional sphere S^{d-1}. The first equation of the system above then

gives $y_M = (\tilde{M}^\top)^{-1}(v_M - |y|^2 x_M)$; hence y lives in the intersection of S^{d-1} and the $(d-r)$-dimensional affine subspace of vectors $z \in \mathbb{R}^d$ with fixed z_M, and this intersection is a $(d-r-1)$-dimensional affine sphere. Reciprocally, any y in this affine sphere will satisfy $|y|^2 x_0 + M^\top y = v$. Hence for $m \geq 2$, the m-twist condition is satisfied only if $d - r - 1 = 0$, *i.e.* $\mathrm{rk}(M) = d - 1$.

Non-degeneracy condition (Non-deg). The inner product case is direct, since $\det(\nabla^2_{xy} c_{\mathrm{ip}}(x,y)) = \det(M)$. Regarding the quadratic cost, since $\nabla^2_{xy} c_{\mathrm{q}}$ is (up to a factor and a transpose) $M + yx^\top$, *i.e.* the sum of the matrix M and of a rank 1 matrix:

- case 1: $\mathrm{rk}\, M \leq d-2$. Then $\mathrm{rk}(M + yx^\top) \leq d-1$ and $\det(M + xy^\top) = 0$ for all (x,y).
- case 2: $\mathrm{rk}\, M = d-1$. Then ???
- case 3: $\mathrm{rk}\, M = d$. Then by the matrix determinant lemma, $\det(M + xy^\top) = (1 + x^\top M^{-1} y)\det(M)$. Hence the determinant of $M + xy^\top$ is non-zero everywhere except on the set

$$\left\{(x,y) \mid x^\top M^{-1} y + 1 = 0\right\}.$$

Remark A.2. In dimension 1, this is indeed the parametric equation of the hyperbolae delimiting the submodularity region. □

A.3 Proof of Proposition 2.10: measurable selection of maps on manifolds

The proof is essentially an adaptation of the one of [FGM10], with additional care required due to the fact that we do not have access to a linear structure on the manifold M. It relies on measurability of set-valued maps (see [RW09, Ch. 5 and 14] and Appendix B.1.4 for a summary).

The crucial point regarding measurability is the following proposition.

Proposition A.1. The set

$$B_{n,k} = \{(u,x),\ T_u(x) \in A_{n,k}\}. \tag{A.2}$$

is measurable.

Its proof relies on a core lemma:

Lemma A.2. Let $F \subset M$ be a closed set. Then the set

$$B_F = \{(u,x),\ T_u(x) \in F\}$$

is measurable.

The key will be to identify this set as the domain of a measurable set-valued map, see Appendix B.1.4.

Proof of Lemma A.2. Observe that $B_F = \{(u,x),\ (\{x\}\times F) \cap \mathrm{gph}(T_u) \neq \varnothing\}$, where $\mathrm{gph}(T_u) = \{(x, T_u(x)),\ x \in M\}$ denotes the topological closure of the graph of the optimal transport map T_u that pushes μ_u onto ν_u. Let $S_1 : (u,x) \mapsto \{x\}\times F$ and $S_2 : (u,x) \mapsto \mathrm{gph}(T_u)$, so that $B_F = \mathrm{dom}(S)$, where $S(x) = S_1(x) \cap S_2(x)$. According to Proposition B.7, given that S_1 and S_2 are closed-valued, if they are measurable, so is S, and so is B_F as the domain of a measurable map. The measurability of these two maps can be easily adapted from the work of [FGM10], we give details for the sake of completeness.

Measurability of S_1: Let $O \subset M \times M$ be open. For any $z \in F$, if $x, z \in S_1^{-1}(O) = \{x,\ \{x\}\times F \cap O \neq \varnothing\}$, we have $\varepsilon > 0$ such that $B(x,\varepsilon) \times \{z\} \subset O$ (since O is open), and thus $B(x,\varepsilon) \times F \cap O \neq \varnothing$, proving that there is a neighborhood of x included in $S_1^{-1}(O)$ which is thus open (thus measurable), hence the measurability of S_1.

Measurability of S_2: Given that $u \mapsto (\mu_u, \nu_u)$ is measurable by assumption, and that measurability is preserved by composition, we want to show that (i) the map $S : (\mu,\nu) \mapsto \Pi(\mu,\nu)$ (the set of optimal transport plans between μ and ν for the quadratic cost on M) is measurable and (ii) the map

$U : \pi \in P(M)^2 \mapsto \operatorname{supp}\pi$ satisfies $U^{-1}(O)$ is open for any open set $O \subset P(M^2)$. From these two points, we get that $(U \circ S)^{-1}(O)$ is measurable, thus the measurability of S_2.

To get (i), observe first that S is closed-valued, so that it is sufficient to prove that $S^{-1}(C)$ is measurable for any closed set $C \subset P(M)^2$ according to Proposition B.6. Let $C \subset P(M^2)$ be closed. Then, $S^{-1}(C) = \{(\mu,\nu),\ \Pi^\star(\mu,\nu) \cap C \neq \varnothing\}$, and consider a sequence $(\mu_n,\nu_n)_n$ in $S^{-1}(C)$ that converges to (μ,ν) for the weak topology. Let $\pi_n \in \Pi^\star(\mu_n,\nu_n) \cap C$. According to [Vil08, Thm. 5.20], $(\pi_n)_n$ admits a weak limit π in $\Pi^\star(\mu,\nu)$, but also since C is closed, $\pi \in C$, so that $(\mu,\nu) \in S^{-1}(C)$ that is closed (hence measurable), proving the measurability of S.

(ii) simply follows from the fact that $U^{-1}(O) = \{\pi,\ \operatorname{supp}\pi \cap O \neq \varnothing\} = \{\pi,\ \pi(O) > 0\}$ that is open. Indeed, the Portmanteau theorem gives that if $\pi_n \to \pi$ (weakly) and $\pi_n(O) = 0$, then $0 = \liminf \pi_n(O) \geq \pi(O) \geq 0$, so $\pi(O) = 0$. The complementary set of $U^{-1}(O)$ is closed, that is $U^{-1}(O)$ is open. □

Proof of Proposition A.1. This follows from the fact that $A_{n,k}$ can be inner-approximated by a sequence of closed set $F_j \subset A_{n,k}$ and the fact that the B_{F_j} are measurable. □

We can now prove our main theorem. The proof is clearly inspired from the one of [FGM10], though it requires, in few places, careful adaptation.

Proof of Proposition 2.10. Recall that we assume that $M = \bigsqcup_n A_{n,k}$. For each n,k, select (in a measurable way) a $a_{n,k}$ in $A_{n,k}$. Then, define the map

$$T^{(k)} : (u,x) \mapsto a_{n,k}, \ \text{ such that } T_u(x) \in A_{n,k}. \tag{A.3}$$

This map is measurable. Indeed, the map $\Phi_k : (u,x) \mapsto A_{n,k}$ where $T_u(x) \in A_{n,k}$ is measurable, because $\Phi_k^{-1}(O) = \bigcup_n B_{n,k} \cap O$ that is measurable.

Now, for two maps $f, g : B \times M \to M$, let D_1 denotes the natural L_1 distance on M, that is

$$D_1(f,g) = \int_B \int_M d(f(u,x), g(u,x))\, \mathrm{d}\mu_u(x)\, \mathrm{d}m(u)\,. \tag{A.4}$$

This yields a complete metric space, and we can observe that $(T^{(k)})_k$ is a Cauchy sequence for this distance. Indeed, for $k \leq j$ two integers, recall that we assume that $(A_{n,j})_n$ is a refinement of $(A_{n,k})_n$, yielding

$$\begin{aligned}
D_1(T^{(k)}, T^{(j)}) &= \int_B \int_M d(T^{(k)}(u,x), T^{(j)}(u,x))\, \mathrm{d}\mu_u(x)\, \mathrm{d}m(u) \\
&= \int_B \int_M \sum_n \sum_{n' : A_{n',j} \subset A_{n,k}} 1_{B_{n',j}}(u,x) \cdot d(a_{n,k}, a_{n',j})\, \mathrm{d}\mu_u(x)\, \mathrm{d}m(u) \\
&= \int_B \int_M \sum_n \sum_{n' : A_{n',j} \subset A_{n,k}} d(a_{n,k}, a_{n',j})\, \mathrm{d}\nu_u(A_{n',j})\, \mathrm{d}m(u) \\
&\leq 2^{-k}
\end{aligned}$$

where we use that for all u, $\int_{x \in M} 1_{B_{n',j}}(u,x)\, \mathrm{d}\mu_u(x) = \nu_u(A_{n',j})$ by construction (recall that $(u,x) \in B_{n',j} \Leftrightarrow T_u(x) \in A_{n',j} \Leftrightarrow x \in \mu_u(T_u^{-1}(A_{n',j})) = T_u \# \mu_u(A_{n',j})$ and T_u transports μ_u onto ν_u), and then that the diameter of the partition $A_{n,k}$ is less than or equal to 2^{-k} and that ν_u and m are probability measures.

Now, let T denote the limit of $(T^{(k)})_k$ (that is measurable). It remains to show that $T(u,x) = T_u(x)$, m-a.e. This can be obtained by proving that

$$\int g(x) f(T(x,u))\, \mathrm{d}\mu_u(x) = \int g(x) f(T_u(x))\, \mathrm{d}\mu_u(x), \tag{A.5}$$

for any pair $f, g : M \to \mathbb{R}$ of bounded Lipschitz-continuous functions [VdV00, Lemma 2.24].

As in [FGM10], let $\|f\| = \sup_{x \neq y} \frac{|f(x)-f(y)|}{d(x,y)} + \sup_x |f(x)|$. The difference between these two terms can be bounded using the partition $(A_{n,k})_n$. We have for m-a.e. u:

$$\left| \int g(x) f(T_u(x)) \, \mathrm{d}\mu_u(x) - \int g(x) f(T(u,x)) \, \mathrm{d}\mu_u(x) \right| \tag{A.6}$$

$$\leq \left| \int g(x) f(T_u(x)) \, \mathrm{d}\mu_u(x) - \int g(x) f(T^{(k)}(u,x)) \, \mathrm{d}\mu_u(x) \right| + \|g\| \|f\| \int d\left(T^{(k)}(u,x), T(u,x)\right) \, \mathrm{d}\mu_u(x). \tag{A.7}$$

Since $T^{(k)} \to T$ in D_1, it implies that up to a subsequence, $\int_x d(T^{(k)}(u,x), T(u,x)) \, \mathrm{d}\mu_u(x) \to 0$ as $k \to \infty$ for m-a.e. u.

To treat the first term and show that it goes to 0 as $k \to \infty$ for a subset of B with full m-measure, we write for m-a.e. u:

$$\begin{aligned}
&\left| \int g(x) f(T_u(x)) \, \mathrm{d}\mu_u(x) - \int g(x) f(T^{(k)}(u,x)) \, \mathrm{d}\mu_u(x) \right| \\
\leq &\int |g(x)| \left| f(T_u(x)) - f(T^{(k)}(u,x)) \right| \, \mathrm{d}\mu_u(x) \\
\leq &\|g\| \|f\| \int d(T_u(x), T^{(k)}(u,x)) \, \mathrm{d}\mu_u(x) \\
\leq &\|g\| \|f\| 2^{-k} \sum_n \nu_u(A_{n,k}) \to 0 \,.
\end{aligned}$$

This concludes the proof. □

Appendix B

Additional notions

Contents

In this chapter, we report some mathematical definitions and results that will be useful for our study. The most important ones are those on measure theory (Appendix B.1) that we need for constructing optimal maps in Chapter 2—especially the disintegration theorem (Appendix B.1.2)—, and the definition of c-convexity (Appendix B.2), crucial in optimal transport theory.

B.1 Notions of measure theory

B.1.1 Basics

Definition B.1 (Atom)**.** *An* atom *is a set of positive measure which contains no nontrivial smaller measurable sets. A measure $\mu \in \mathcal{P}(\mathcal{X})$ is therefore said to be* atomless *if for any Borel set A with $\mu(A) > 0$ there exists a measurable subset B of A such that $\mu(A) > \mu(B) > 0$.*

Example B.1. In $\mathbb{R}^d$, an atomless measure is a measure that does not put mass on any point.

Definition B.2 (Support)**.** *The* support $\operatorname{supp}\mu \subset \mathcal{X}$ *of a measure $\mu \in \mathcal{P}(\mathcal{X})$ is the smallest closed set C such that $\mathcal{X} \setminus C$ has μ-measure zero.*

Definition B.3 (Absolute continuity)**.** *A measure $\mu \in \mathcal{P}(\mathcal{X})$ is* absolutely continuous with respect to a measure λ *(often the Lebesgue, Hausdorff or volume measures), written $\mu \ll \lambda$, if for every λ-measurable set A, $\lambda(A) = 0$ implies $\mu(A) = 0$.*

Theorem B.1 (Radon–Nikodym theorem)**.** In the same setting where $\mu \ll \lambda$, there exists a measurable function $\rho : X \to [0, \infty)$ such that for any measurable set $A \subset \mathcal{X}$,

$$\mu(A) = \int_A \rho \, \mathrm{d}\lambda \, .$$

The function ρ satisfying the above equality is uniquely defined up to a λ-null set. ρ is commonly written $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ and is called the *Radon–Nikodym derivative*. We say that μ has a *density* ρ with respect to λ.

Definition B.4 (Metric measure space [Mém11])**.** *A* metric measure (mm) space *is a triplet of the form $(\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X})$ where $(\mathcal{X}, d_\mathcal{X})$ is a compact metric space and $\mu_\mathcal{X}$ is a probability measure on $\mathcal{X}$ with full support, i.e.* $\mathrm{supp}(\mu_\mathcal{X}) = \mathcal{X}$.

Two mm-spaces $(\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X})$ and $(\mathcal{X}, d_\mathcal{X}, \mu_\mathcal{X})$ are said to be strongly isomorphic *if there exists an isometry (bijective and distance-preserving) $\varphi : \mathcal{X} \to \mathcal{Y}$ such that $\varphi_* \mu_\mathcal{X} = \mu_\mathcal{Y}$. They are said to be* weakly isomorphic *if there exists $(\mathcal{Z}, c_\mathcal{Z}, m)$ with $\mathrm{supp}(m) = \mathcal{Z}$ and "strong isomorphisms" $\varphi : \mathcal{Z} \to \mathcal{X}$ and $\psi : \mathcal{Z} \to \mathcal{Y}$ (with quotation marks since $c_\mathcal{X}$ and $c_\mathcal{Y}$ do not give $\mathcal{X}$ and $\mathcal{Y}$ a metric space structure).*

Definition B.5 (Measure restriction)**.** *Given a measure $\mu \in \mathcal{P}(\mathcal{X})$ and a set $A \subset \mathcal{X}$, the* restriction of μ to A*, written $\mu \llcorner A$, is defined by*

$$\text{for all Borel set } B, \quad (\mu \llcorner A)(B) = \mu(A \cap B) \, .$$

B.1.2 Measure disintegration

Definition B.6 (Measure disintegration)**.** *Let $\mathcal{X}$ and $\mathcal{Z}$ be two Radon spaces, $\mu \in P(\mathcal{X})$ and $\varphi : \mathcal{X} \to \mathcal{Z}$ a Borel-measurable function. A family of probability measures $\{\mu_u\}_{u \in \mathcal{Z}} \subset P(\mathcal{X})$ is a* disintegration *of μ by φ if:*

(i) the function $u \mapsto \mu_u$ is Borel-measurable;
(ii) μ_u lives on the fiber $\varphi^{-1}(\{u\})$: for $\varphi_*\mu$-a.e. $u \in \mathcal{Z}$,

$$\mu_u(\mathcal{X} \setminus \varphi^{-1}(\{u\})) = 0 \, ,$$

and so $\mu_u(B) = \mu_u(B \cap \varphi^{-1}(\{u\}))$ for any Borel $B \subset \mathcal{X}$;
(iii) for every measurable function $f : \mathcal{X} \to [0, \infty]$,

$$\int_\mathcal{X} f(x) \, \mathrm{d}\mu(x) = \int_\mathcal{Z} \left(\int_{\varphi^{-1}(\{u\})} f(x) \, \mathrm{d}\mu_u(x) \right) \mathrm{d}(\varphi_*\mu)(u) \, .$$

In particular, for any Borel $B \subset \mathcal{X}$, taking f to be the indicator function of B,

$$\mu(B) = \int_\mathcal{Z} \mu_u(B) \, \mathrm{d}(\varphi_*\mu)(u) \, .$$

Theorem B.2 (Disintegration theorem)**.** Let $\mathcal{X}$ and $\mathcal{Z}$ be two Radon spaces, $\mu \in P(\mathcal{X})$ and $\varphi : \mathcal{X} \to \mathcal{Z}$ a Borel-measurable function. There exists a $\varphi_*\mu$-a.e. uniquely determined family of probability measures $\{\mu_u\}_{u \in \mathcal{Z}} \subset P(\mathcal{X})$ that provides a disintegration of μ by φ.

Remark B.1 (Disintegration in a product space)**.** A special case of the disintegration theorem is when $\mathcal{X}$ is a product space $\mathcal{X}_1 \times \mathcal{X}_2$ and $\varphi = p = P^1 : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{X}_1$ is the projection on the first component.

[0] Polish spaces are Radon spaces.

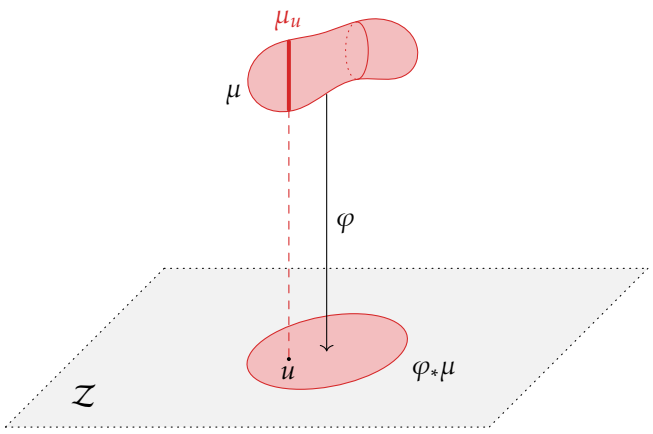


Figure B.1: Disintegration of a measure $\mu \in \mathcal{P}(\mathcal{X})$ into a family of measures $\{\mu_u\}_{u\in\mathcal{Z}}$.

Each $p^{-1}(x_1)$ can be identified with $\mathcal{X}_2$ and the measures $(\mu_{x_1})_{x_1\in\mathcal{X}_1}$ can be seen as measures of $\mathcal{P}(\mathcal{X}_2)$; from a random variable perspective, they actually are the *conditional probability measures* of μ associated to the events $\{X_1 = x_1\}_{x_1\in\mathcal{X}_1}$. Hence, if μ has a density ρ_μ w.r.t. a reference measure $\lambda_\mathcal{X} = \lambda_1 \otimes \lambda_2$, then μ_{x_1} has a density w.r.t. λ_2 on $\mathcal{X}_2$, given by

$$\rho_{\mu_{x_1}}(x_2) = \frac{\rho_\mu(x_1, x_2)}{\rho_{p_*\mu}(x_1)} = \frac{\rho_\mu(x_1, x_2)}{\int_{\mathcal{X}_2} \rho_\mu(x_1, x_2')\,\mathrm{d}x_2'}\,,$$

by the conditional density function formula.

B.1.3 Some absolute continuity results

Proposition B.3 (Density of projection in a product space)**.** Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be Polish spaces equipped with reference measures λ_1 and λ_2, and $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ equipped with the reference measure $\lambda = \lambda_1 \otimes \lambda_2$. Let $\mu \in \mathcal{P}(\mathcal{X})$ be a measure with a density ρ_μ w.r.t. λ, p the standard projection on $\mathcal{X}_1$, and $\{\mu_{x_1}\}_{x_1\in\mathcal{X}_1}$ a disintegration of μ. Then $p_*\mu \ll \lambda_1$ on $\mathcal{X}_1$.

Proof. For all measurable f:

$$\begin{aligned}
\int_{\mathcal{X}_1} f(x_1)\,\mathrm{d}(p_*\mu)(x_1) &= \int_{\mathcal{X}} f(p(x))\rho_\mu(x)\,\mathrm{d}\lambda(x) && \text{(pushforward measure)}\\
&= \iint_{\mathcal{X}_1\times\mathcal{X}_2} f(p(x_1,x_2))\rho_\mu(x_1,x_2)\,\mathrm{d}\lambda_1(x_1)\,\mathrm{d}\lambda_2(x_2) && \text{as } \lambda = \lambda_1\otimes\lambda_2\\
&= \iint_{\mathcal{X}_1\times\mathcal{X}_2} f(x_1)\rho_\mu(x_1,x_2)\,\mathrm{d}\lambda_1(x_1)\,\mathrm{d}\lambda_2(x_2)\\
&= \int_{\mathcal{X}_1} f(x_1)\left(\int_{\mathcal{X}_2} \rho_\mu(x_1,x_2)\,\mathrm{d}\lambda_2(x_2)\right)\mathrm{d}\lambda_1(x_1)\,,
\end{aligned}$$

which means that $p_*\mu$ has a density w.r.t. λ_1 on $\mathcal{X}_1$, given by

$$\text{for all } x_1 \in \mathcal{X}_1, \quad \rho_{p_*\mu}(x_1) = \int_{\mathcal{X}_2} \rho_\mu(x_1,x_2)\,\mathrm{d}\lambda_2(x_2)\,. \qquad \square$$

Proposition B.4 (Density of diffeomorphism)**.** Let $\mathcal{X}$ be a Polish space equipped with a reference measure λ and $\psi : \mathcal{X} \to \mathcal{X}$ a diffeomorphism. Let $\mu \in \mathcal{P}(\mathcal{X})$ a measure such that $\mu \ll \lambda$. Then $\psi_*\mu \ll \lambda$.

Proof. The change of variable formula gives

$$\text{for all } y \in \mathcal{X}, \quad \rho_{\psi_*\mu}(y) = \rho_\mu(\psi^{-1}y)|\det D\psi(\psi^{-1}y)|^{-1}\,. \qquad \square$$

Proposition B.5 (Pushing absolute continuity)**.** Let $\varphi : \mathcal{X} \to \mathcal{Y}$ be measurable and $\mu, \lambda \in \mathcal{P}(\mathcal{X})^2$. If $\mu \ll \lambda$, then $\varphi_*\mu \ll \varphi_*\lambda$.

Proof. If a Borel set $B \subset \mathcal{Y}$ is such that $\varphi_*\lambda(B) = 0$, i.e. $\lambda(\varphi^{-1}(B)) = 0$, then $\mu(\varphi^{-1}(B)) = 0$ since $\mu \ll \lambda$. Hence $\varphi_*\mu(B) = 0$. Furthermore, we can express the density of $\varphi_*\mu$ w.r.t. $\varphi_*\lambda$:

$$\begin{aligned}
\int_{\mathcal{Y}} f(y)\,\mathrm{d}(\varphi_*\mu)(y) &= \int_{\mathcal{X}} f(\varphi(x))\,\mathrm{d}\mu(x) && \text{(pushforward)}\\
&= \int_{\mathcal{X}} f(\varphi(x))\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)\,\mathrm{d}\lambda(x)\\
&= \int_{\mathcal{Y}}\int_{\varphi^{-1}(u)} f(\varphi(x))\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)\,\mathrm{d}\lambda_u(x)\,\mathrm{d}(\varphi_*\lambda)(u) && \text{(disintegration)}
\end{aligned}$$

Hence

$$\frac{\mathrm{d}(\varphi_*\mu)}{\mathrm{d}(\varphi_*\lambda)}(u) = \int_{\varphi^{-1}(u)} f(\varphi(x))\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)\,\mathrm{d}\lambda_u(x)\,. \qquad \square$$

This is already quite useful, since in the two following case we can obtain the absolute continuity of the pushforward measure in $\mathbb{R}^d$:

with p, the orthogonal projection on $\mathbb{R}^h$:	$p_*\mu \ll p_*\mathcal{L}^n = \mathcal{L}^h$;
with the norm $n(x) = \|x\|$:	$n_*\mu \ll n_*\mathcal{L}^n = \frac{4}{3}\pi(\cdot)^2\mathcal{L}^1 \ll \mathcal{L}^1$;
with the squared norm $\tilde{n}(x) = \|x\|^2$:	$\tilde{n}_*\mu \ll \tilde{n}_*\mathcal{L}^n = \frac{4}{2}\pi\sqrt{\cdot}\mathcal{L}^1 \ll \mathcal{L}^1$.

B.1.4 Measurability of set-valued maps

Let X, U be two topological spaces, and let $\mathcal{A}$ denote the Borel σ-algebra on X. A set-valued map S is a map from X to $P(U)$ (the set of subsets of U). This will be denoted by $S : X \rightrightarrows U$. The idea is to introduce notations which are consistent with the case where $S(x) = \{u\}$ for all x in X, where we want to retrieve the standard case of maps $X \to U$. Definitions are taken from [RW09], where measurability is studied when $U = \mathbb{R}^n$. Most results and proofs adapt to a more general setting—in particular when U is a complete Riemannian manifold M. For the sake of completeness, we provide all the proofs, and highlight those that require specific care by replacing $\mathbb{R}^n$ by such a manifold.

Of importance for our proofs, we define:

- The *pre-image* of a set $B \subset U$ is given by
$$S^{-1}(U) = \{x \in X,\ S(x) \cap B \neq \varnothing\}.$$
- The *domain* of S is $S^{-1}(U)$, that is $\{x \in X,\ S(x) \neq \varnothing\}$.

We will often use the following relation: if a set A can be written as $A = \bigcup A_k$, then $S^{-1}(A) = \bigcup S^{-1}(A_k)$. Indeed, $x \in S^{-1}(A) \Leftrightarrow S(x) \cap A \neq \varnothing \Leftrightarrow \exists k,\ S(x) \cap A_k \neq \varnothing \Leftrightarrow \exists k,\ x \in S^{-1}(A_k) \Leftrightarrow x \in \bigcup_k S^{-1}(A_k)$.

A set-valued map $S : X \rightrightarrows U$ is said to be *measurable* if, for any open set $O \subset U$,

$$S^{-1}(O) \in \mathcal{A}. \tag{B.1}$$

Note that if S is measurable (as a set-valued map), then its domain must be measurable as well (as an element of $\mathcal{A}$). We say that $S : X \rightrightarrows U$ is *closed-valued* if $S(x)$ is a closed subset of U for all $x \in X$.

Proposition B.6 (Theorem 14.3.c in [RW09])**.** A closed-valued map S is measurable if and only if $S^{-1}(B) \in \mathcal{A}$ for all $B \subset U$ that are either:
(a) open (the definition);
(b) compact;
(c) closed.

Proof of Proposition B.6.
- (a) $\Rightarrow$ (b): For a compact $B \subset U$, let $B_k = \{x \in U,\ d(x, B) < k^{-1}\}$, $k \geq 0$ (that is open). Note that $x \in S^{-1}(B) \Leftrightarrow S(x) \cap B \neq \varnothing \Leftrightarrow S(x) \cap B_k \neq \varnothing$ for all k because $S(x)$ is a closed set. Hence $S^{-1}(B) = \bigcap_k S^{-1}(B_k)$. All the $S^{-1}(B_k)$ are measurable, so is $S^{-1}(B)$ as a countable intersection of measurable sets.
- (b) $\Rightarrow$ (a): Fix O an open set of U. As we assume U to be a complete separable Riemannian manifold, O can be written as a countable union of compact balls: $O = \bigcup_n \overline{B(x_n, r_n)}$.
- (b) $\Rightarrow$ (c): Immediate.
- (c) $\Rightarrow$ (b): A closed set B can be obtained as a countable union of compact sets by letting $B = \bigcup_n B \cap \overline{B(x_0, n)}$ for some x_0. Hence $S^{-1}(B) = \bigcup_n S^{-1}(B \cap \overline{B(x_0, n)})$ is in $\mathcal{A}$. □

Now, we introduce a proposition on operations that preserve measurability of closed-set valued maps. The proof requires adaptation from the one of [RW09] because the latter uses explicitly the fact that one can compute Minkowski sums of sets (which may not make sense on a manifold).

Proposition B.7 (Proposition 14.11 in [RW09], adapted to the manifold case)**.** Let S_1 and $S_2 : X \rightrightarrows U$ be two measurable closed-set valued maps. Then
- $P : x \mapsto S_1(x) \times S_2(x)$ is measurable as a closed-valued map in $U \times U$ (equipped with the product topology).
- $Q : x \mapsto S_1(x) \cap S_2(x)$ is measurable.

Proof. The first point can be proved in the same spirit as the proof proposed by Rockafellar and Wets. Namely, let O' be an open set in $U \times U$. By definition of the product topology, O' can be obtained as $\bigcup_n O_1^{(n)} \times O_2^{(n)}$ where $O_1^{(n)}$ and $O_2^{(n)}$ are open sets in U. Then $P^{-1}(O') = \bigcup_n P^{-1}(O_1^{(n)} \times O_2^{(n)})$. Now, observe that $P^{-1}(A \times B) = \{x,\ S_1(x) \times S_2(x) \in A \times B = \{x,\ S_1(x) \in A \text{ and } S_2(x) \in B\} = S_1^{-1}(A) \cap S_2^{-1}(B)$, so that finally, $P^{-1}(O') = \bigcup_n S_1^{-1}(O_1^{(n)}) \cap S_2^{-1}(O_2^{(n)})$ that is measurable as a countable union of (finite) intersection of measurable sets (given that S_1, S_2 are measurables). Note that this does not require S_1, S_2 to be closed-valued.

Now, let us focus on the second point, that requires more attention. Thanks to the previous proposition, it is sufficient to show that $Q^{-1}(C) \in \mathcal{A}$ for any compact set $C \subset U$. In [RW09], this is done by writing $Q^{-1}(C) = \{x,\ S_1(x) \cap S_2(x) \cap C \neq \varnothing\} = \{x,\ R_1(x) \cap R_2(x) \neq \varnothing\} = \{x,\ 0 \in (R_1(x) - R_2(x))\} = (R_1 - R_2)^{-1}(\{0\})$, where $R_j(x) = S_j(x) \cap C$ (that is also closed valued), and using the fact that the (Minkowski) difference of measurable closed-valued maps is measurable as well [RW09, Prop. 14.11.c].

To adapt this idea (we cannot consider Minkowski difference in our setting), we introduce the diagonal $\Delta = \{(u, u),\ u \in U\} \subset U \times U$. Now, observe that $R_1(x) \cap R_2(x) \neq \varnothing \Leftrightarrow (R_1(x) \times R_2(x)) \cap \Delta \neq \varnothing$, that is $x \in R^{-1}(\Delta)$, where $R(x) = R_1(x) \times R_2(x)$. Now, since the maps R_1 and R_2 are measurable closed-valued maps (inherited from S_1, S_2), so is R according to the previous point. And since Δ is closed, $R^{-1}(\Delta) = Q^{-1}(C)$ is measurable. □

B.2 A bit of convex analysis

In this section, we define some notions of convex analysis that are crucial in optimal transport theory. We refer the reader to [Vil08, MG11] or any optimal transport textbook for more information on this matter. We first recall the definition of the *subdifferential* of a function $f : \mathcal{X} \to \overline{\mathbb{R}}$ defined on $\mathcal{X}$ in duality with a space $\mathcal{Y}$:

$$\partial f \triangleq \left\{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid f(z) - f(x) \geq \langle y, z - x\rangle + o(|z - x|)\right\}, \tag{B.2}$$

$$\partial f(x) \triangleq \left\{y \in \mathcal{Y} \mid f(z) - f(x) \geq \langle y, z - x\rangle + o(|z - x|)\right\} \quad \text{for } x \in \mathcal{X}. \tag{B.3}$$

It consists of the set of (point, slope) pairs for which y is the slope of a hyperplane supporting the graph of f at $(x, f(x))$. We also recall the *Legendre transform* of f, defined as

$$\text{for all } y \in \mathcal{Y}, \quad f^*(y) \triangleq \sup_{x \in \mathcal{X}} x \cdot y - f(x) \tag{B.4}$$

when this value is finite, and the following important characterization of the subdifferential in terms of Legendre transform:

$$\text{for all } x, y \in \mathcal{X} \times \mathcal{Y}, \quad f(x) + f^*(y) \geq x \cdot y \quad \text{with equality iff } y \in \partial f(x). \tag{B.5}$$

Moreover, for any lower semi-continuous (l.s.c.) function f, one has $(f^*)^*$ if and only if f is convex. For a (cost) function $c : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$, we will now define the notions of *c-convexity*, *c-subdifferential* and *c-transform* and see that they particularize respectively into those of convexity, subdifferential and Legendre transform in the case where $c(x, y) = -x \cdot y$.

Definition B.7 (c-convexity)**.** *Let $c : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$ and $f : \mathcal{X} \to \overline{\mathbb{R}}$. We define the* c-transform *of f as*

$$\text{for all } y \in \mathcal{Y}, \quad f^c(y) \triangleq \sup_{x \in \mathcal{X}} -c(x, y) - f(x) \tag{B.6}$$

and its c-subdifferential *as*

$$\partial_c f \triangleq \left\{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid f(x) + f^c(y) = -c(x, y)\right\}, \tag{B.7}$$

which are respectively the generalizations of (B.4) and (B.5) to an arbitrary c. We say that f is c-convex *if it is not identically $+\infty$ and if $(f^c)^{\overline{c}} = f$, where $g^{\overline{c}}$ is the $\overline{c}$-transform of g with $\overline{c}(y, x) = c(x, y)$ the symmetrized cost, i.e. for all $x \in \mathcal{X}$, $g^{\overline{c}}(x) \triangleq \sup_{y \in \mathcal{Y}} -c(x, y) - g(y)$.*

B.3 Geometry

B.3.1 (One) general notion(s)

Definition B.8 (Clarke's tangent cone)**.** *Let C be a nonempty closed subset of the Banach space B. The* Clarke's tangent cone *to C at $x \in C$, denoted by $T_C(x)$ consists of all vectors $v \in B$ such that for any sequence $\{t_n\}_{n\geq 0} \subset \mathbb{R}$ tending to zero and any sequence $\{x_n\}_{n\geq 0} \subset C$ tending to x, there exists a sequence $\{v_n\}_{n\geq 0} \subset B$ tending to v such that for all $n \geq 0$, $x_n + t_n v_n \in C$.*

B.3.2 (Very few) notions of differential geometry

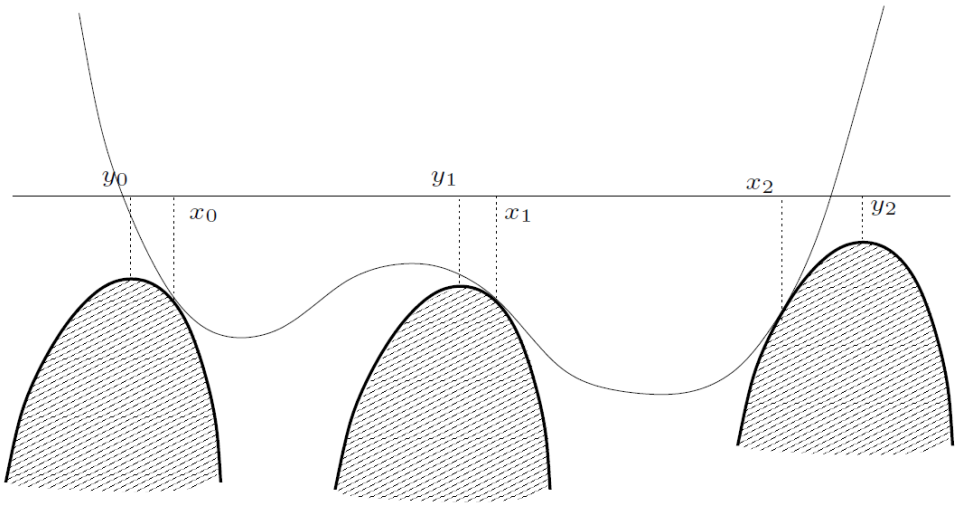


Figure B.2: A c-convex function. As put in [Vil08]: "A c-convex function is a function whose graph you can entirely *caress* from below with a tool whose shape is the negative of the cost function (this shape might vary with the point y). In the picture $y_i \in \partial_x \psi(x_i)$."

Definition B.9 (Riemannian manifold)**.** *A* manifold *M of dimension d is a space which is locally homeomorphic to $\mathbb{R}^d$. At each point $p \in M$ is a* tangent space *$T_pM = \{$velocity of curves through $p\}$, and M is* smooth *if this whole structure is. A* Riemannian metric *is a smoothly varying choice $p \mapsto \langle \cdot, \cdot \rangle_p$, an inner product on T_pM; $(M, \langle \cdot, \cdot \rangle)$ is then a* Riemannian manifold.

For any curve γ, we then have $\|\dot\gamma(t)\|^2_{\gamma(t)} = \langle \dot\gamma(t), \dot\gamma(t) \rangle_{\gamma(t)}$. The distance function *is defined as*

$$d(p, q) = \inf\{ \int_0^1 \|\dot\gamma(t)\|^2 \, \mathrm{d}t \mid \gamma(0) = p, \gamma(1) = q \}.$$

If the infimum is achieved by a $\gamma^\star$, then $\gamma^\star$ is called a geodesic *between p and q.*

Definition B.10 (Exponential map)**.** *Let M be a differentiable manifold and $p \in M$. Let $v \in T_pM$ be a tangent vector to M at p. Then there is a unique geodesic γ_v satisfying $\gamma_v(0) = p$ with initial tangent vector $\gamma_v'(0) = v$. The corresponding* exponential map *is defined by* $\exp_p(v) = \gamma_v(1)$.

Intuitively speaking, the exponential map takes a given tangent vector v to the manifold at point p, runs along the geodesic starting at p and goes in the direction of v for a unit time.

B.3.3 (Even fewer) notions of Alexandrov geometry

Definition B.11 (Sectional curvature, synthetic (Alexandrov) definition)**.** *A geodesic space $(\mathcal{X}, d)$ is said to have* nonnegative Alexandrov curvature *if triangles in $\mathcal{X}$ are no more "skinny" than reference triangles drawn on the model space $\mathbb{R}^2$. More precisely: if xyz is a triangle in $\mathcal{X}$, $x_0y_0z_0$ is a triangle drawn on $\mathbb{R}^2$ with*

$$\begin{cases} d(x_0, y_0) = d(x, y) \\ d(y_0, z_0) = d(y, z) \\ d(z_0, x_0) = d(z, x), \end{cases}$$

and if x' is a midpoint between y and z and x'_0 a midpoint between y_0 and z_0, then one should have $d(x_0, x'_0) \leq d(x, x')$.

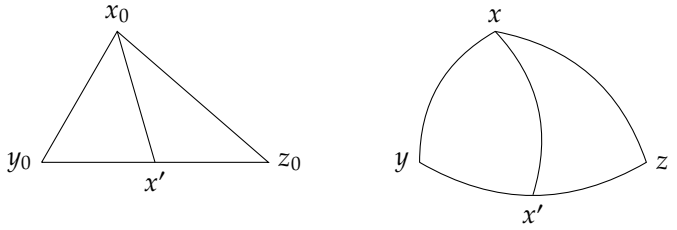


Figure B.3: Synthetic definition of Alexandrov curvature by triangles.

To define what it means (synthetically) to have a curvature bounded below by $K \in \mathbb{R}$, not necessarily 0, we replace $\mathbb{R}^2$ by these model spaces:

- the sphere $S^2(1/\sqrt{K})$ which has constant sectional curvature $K > 0$;
- the plane $\mathbb{R}^2$ which has constant sectional curvature $K = 0$;
- the hyperbolic space $\mathbb{H}^2(1/\sqrt{|K|})$ which has constant sectional curvature $K < 0$.

Definition B.12 (Sectional curvature, analytic definition)**.** *[We do not give here the analytic definition of the sectional curvature. For the sake of having intuition on the conditions of Proposition 1.4, we give the (thus incomplete) definition of having an* asymptotically nonnegative curvature*] Let M be a Riemannian manifold. We say that M has* asymptotically nonnegative curvature *if all sectional curvatures σ_x at point x satisfy*

$$\sigma_x \geq -\frac{C}{d(x_0, x)^2}$$

for some positive constant C and some $x_0 \in M$.

Toponogov's theorem makes the connection between both definitions. In particular, a complete simply connected Riemannian manifold has:

- (analytic) non-negative sectional curvature if and only if the function $f_p(x) = d(p, x)^2$ is 1-concave for all points p;
- (analytic) non-positive sectional curvature if and only if the function $f_p(x) = d(p, x)^2$ is 1-convex for all points p.

B.4 Other notions

B.4.1 Approximate differentiability

Definition B.13 (Approximate differentiability, [Vil08])**.** *Let U be an open set of a Riemannian manifold M, and let $f : U \to \mathbb{R} \cup \{\pm\infty\}$ be a measurable function. Then f is said to be* approximately differentiable *at $x \in U$ if there is a measurable function $\tilde{f} : U \to \mathbb{R}$, differentiable at x, such that the set $\{\tilde{f} = f\}$ has density 1 at x; in other words,*

$$\lim_{r \to 0} \frac{\mathrm{vol}\left[\left\{z \in B_r(x); f(z) = \tilde{f}(z)\right\}\right]}{\mathrm{vol}\left[B_r(x)\right]} = 1\,.$$

Then one defines the approximate gradient of f at x by the formula

$$\tilde{\nabla} f(x) = \nabla \tilde{f}(x)\,.$$

B.4.2 General definition of submodularity

Definition B.14. *Submodularity is more generally defined for functions defined on subsets of $\mathbb{R}^n$ of the form $\mathcal{Z} = \Pi_{i=1}^n \mathcal{Z}_i$ with each $\mathcal{Z}_i$ a compact subset of $\mathbb{R}$ [Bac19]. A function $H : \mathcal{Z} \to \mathbb{R}$ is then submodular if for all z_1, z_2 in $\mathcal{Z}$, $H(z_1) + H(z_2) \geq H(\max\{z_1, z_2\}) + H(\min\{z_1, z_2\})$, where the min and max operations are applied component-wise. In our case where c is a function over $\mathbb{R} \times \mathbb{R}$, not necessarily differentiable, this more general formulation of the submodularity becomes:*

$$\text{for all } x, y \in \mathbb{R}, \text{ for all } \alpha, \beta \geq 0, \quad c(x+\alpha, y) + c(x, y+\beta) \geq c(x, y) + c(x+\alpha, y+\beta)$$

and we recover the definition of Sec. 1.2.2.

B.4.3 Numbered limb system

The notion of numbered limb system has been introduced in [HW95] and applied to optimal transport by [AKM11, CMN10].

Definition B.15 (Numbered limb system)**.** *Let X and Y be Borel subsets of complete separable metric spaces. A relation $S \subset X \times Y$ is a* numbered limb system *if there is a countable disjoint decomposition of $X = \cup_{i=0}^{\infty} I_{2i+1}$ and of $Y = \cup_{i=0}^{\infty} I_{2i}$ with a sequence of maps $f_{2i} : \mathrm{Dom}(f_{2i}) \subset Y \longrightarrow X$ and $f_{2i+1} : \mathrm{Dom}(f_{2i+1}) \subset X \longrightarrow Y$ such that $S = \cup_{i=1}^{\infty} \mathrm{Graph}(f_{2i-1}) \cup \mathrm{Antigraph}(f_{2i})$, with $\mathrm{Dom}(f_k) \cup \mathrm{Ran}(f_{k+1}) \subset I_k$ for each $k \geq 0$. The system has (at most) N limbs if $\mathrm{Dom}(f_k) = \varnothing$ for all $k > N$.*

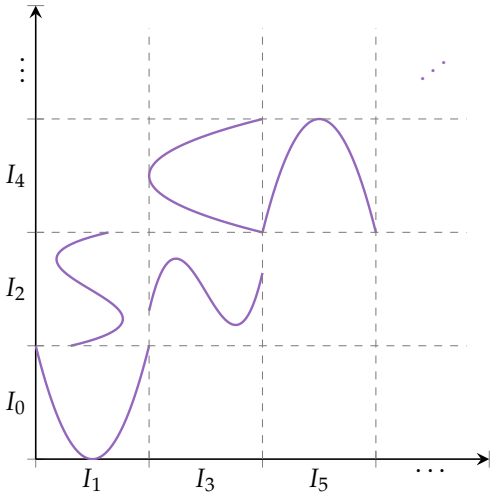
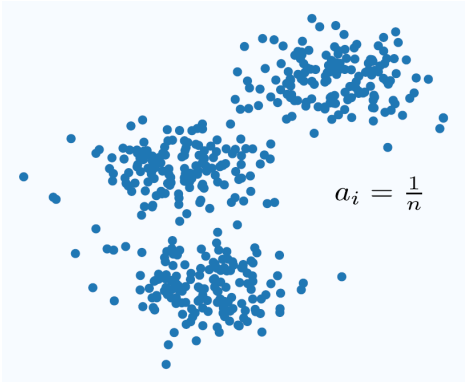
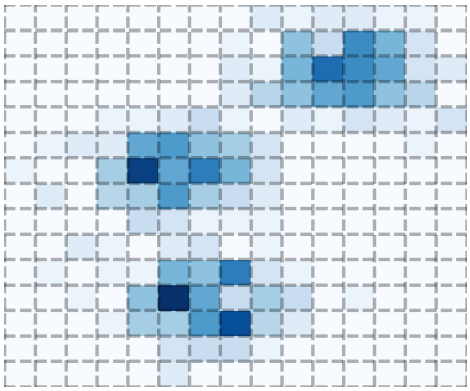


Figure B.4: Representation of a numbered limb system with N limbs. The subsets I_k are represented connected for visual convenience but do not need to be.

B.4.4 Lagrangian and Eulerian discretizations



(a) Lagrangian formulation (point cloud). The discrete probability measure is $\sum_{i=1}^{n} \frac{1}{n} \delta_{x_i}$.



(b) Eulerian formulation (histogram). The discrete probability measure is $\sum_{i=1}^{n} a_i \delta_{\hat{x}_i}$ with a a probability vector and $(\hat{x}_i)$ a regular grid on the domain.

Figure B.5: Associating a discrete probability measure to a dataset $(x_i)_{i=1}^{n}$: Lagrangian and Eulerian methods. Image taken from [Vay20].

Bibliography

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005. 1, 25, 27

[AKM11] Najma Ahmad, Hwa Kil Kim, and Robert J McCann. Optimal transportation, topology and uniqueness. *Bulletin of Mathematical Sciences*, 1(1):13–32, 2011. 10, 11, 46, 61

[AMJ18] David Alvarez-Melis and Tommi S Jaakkola. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018. 1

[AMJJ19] 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. 18

[Bac19] Francis Bach. Submodular functions: from discrete to continuous domains. *Mathematical Programming*, 175(1):419–459, 2019. 61

[BB01] Guy Bouchitté and Giuseppe Buttazzo. Characterization of optimal shapes and masses through monge-kantorovich equation. *Journal of the European Mathematical Society*, 3(2):139–168, 2001. 1

[BB07] Patrick Bernard and Boris Buffoni. Optimal mass transportation and mather theory. *Journal of the European Mathematical Society*, 9(1):85–121, 2007. 1

[BBK06] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006. 1

[BHS22] Robert Beinert, Cosmas Heiss, and Gabriele Steidl. On assignment problems related to gromov-wasserstein distances on the real line. *arXiv preprint arXiv:2205.09006*, 2022. 21, 22, 35, 36, 40, 45

[Bre87] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987. 1

[BTBD20] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020. 36

[Car08] Guillaume Carlier. Remarks on toland's duality, convexity constraint and optimal transport. *Pacific Journal of Optimization*, 4(3):423–432, 2008. 12

[CE07] Guillaume Carlier and Ivar Ekeland. Equilibrium structure of a bidimensional asymmetric city. *Nonlinear Analysis: Real World Applications*, 8(3):725–748, 2007. 1

[CFHR17] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[CM19] Samir Chowdhury and Facundo Mémoli. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019. 17

[CMN10] Pierre-André Chiappori, Robert J McCann, and Lars P Nesheim. Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness. *Economic Theory*, 42(2):317–354, 2010. 1, 10, 11, 38, 61

[Cul06] Michael John Priestley Cullen. *A mathematical theory of large-scale atmosphere/ocean flow*. World Scientific, 2006. 1

[DSS+20] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, 2020. 1

[DSSS22] Pınar Demetçi, Rebecca Santorella, Björn Sandstede, and Ritambhara Singh. Unsupervised integration of single-cell multi-omics datasets with disproportionate cell-type representation. In *International Conference on Research in Computational Molecular Biology*, pages 3–19. Springer, 2022. 1

[FGM10] Joaquin Fontbona, Hélène Guérin, and Sylvie Méléard. Measurability of optimal transportation and strong coupling of martingale measures. *Electronic communications in probability*, 15:124–133, 2010. 27, 29, 50, 51, 52

[FKM11] Alessio Figalli, Young-Heon Kim, and Robert J McCann. When is multidimensional screening a convex program? *Journal of Economic Theory*, 146(2):454–478, 2011. 1

[FMMS02] Uriel Frisch, Sabino Matarrese, Roya Mohayaee, and Andrei Sobolevski. A reconstruction of the initial conditions of the universe by optimal mass transportation. *Nature*, 417(6886):260–262, 2002. 1

[FMP10] Alessio Figalli, Francesco Maggi, and Aldo Pratelli. A mass transportation approach to quantitative isoperimetric inequalities. *Inventiones mathematicae*, 182(1):167–211, 2010. 1

[GDBFL19] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR, 2019. 1

[GJB19] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019. 1

[GKPS99] Mikhael Gromov, Misha Katz, Pierre Pansu, and Stephen Semmes. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152. Springer, 1999. 15

[GO03] Tilmann Glimm and Vladimir Oliker. Optical design of single reflector systems and the monge–kantorovich mass transfer problem. *Journal of Mathematical Sciences*, 117(3):4096–4108, 2003. 1

[GPC18] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018. 1

[GSR+17] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017. 1

[HT01] Steven Haker and Allen Tannenbaum. Optimal mass transport and image registration. In *Proceedings IEEE Workshop on Variational and Level Set Methods in Computer Vision*, pages 29–36. IEEE, 2001. 1

[HW95] Kevin Hestir and Stanley C Williams. Supports of doubly stochastic measures. *Bernoulli*, 1(3):217–243, 1995. 11, 61

[Kan42] L Kantorovich. On the translocation of masses. *Dokl. Akad. Nauk. USSR*, 37(7–8):227–229, 1942. 6

[KB57] Tjalling C Koopmans and Martin Beckmann. Assignment problems and the location of economic activities. *Econometrica: journal of the Econometric Society*, pages 53–76, 1957. 18

[Kon76] Hiroshi Konno. Maximization of a convex quadratic function under linear constraints. *Mathematical programming*, 11(1):117–127, 1976. 19

[KPT+17] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017. 1

[Lev99] Vladimir Levin. Abstract cyclical monotonicity and monge solutions for the general monge–kantorovich problem. *Set-Valued Analysis*, 7(1):7–32, 1999. 14

[Loe09] Grégoire Loeper. On the regularity of solutions of optimal transportation problems. *Acta mathematica*, 202(2):241–283, 2009. 1

[LV09] John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, pages 903–991, 2009. 1

[McC94] Robert John McCann. *A convexity theory for interacting gases and equilibrium crystals*. Princeton University, 1994. 1

[McC12] Robert J McCann. A glimpse into the differential topology and geometry of optimal transport. *arXiv preprint arXiv:1207.1867*, 2012. 10, 14

[Mém11] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011. 1, 5, 14, 15, 54

[MG11] Robert J McCann and Nestor Guillen. Five lectures on optimal transportation: geometry, regularity and applications. *Analysis and geometry of metric measure spaces: lecture notes of the séminaire de Mathématiques Supérieure (SMS) Montréal*, pages 145–180, 2011. 10, 14, 58

[ML18] Haggai Maron and Yaron Lipman. (probably) concave graph matching. *Advances in Neural Information Processing Systems*, 31, 2018. 17

[Moa16] Abbas Moameni. A characterization for solutions of the monge-kantorovich mass transport problem. *Mathematische Annalen*, 365(3):1279–1304, 2016. 10, 14

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781. 1, 5

[MPW12] Robert J McCann, Brendan Pass, and Micah Warren. Rectifiability of optimal transportation plans. *Canadian Journal of Mathematics*, 64(4):924–934, 2012. 13, 14

[MR20] Abbas Moameni and Ludovic Rifford. Uniquely minimizing costs for the kantorovitch problem. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 29, pages 507–563, 2020. 14

[MV05] Francesco Maggi and Cédric Villani. Balls have the worst best sobolev inequalities. *The Journal of Geometric Analysis*, 15(1):83–121, 2005. 1

[PC19] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 5, 15, 46

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 36

[Rac98] ST Rachev. L. r uschendorf. mass transportation problems. *Probab. Appl. Springer-Verlag, New York*, 1998. 1

[RTG98] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE, 1998. 1

[RW09] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009. 50, 56, 57

[San15] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015. 5, 8, 10, 12, 14, 27

[SST+19] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019. 1

[Stu06] Karl-Theodor Sturm. On the geometry of metric measure spaces. *Acta mathematica*, 196(1):65–131, 2006. 15

[Stu12] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012. 2, 5, 14, 17, 20, 21, 22, 45

[SVP21] Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021. 1, 19

[Vay20] Titouan Vayer. A contribution to optimal transport on incomparable spaces. *arXiv preprint arXiv:2011.04447*, 2020. 2, 17, 21, 22, 30, 31, 35, 45, 46, 62

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. 52

[VFC+19] Titouan Vayer, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced gromov-wasserstein. *Advances in Neural Information Processing Systems*, 32, 2019. 45

[Vil08] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2008. 5, 9, 10, 12, 46, 51, 58, 59, 60

[Xia07] Qinglan Xia. The formation of a tree leaf. *ESAIM: Control, Optimisation and Calculus of Variations*, 13(2):359–377, 2007. 1

[XLC19] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019. 1