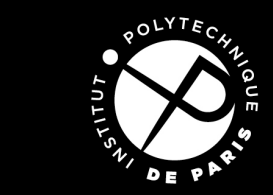
Optimal transport in high dimension : obtaining regularity and robustness using convexity and projections

François-Pierre Paty



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAG003

Thèse de doctorat

ENSAE

IP PARIS

Optimal Transport in High Dimension: Obtaining Regularity and Robustness using Convexity and Projections

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à École nationale de la statistique et de l'administration économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 29 juin 2021, par

FRANÇOIS-PIERRE PATY

Composition du Jury :

Guillaume Lecué Professeur, ENSAE Paris	Président
Jérôme Malick Directeur de recherche, Université Grenoble Alpes	Rapporteur
François-Xavier Vialard Professeur, Université Gustave Eiffel	Rapporteur
Laetitia Chapel Maîtresse de conférences, Université de Bretagne Sud	Examinatrice
Giovanni Conforti Professeur assistant, Institut Polytechnique de Paris	Examinateur
Umut Şimşekli Chargé de recherche, INRIA et École Normale Supérieure de Paris	Examinateur
Marco Cuturi Professeur, ENSAE Paris et Google Brain	Directeur de thèse

Abstract

Over the past few years, optimal transport has gained popularity in machine learning as a way to compare probability distributions. Unlike more classical dissimilarities for probability measures, such as the Kullback-Leibler divergence, optimal transport distances (or Wasserstein distances) can deal with distributions of disjoint supports by taking into account the geometry of the underlying ground space. This strength is, however, hampered by the fact that these distances are usually computed by solving a linear program, resulting, when this ground space is high-dimensional, in well documented statistical challenges, usually referred to as the "curse" of dimensionality. Finding new methodologies that can mitigate this issue is therefore crucial if one wants optimal transport-based algorithms to perform well on real data.

Beyond this purely metric aspect, another appealing feature of optimal transport theory is that it provides mathematical tools to study maps that are able to morph (or push-forward) a measure into another. Such maps are playing an increasingly important role in various areas of science (biology, neuroimaging) or subdomains in machine learning (generative models, domain adaptation), to name a few. Estimating such morphings, or maps, that are both optimal and able to generalize outside the data, is an open problem.

In this thesis, we propose a new estimation framework to compute proxies to the Wasserstein distance. That framework aims at handling high-dimensionality by taking advantage of the low-dimensional structures hidden in the distributions. This can be achieved by projecting the measures onto a subspace chosen so as to maximize the Wasserstein distance between their projections. In addition to this novel methodology, we show that this framework falls into a broader connection between regularization when computing Wasserstein distances and adversarial robustness.

In the next contribution, we start from the same problem, estimation of optimal transport in high dimensions, but adopt a different perspective: rather than changing the ground cost, we go back to the more fundamental Monge perspective on optimal transport and use the Brenier theorem and Caffarelli's regularity theory to propose a new estimation procedure to characterize maps that are Lipschitz and gradients of strongly convex functions.

Résumé

Au cours des dernières années, le transport optimal a gagné en popularité en apprentissage automatique comme moyen de comparer des mesures de probabilité. Contrairement aux dissimilarités plus classiques pour les distributions de probabilité, telles que la divergence de Kullback-Leibler, les distances de transport optimal (ou distances de Wasserstein) permettent de comparer des distributions dont les supports sont disjoints en prenant en compte la géométrie de l'espace sous-jacent. Cet avantage est cependant entravé par le fait que ces distances sont généralement calculées en résolvant un programme linéaire, ce qui pose, lorsque l'espace sous-jacent est de grande dimension, des défis statistiques bien documentés et auxquels on se réfère communément sous le nom de "fléau" de la dimension. Trouver de nouvelles méthodologies qui puissent atténuer ce problème est donc un enjeu crucial si l'on veut que les algorithmes fondés sur le transport optimal puissent fonctionner en pratique.

Au-delà de cet aspect purement métrique, un autre intérêt de la théorie du transport optimal réside en ce qu'elle fournit des outils mathématiques pour étudier des cartes qui peuvent transformer, ou transporter, une mesure en une autre. De telles cartes jouent un rôle de plus en plus important dans divers domaines des sciences (biologie, imagerie cérébrale) ou sous-domaines de l'apprentissage automatique (modèles génératifs, adaptation de domaine), entre autres. Estimer de telles transformations qui soient à la fois optimales et qui puissent être généralisées en dehors des simples données, est un problème ouvert.

Dans cette thèse, nous proposons un nouveau cadre d'estimation pour calculer des variantes des distances de Wasserstein. Le but est d'amoindrir les effets de la haute dimension en tirant partie des structures de faible dimension cachées dans les distributions. Cela peut se faire en projetant les mesures sur un sous-espace choisi de telle sorte à maximiser la distance de Wasserstein entre leurs projections. Outre cette nouvelle méthodologie, nous montrons que ce cadre d'étude s'inscrit plus largement dans un lien entre la régularisation des distances de Wasserstein et la robustesse.

Dans la contribution suivante, nous partons du même problème d'estimation du transport optimal en grande dimension, mais adoptons une perspective différente : plutôt que de modifier la fonction de coût, nous revenons au point de vue plus fondamental de Monge et proposons d'utiliser le théorème de Brenier et la théorie de la régularité de Caffarelli pour définir une nouvelle procédure d'estimation des cartes de transport lipschitziennes qui soient le gradient d'une fonction fortement convexe.

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, Marco Cuturi, pour m'avoir guidé à travers les chemins escarpés de la recherche. Merci d'avoir su me convaincre d'entreprendre une thèse : ce fut une expérience plus riche que je ne l'avais prévu. Merci de m'avoir laissé une grande liberté dans mon travail, et de m'avoir initié au domaine académique avec tant de générosité.

Je remercie Jérôme Malick et François-Xavier Vialard d'avoir eu la gentillesse d'accepter de rapporter ma thèse et d'avoir pris le temps de relire précisément ce manuscrit. Leurs remarques auront contribué à élever la qualité de ces pages. Je veux aussi remercier Laetitia Chapel, Giovanni Conforti, Guillaume Lecué et Umut Şimşekli qui me font l'amitié de constituer mon jury : c'est un honneur de présenter mes travaux à ces chercheurs qui m'ont inspiré.

Merci à tous les doctorants et chercheurs avec lesquels j'ai eu l'occasion, au cours de ma thèse, d'échanger des idées, de collaborer, et parfois d'écrire des articles. J'ai beaucoup appris de ces discussions variées et enrichissantes. Merci à Alexandre d'Aspremont, avec lequel j'ai eu le grand plaisir de travailler sur ce qui constitue la seconde partie de ce manuscrit.

Je remercie les doctorants et les professeurs que j'ai eu le privilège de côtoyer au CREST et dont la bienveillance et la bonhomie quotidiennes ont largement contribué à faire du laboratoire un lieu agréable de travail. Merci à Pascale Deniau et Edith Verger pour m'avoir aidé dans mes démarches administratives et pour leur constante bonne humeur. Merci à Nicolas et Martin d'avoir été des coorganisateurs dévoués du séminaire. Merci à Meyer d'être un modèle de productivité. Merci à mes grands frères de thèse, Boris et Théo, pour leurs conseils avisés et dont l'amitié m'est précieuse. Grazie a Giulia per essere una così buona amica che mi capisce così bene. Ogni volta che, d'estate, vedrò lo Scorpione, mi ricorderò questi arcipelaghi siderali di Cortona.

Je tiens particulièrement à remercier l'École, principale artisane à travers moi de ces travaux, et ses professeurs, mes professeurs, qui tous m'ont transmis bien davantage que des connaissances : curiosité, ouverture d'esprit, rigueur

intellectuelle, dépassement de soi. Pour leur infinie bienveillance et l'importance décisive que leur rencontre eut sur moi, je veux singulièrement remercier Brigitte Thiébaut, Claire Busquet, Amélie Collot, Franck Henry, Frédéric Harymbat, Marie-Christine Godefroy, Yves Duval et Marie Clément.

Quoique ce manuscrit vienne conclure ma thèse, il ne représente pas tout à fait l'ensemble du travail que j'ai fourni ces trois dernières années, puisqu'il éclipse ce qui fut pour moi, peut-être davantage que la recherche elle-même, l'activité la plus indispensable : l'enseignement. Je veux donc remercier chaleureusement les étudiants de l'ENSAE auxquels j'ai eu le plaisir et l'honneur de transmettre quelques bribes de mathématiques ; ils m'ont en retour enseigné non seulement des mathématiques, puisque c'est en expliquant que l'on comprend profondément les choses, mais encore qu'il m'est plus cher de partager que de chercher. La joie d'enseigner me fut plus d'une fois salvatrice.

Merci à mes amis de Bayeux, de prépa, d'école, d'athlétisme et d'ailleurs, qui me pardonneront de ne pas citer la liste de leurs noms puisqu'ils se reconnaîtront, pour avoir été de solides soutiens. Merci à Gabriel d'avoir été un colocataire et un ami indéfectible. Merci à mes plus proches amis, Armand, Clément, Théo et Thibaut, que notre amitié soit à vie. Merci enfin à ma famille : à mon frère que j'admire plus qu'il ne le sait, et à mes parents qui m'ont tout donné.

Contents

Chapter 1

Introduction

Optimal transport (OT) dates back to the end of the 18th century, when French mathematician Monge [1781] proposed to solve the problem of *déblais* and *remblais*. Yet, the mathematical formulation of Monge was rapidly found to meet its limits in the lack of provable existence of a solution to his problem. It is only after 150 years that OT enjoyed a resurgence, when Kantorovich [1942] understood the suitable framework that would allow to solve Monge's problem and give rise to fundamental tools and theories in pure and applied mathematics. In the last few years, OT has also found new applications in statistics and machine learning as a way to analyze data: in supervised machine learning [Frogner et al., 2015, Abadeh et al., 2015, Courty et al., 2016], computer graphics [Solomon et al., 2015, Bonneel et al., 2016], imaging [Rabin and Papadakis, 2015, Cuturi and Peyré, 2016], generative models [Arjovsky et al., 2017, Salimans et al., 2018, Genevay et al., 2018], biology [Hashimoto et al., 2016, Schiebinger et al., 2019, Huizing et al., 2021] or natural language processing [Grave et al., 2019, Alaux et al., 2019].

1.1 Background on optimal transport

In this section, we give an introduction to optimal transport and present the main results upon which this thesis is built. Inspired by the reference books of Villani [2003, 2009], Santambrogio [2015], Peyré and Cuturi [2019], we will start by a general presentation to the optimal transport problem as originally introduced by Monge [1781] and further relaxed by Kantorovich [1942]. We will focus on the Euclidean case where the ground-cost function is a power of the Euclidean distance, giving rise to the rich Wasserstein geometry. In particular, we will take an even closer look on the so-called "quadratic case", that is the case where the ground-cost function is the squared Euclidean distance. We finally present classical algorithms used to numerically solve or approximate the optimal transport problem.

1.1.1 Monge Problem

The original optimal transport problem, as introduced by Monge [1781], is a practical one: given a pile of earth of a determined shape, and a target location for this earth to be transported to, what is the optimal way to transport the earth from one place to the other?

Mathematically, a pile of earth lying on the ground can be represented by a probability measure $\mu \in \mathscr{P}(\mathcal{X})$, where $\mathcal{X} = \mathbb{R}^2$: for any Borel set $A \subset \mathcal{X}$, $\mu(A)$ represents the total mass of earth lying over A. Transporting the earth from $\mu \in \mathscr{P}(\mathcal{X})$ to $\nu \in \mathscr{P}(\mathcal{X})$ means that there exists a function $T : \mathcal{X} \to \mathcal{X}$ such that for any Borel $A \subset \mathcal{X}$ (a target location), the initial mass $\mu\left(T^{-1}(A)\right)$ being sent to A should be equal to the target mass $\nu(A)$ lying over A. If a function $T : \mathcal{X} \to \mathcal{X}$ verifies this condition, we shall say that T is a valid transportation map sending μ to ν, and we will write: $T_\sharp \mu = \nu$.

The question boils down to searching for a valid transportation map $T : \mathcal{X} \to \mathcal{X}$ that will be optimal in some sense. Monge originally proposed to minimize the sum of all the displacements, that is:

$$\inf_{\substack{T:\mathcal{X}\to\mathcal{X} \\ \text{s.t. } T_\sharp\mu=\nu}} \int_{\mathcal{X}} \|x - T(x)\| \, d\mu(x). \tag{MP}$$

We can further generalize this formulation. First, we can consider that the ground space is $\mathcal{X} = \mathbb{R}^d$, $d \geq 1$, with no complication. Then, instead of minimizing the sum of all the displacements, we can consider a more general objective defined by $\int_{\mathcal{X}} c(x, T(x)) \, d\mu(x)$ where $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is continuous: $c(x, y)$ represents the cost of moving one unit of mass from a point $x \in \mathcal{X}$ to a point $y \in \mathcal{X}$. Note that even more general settings can be considered (*e.g.* we can take a general Polish space for $\mathcal{X}$).

One tricky problem with the formulation of Monge is the existence of a minimizer in (MP). In fact, even before that, given two distributions $\mu, \nu \in \mathscr{P}(\mathcal{X})$, the existence of a valid transportation map sending μ to ν is not always guaranteed. Indeed, take the case where $\mu = \delta_x$ is a Dirac mass located at some $x \in \mathcal{X}$ and $\nu \in \mathscr{P}(\mathcal{X})$ is any probability measure but not a Dirac mass. Then for any $T : \mathcal{X} \to \mathcal{X}$, $T_\sharp \mu = \delta_{T(x)}$ is a Dirac mass and therefore cannot be equal to ν. These problems of existence led to a relaxation of the Monge problem.

1.1.2 Kantorovich Problem

The existence problems in the formulation of Monge basically stem from the fact that the mass that initially stands at some point $x \in \mathcal{X}$ should all be sent to the same location $T(x)$. This condition was relaxed by Kantorovich [1942], who proposed another mathematical formulation of the optimal transport problem: the mass lying initially at some point $x \in \mathcal{X}$ will now be allowed to be split into

several (possibly infinitely many) target locations. Instead of a *transportation map*, we will consider a *transportation plan*, *i.e.* a probability distribution $\pi \in \mathscr{P}(\mathcal{X}^2)$ which is a coupling of μ and ν: $\pi(A, B)$ represents the amount of mass being transported from a Borel set $A \subset \mathcal{X}$ to a Borel set $B \subset \mathcal{X}$. We also have to ensure that the total amount of mass $\int_{x\in A}\int_{y\in\mathcal{X}} d\pi(x,y)$ being sent from $A \subset \mathcal{X}$ is equal to the initial amount of mass $\mu(A)$ over A, and likewise, that the total amount of mass $\int_{y\in B}\int_{x\in\mathcal{X}} d\pi(x,y)$ arriving to $B \subset \mathcal{X}$ is equal to the target amount of mass $\nu(B)$ over B. This means that π has marginals μ and ν, which we will denote by $\pi \in \Pi(\mu, \nu)$. Finally, we can formulate the Kantorovich [1942] problem:

$$\mathscr{T}_c(\mu,\nu) \stackrel{\text{def}}{=} \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}^2} c(x,y)\, d\pi(x,y). \tag{KP}$$

Existence of a minimizer in (KP) holds in very general settings, and in particular as soon as $\mathcal{X} = \mathbb{R}^d$ and c is continuous [Santambrogio, 2015, Theorem 1.7].

A link between the problems of Monge and Kantorovich A natural question is the one of understanding when (MP) and (KP) are in fact equivalent. It could seem at first that as soon as the Monge problem admits a minimizer $T_\star$ such that $\int_{\mathcal{X}} c\,(x, T_\star(x))\, d\mu(x) < +\infty$, it should automatically provide a solution $\pi = (\mathrm{Id}, T)_\sharp\mu$ to (KP). As noted in [Lacombe, 2020, Remark 2.18], this does not hold in all generality. When μ is atomless though, even if the infimum in (MP) may not be attained, Pratelli [2007] proved that the values of (MP) and (KP) are equal.

Wasserstein distances A special case of problem (KP) is when c is the power of the Euclidean distance, *i.e.* $c(x, y) = \|x - y\|^p$ for some $p \geq 1$. In this case, $\mathscr{T}_c^{1/p}$ defines a distance over

$$\mathscr{P}_p(\mathcal{X}) \stackrel{\text{def}}{=} \left\{\mu \in \mathscr{P}(\mathcal{X}), \int_{\mathcal{X}} \|x\|^p\, d\mu(x) < \infty\right\},$$

see *e.g.* [Santambrogio, 2015, Proposition 5.1]. This distance is called the *Wasserstein* distance (of order p), or p-Wasserstein distance, and we will write $W_p \stackrel{\text{def}}{=} \mathscr{T}_c^{1/p}$. The metric space $(\mathscr{P}_p(\mathcal{X}), W_p)$ is a geodesic space: for any $\mu, \nu \in \mathscr{P}_p(\mathcal{X})$, there exists a path connecting μ and ν with length $W_p(\mu, \nu)$, see [Santambrogio, 2015, Theorem 5.27].

Barycentric projection For a transportation plan $\pi \in \Pi(\mu, \nu)$ and a cost function c, the barycentric projection of π is the map $\bar{\pi} : \mathcal{X} \to \mathcal{X}$ defined by:

$$\bar{\pi}(x) = \arg\min_{z\in\mathcal{X}} \mathbb{E}_{(X,Y)\sim\pi}\left[c(z, Y)|X = x\right].$$

In particular when $\mathcal{X} = \mathbb{R}^d$ and $c(x, y) = \|x - y\|^2$:

$$\bar{\pi}(x) = \mathbb{E}_{(X,Y)\sim\pi}\left[Y|X = x\right].$$

In this quadratic case, if π is an optimal transport plan in (KP), then $\bar{\pi}$ is an optimal transport map sending μ to $\bar{\pi}_\sharp\mu$ [Ambrosio et al., 2006, Theorem 12.4.4].

1.1.3 Some Useful Properties

Duality Problem (KP) has a linear objective and has linear equality and inequality constraints, hence is a linear program. An important tool to study linear programs is duality. The Kantorovich problem admits the following dual formulation [Santambrogio, 2015, Proposition 1.11 & Theorem 1.39]:

$$\mathscr{T}_c(\mu, \nu) = \max_{\substack{\phi,\psi\in\mathcal{C}(\mathcal{X}) \\ \phi\oplus\psi\leq c}} \int_{\mathcal{X}} \phi \, d\mu + \int_{\mathcal{X}} \psi \, d\nu. \tag{Dual-KP}$$

Brenier theorem In the 2-Wasserstein setting, when $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ and μ has a density with respect to the Lebesgue measure $\mathscr{L}^d$, there exists a unique solution $T_\star$ to (MP). We will say that $T_\star$ is the optimal transport map, or the Monge map. Moreover, Brenier [1987] proved that this map is the gradient of a convex function $u_\star : \mathbb{R}^d \to \mathbb{R}$: $T_\star = \nabla u_\star$. We will say that $u_\star$ is an optimal Brenier potential of the problem.

Regularity of the Monge map When there exists an optimal Monge map $T_\star$ sending one measure μ to ν, *e.g.* when μ has a density with respect to the Lebesgue measure and the ground-cost function is the squared Euclidean distance, one question is to determine the regularity of $T_\star$. More precisely, given some assumptions on μ and ν which guarantee the existence of $T_\star$, what can be said about its regularity?

Let us first remark that no regularity can be expected in all generality, calling for specific assumptions on μ and ν. Consider the uniform measure μ over the ball $B(0, 1) \subset \mathbb{R}^d$ and $\nu = T_\sharp\mu$ where $T : \mathbb{R}^d \to \mathbb{R}^d$ is defined by $T(x) = (x_1 + 2\,\mathrm{sign}(x_1), x_2, \ldots, x_d)$ (see Figure 1.1 above). Since T is the subgradient of the convex potential $u : x \mapsto \frac{1}{2}\|x\|^2 + 2|x_1|$, the Brenier theorem implies it is the optimal Monge map sending μ to ν. But T is discontinuous on the hyperplane of equation $x_1 = 0$. As noticed by Caffarelli [1992], one key ingredient to obtain regularity of the map is the *convexity* of the support of ν.

Let us now place ourselves in the quadratic case, and suppose that μ and ν have respective densities f and g with respect to the Lebesgue measure. The existence of the optimal Monge map $T_\star$ is guaranteed by the Brenier theorem and $T_\star = \nabla u_\star$ where $u_\star$ is convex. In this case, the condition that $T_\sharp\mu = \nu$

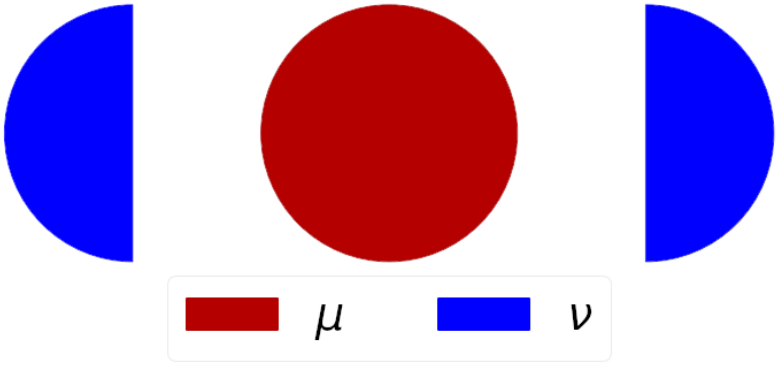


Figure 1.1: Two measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^2)$ such that the optimal Monge map sending μ onto ν is discontinuous.

in (MP) can be written as

$$\det \nabla T(x) = \frac{f(x)}{g\left(T(x)\right)}$$

and by plugging the Brenier potential $T = \nabla u$:

$$\det \nabla^2 u(x) = \frac{f(x)}{g\left(\nabla u(x)\right)}$$

which is known as (a particular case of the more general) *Monge-Ampère equation*. This equation is highly non-linear, which makes its study more intricate. One observation though is that the regularity of the Brenier potential u (hence the regularity of the Monge map $T = \nabla u$) will depend on the regularity of the densities f and g.

A first regularity result given by Caffarelli [2000] states that for $\mu \propto e^V \gamma_d$ and $\nu \propto e^{-W} \gamma_d$ where γ_d is the d-dimensional standard Gaussian measure and V, W are convex potentials, the associated Monge map is 1-Lispchitz, *i.e.* the associated convex Brenier potential is 1-smooth. Of course, weaker assumptions on μ and ν are possible, but the regularity of the Monge map will be weaker too. Classical theorems assume the convexity of the support of ν and that the densities are bounded away from zero and infinity. In this case, some Hölder regularity on u can be proved, see *e.g.* [Figalli, 2017, Theorem 4.23] for a precise statement. General results about the regularity of the Monge maps and the Monge-Ampère equation can be found in [Philippis, 2013, Figalli, 2017].

1.1.4 Computational Aspects

In this subsection, we mainly consider the so-called "discrete case", that is the case where μ and ν have finite discrete supports:

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

where $x_1, \dots, x_n, y_1, \dots, y_m \in \mathcal{X} = \mathbb{R}^d$ and a, b are probability weights.

We introduce the cost matrix $C = [c(x_i, y_j)]_{1 \leq i \leq n, 1 \leq j \leq m}$. Then the Kantorovich problem (KP) writes:

$$\min_{P \in \mathcal{U}(a,b)} \langle C, P \rangle \tag{Discrete-OT}$$

where $\mathcal{U}(a,b) = \left\{ P \in \mathbb{R}_+^{n \times m}, P\mathbf{1}_m = a, P^\top \mathbf{1}_n = b \right\}$ is the set of admissible transport plans. Problem (Discrete-OT) is a linear program and can be solved by the network simplex algorithm in $\mathcal{O}\left((n+m)nm \log(n+m)\right)$, see *e.g.* [Ahuja et al., 1988]. This means that OT in a discrete setting can be computed exactly, at the cost of a cubic complexity which is prohibitive in large-scale machine learning settings. This algorithmic complexity has led to two different approaches: considering settings where the Wasserstein distances are available in closed-form expression, and introducing new variants of the classical OT problem that are more computationally tractable.

1.1.4.1 Closed-form solutions to the Monge problem

Bures-Wasserstein metric When $\mu = \mathcal{N}(a, A)$ and $\nu = \mathcal{N}(b, B)$ are Gaussian distributions, Dowson and Landau [1982], Olkin and Pukelsheim [1982], Givens et al. [1984] showed that the 2-Wasserstein distance between them is available in closed-form:

$$W_2^2(\mu, \nu) = \|a - b\|^2 + \mathfrak{B}^2(A, B)$$

where $\mathfrak{B}^2(A,B) \stackrel{\text{def}}{=} \operatorname{trace} A + \operatorname{trace} B - 2\operatorname{trace}(A^{1/2} B A^{1/2})^{1/2}$ is the Bures [1969] metric between positive semi-definite (PSD) matrices. This formula holds in the more general setting where the measures belong to the same family of elliptically-contoured distributions, *i.e.* densities with elliptical level sets, with a, b being the mean vectors and A, B the covariance matrices of the measures [Gelbrich, 1990]. The Monge map has the following closed-form affine expression:

$$T : x \mapsto b + A^{-1/2} \left(A^{1/2} B A^{1/2} \right)^{1/2} A^{-1/2} (x - a).$$

Univariate case In dimension $d = 1$, if the cost function is of the form $c(x, y) = h(x - y)$, $h : \mathbb{R} \to \mathbb{R}$ being a convex function, the following map is optimal in the Monge problem (see [Santambrogio, 2015, Theorem 2.9]):

$$T : x \mapsto F_\mu^{\leftarrow} \circ F_\nu(x)$$

where $F_\mu^{\leftarrow}$ is the generalized inverse of the cumulative distribution function (c.d.f) F_μ of μ. In the discrete setting, the c.d.f of μ and ν are piecewise constant, so T can be computed exactly in a mere $\mathcal{O}(n \log(n) + m \log(m))$ time.

1.1.4.2 Variants of optimal transport

Sliced Wasserstein distance The one dimensional case inspired an approximation of the Wasserstein distance in $\mathbb{R}^d$: the Sliced Wasserstein (SW) distance defined by Rabin et al. [2011]. First, project the measures $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$ over a line $\theta \in \mathbb{S}^{d-1}$, then compute the Wasserstein distance W_p between these one dimensional projections, and finally average this value over all directions $\theta \in \mathbb{S}^{d-1}$:

$$SW_p(\mu, \nu) \overset{\text{def}}{=} \int_{\mathbb{S}^{d-1}} W_p\left(P_{\theta\sharp}\mu, P_{\theta\sharp}\nu\right) d\theta,$$

where P_θ is the orthogonal projection onto the line of direction $\theta \in \mathbb{S}^{d-1}$. In practice, the integral over the sphere $\mathbb{S}^{d-1}$ is approximated by Monte-Carlo integration by randomly sampling different directions on the sphere before averaging the associated Wasserstein values.

Entropy-regularized optimal transport Another approach to deal with the computational difficulties of (Discrete-OT) relies in its regularization. Cuturi [2013] proposed to add a small entropic penalty to its objective

$$\min_{P \in \mathcal{U}(a,b)} \langle C, P \rangle + \gamma \sum_{i,j} P_{ij} \log(P_{ij}) \qquad \text{(Discrete-Entropic-OT)}$$

where $\gamma > 0$ is the regularization strength. Note that we could equivalently minimize

$$\langle C, P \rangle + \gamma \sum_{i,j} P_{ij} \log\left(\frac{P_{ij}}{a_i b_j}\right) = \langle C, P \rangle + \gamma \, \mathrm{KL}(P || a \otimes b)$$

where $(a \otimes b)_{ij} = a_i b_j$ and KL is the Kullback-Leibler divergence. Indeed:

$$\mathrm{KL}(P||a \otimes b) = \sum_{i,j} P_{ij} \log\left(\frac{P_{ij}}{a_i b_j}\right) = \sum_{i,j} P_{ij} \log P_{ij} - \sum_i a_i \log a_i - \sum_j b_j \log b_j$$

and the two last sums do not depend on the transport plan, hence play no role in the minimization.

In the general Kantorovich problem (KP), entropy-regularized optimal transport writes:

$$\mathscr{S}_c^\gamma(\mu,\nu) \stackrel{\text{def}}{=} \min_{\pi\in\Pi(\mu,\nu)} \int c\, d\pi + \gamma\, \mathrm{KL}(\pi||\mu\otimes\nu)$$

where $\mu\otimes\nu$ is the independent coupling of μ and ν.

As shown in [Peyré and Cuturi, 2019, Proposition 4.3], (Discrete-Entropic-OT) admits a unique solution of the form $P = \mathrm{diag}(u)K\,\mathrm{diag}(v)$ where $K = \left[e^{-C_{ij}/\gamma}\right]_{ij} \in \mathbb{R}_+^{n\times m}$ and $(u,v) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$ verify the following fixed-point iterations (the operations are taken elementwise):

$$u = a \oslash (Kv) \quad \text{and} \quad v = b \oslash (K^\top u).$$

Alternating between solving these equations for u and v defines the Sinkhorn [1964] algorithm.

1.2 Outline and contributions of this thesis

1.2.1 OT and the curse of dimensionality

I started my PhD studies in September 2018. Back then, one critical problem in the estimation of Wasserstein distances was the curse of dimensionality it suffers from. It is know since [Dudley, 1969] that given a probability distribution $\mu \in \mathscr{P}_p(\mathbb{R}^d)$ and its empirical version $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ where $x_1, \dots, x_n \sim \mu$ are i.i.d samples from μ, $\mathbb{E}[W_p(\hat{\mu}_n, \mu)] = \mathcal{O}\left(n^{-1/d}\right)$. Non-asymptotic bounds and concentration results have also been proved by Fournier and Guillin [2015]. In [Weed and Bach, 2019], whose preprint was first published in 2017, the authors show improved asymptotic rates when the support of the measure μ is a low-dimensional manifold: in this case, $\mathbb{E}[W_p(\hat{\mu}_n, \mu)] = \mathcal{O}\left(n^{-1/k}\right)$ where k is the dimension of the manifold $\mathrm{supp}(\mu)$. While it could seem to break the curse of dimensionality, the dimension of $\mathrm{supp}(\mu)$ cannot always be supposed to be low-dimensional: even if it is of dimension as low as $k = 10$ in a space of dimension $d = 10^6$ (think of images), a rate of $\mathcal{O}\left(n^{-1/10}\right)$ is still very slow. Note that using the triangular inequality for W_p, all these results imply similar bounds on $\mathbb{E}\left[|W_p(\hat{\mu}_n, \hat{\nu}_n) - W_p(\mu, \nu)|\right]$ where $\nu \in \mathscr{P}_p(\mathbb{R}^d)$ and $\hat{\nu}_n$ is its empirical version. Finally, Genevay et al. [2019] showed at the beginning of my research studies that entropy-regularized optimal transport mitigates the curse of dimensionality: for a Lipschitz cost function c and at fixed regularization strength $\gamma > 0$, $\mathbb{E}\left[|\mathscr{S}_c^\gamma(\hat{\mu}_n, \hat{\nu}_n) - \mathscr{S}_c^\gamma(\mu, \nu)|\right] = \mathcal{O}\left(1/\sqrt{n}\right)$. Yet, this result deteriorates badly as $\gamma \to 0$, and holds for any measures μ, ν over a compact set: the potentially low-dimensional structures of the measures would not be taken into account even when appropriate. The question of finding algorithmic approaches that could mitigate this curse of dimensionality while taking advantage of the low-dimensional structures hidden in the measures was therefore still partly open when I started the PhD.

1.2.2 Robustness through projections

One approach though would exploit the low-dimensional structures of the measures: the Sliced Wasserstein (SW) distance introduced by Rabin et al. [2011]. The main reason which led to the definition of the SW distance is an algorithmic one, since computing the Wasserstein distance between one-dimensional projections of the measures has a log-linear complexity, in stark contrast with the cubic complexity of (Discrete-OT). While benefiting on the algorithmic side, it seemed that projecting measures with high-dimensional supports onto one-dimensional lines would cause to lose a large part of the information contained in these measures, or at least that the algorithmic-driven choice of lines was a bit ad hoc. Even worse, when averaging the values $W_p\left(P_{\theta\sharp}\mu, P_{\theta\sharp}\nu\right)$ over the sphere $\theta \in \mathbb{S}^{d-1}$, most of the projections do not discriminate well between the measures and are therefore polluting the average

value. It thus appeared clear that projecting measures onto low-dimensional subspaces instead of lines, and selecting such subspaces by optimizing some criterion, should provide interesting research directions. In Principal Component Analysis (PCA) [Pearson, 1901], the data is projected onto the low-dimensional subspace that maximizes the variance of the projected points. Translating this fact into the setting of measures and optimal transport was the key to defining robust variants to the Wasserstein distances that would make use of the low-dimensional structures of the measures.

This very problem is naturally rewritten as the maximization of the Wasserstein distance with respect to the ground-cost function. Choosing which ground-cost function to use when computing optimal transport is in fact an important task: as noticed by Genevay et al. [2018, Section 3], learning the cost function in an adversarial way is crucial when computing optimal transport in high-dimensional spaces. But how to interpret such an adversarial choice? In our case, the link with the low-dimensional subspaces makes it clearer: the problem can naturally be reinterpreted as an optimal transport problem where the linear objective is replaced by a convex spectral objective, in the flavour of PCA where the optimal subspace can be retrieved by the eigendecomposition of the covariance matrix. This brought to light what seemed at first a feeble link between maximizing the Wasserstein distance with respect to the ground-cost function on the one hand, and convexifying the classical optimal transport objective on the other. Elucidating this link became my new research direction, and has proved fruitful: any convexification of the traditional optimal transport problem can be recast as a ground-cost maximization problem using Legendre-Fenchel duality. This, in turns, means that convex regularizations of optimal transport, and in particular entropy-regularized optimal transport, are robust to the ground-cost function. This shed new light on why regularization works in practice: regularizing automatically implies a robust choice of the ground-cost function, and conversely, maximizing over the ground-cost function as in [Genevay et al., 2018] is an implicit regularization of optimal transport.

1.2.3 On the difficulty of estimating Monge maps

While the curse of dimensionality and the robust choice of a ground-cost function are about the optimal transport *value*, some applications are about the optimal transport *map*. For example, Courty et al. [2016] align two datasets by optimally transporting one onto the other; in [Schiebinger et al., 2019], the authors use optimal transport maps to track the evolution of a cell colony when only snapshots are available. In these applications, having access to a map instead of a plan is crucial: one point should be transported to one unique destination. Even more so is the need for that map to be defined on the whole space, to allow for generalization outside the data. Unfortunately, the Monge problem does not always admit a minimizer, and in fact, no map can transport a discrete measure to another with a larger support size. A first approach

proposed by Courty et al. [2016] is to compute an optimal transport plan and deduce a map from it by considering its barycentric projection, which is known to optimally transport the initial measure onto its image for the quadratic cost function [Ambrosio et al., 2006, Theorem 12.4.4]. In this case, the obtained map is only defined on the data and cannot generalize to new data points. To define a map on the whole space, Perrot et al. [2016] proposed to learn it from a parametric family (linear functions, or non-linear functions using kernels). But how to ensure it induces an optimal transport? The authors proposed to jointly learn a plan and a map, such that the plan is approximately minimizing the optimal transport cost and the map is approximately its barycentric projection. This map is therefore not an optimal transport map, but only approximately one. Two years later, Seguy et al. [2018] continued this line of work by proposing to model the map by a deep neural network. Again, the optimality of the transport induced by this map cannot be guaranteed.

1.2.4 Monge maps using convexity and regularity

How can those two specifications on the map—inducing a transport which is optimal and being defined on the whole space—be reconciliated? As no answer was available when I started the PhD, I decided to follow this research lead. Instead of defining a map from a transport plan using the barycentric projection, I started looking directly for a map that would be optimal *by design*. In the quadratic Wasserstein case, the Brenier [1987] theorem gives such an optimality condition: a map $T : \mathbb{R}^d \to \mathbb{R}^d$ is optimal if, and only if, it is the gradient of a convex potential. Reframing the mapping estimation procedure in terms of convex Brenier potentials guarantees the optimality of the learnt map. Going one step further, we know by the regularity theory of optimal transport that we can expect at least some regularity on the Monge map. Adding such regularity prior during the estimation should help the generalization: by constraining the map to be bi-Lipschitz (*i.e.* the Brenier potential to be strongly convex and smooth), similar data points should be mapped to similar locations. The question remains as to how such strongly convex and smooth potentials—and their gradients—can generalize to the whole space. To overcome this difficulty, the convex interpolation literature (see [Taylor et al., 2017, Section 3] for an overview on the subject) comes in handy: it provides algorithmic methods that can recover a smooth and strongly convex function $u : \mathbb{R}^d \to \mathbb{R}$ with prescribed values $u(x_i)$ and/or values of its gradient $\nabla u(x_i)$ on a finite number of points $x_1, \dots, x_n \in \mathbb{R}^d$. Combining the approach *à la Brenier* with these methods allow for an estimation procedure that results in an optimal transport map defined on the whole space and which is bi-Lipschitz.

1.2.5 Contributions

The main objective of this thesis is to contribute to the understanding of regularized optimal transport, as well as proposing novel regularized versions of optimal transport distances. This thesis is divided into two thematic parts:

- **Part I: Ground-Cost Robustness**
  Optimal transport is highly-dependent on the choice of the ground-cost function [Genevay et al., 2018, Section 3]. In practice, it is often chosen to be the Euclidean distance or the squared Euclidean distance, leading to the famous 1- and 2-Wasserstein distances respectively. But Wasserstein distances suffer from the curse of dimensionality [Dudley, 1969, Weed and Bach, 2019], calling for some sort of regularization. In this part, we propose to maximize the optimal transport problem over the ground-cost function inside some class of functions. In chapter 2, the class of admissible ground-costs is the set of Mahalanobis distances, giving rise to a robust variant of the 2-Wasserstein distance which we show to have connections with dimensionality reduction. In chapter 3, we prove that maximizing the optimal transport problem over the ground-cost function actually corresponds to regularizing optimal transport, generalizing a result of chapter 2.

- **Part II: Regularity-Constrained Maps**
  In different applications, having access to the optimal transport map sending one measure onto another is crucial [Courty et al., 2016, Schiebinger et al., 2019]. In practice though, when only a finite number of data is known, such a map does not necessarily exists or is only defined on the data points. In order to extend such a map to the whole space, we propose to learn a map that is optimal *by design* and that is constrained to be bi-Lipschitz, allowing for generalization out of the data samples.

We now dive into more details about each chapter of the thesis: for each one, we outline the related works and our original contributions.

Chapter 2: Subspace Robust Wasserstein Distances

This chapter is based on [Paty and Cuturi, 2019].

Optimal transport suffers from the curse of dimensionality [Dudley, 1969, Weed and Bach, 2019]. Classical approaches to this problem involve projections on random real lines [Rabin et al., 2011] or a preliminary dimensionality reduction of the measures [Schiebinger et al., 2019]. In this chapter, we propose to perform dimensionality reduction and optimal transport at the same time. This leads to a robust variant of the quadratic Wasserstein distance, that we call the *subspace robust Wasserstein distance*.

Related work Wasserstein distances suffer from the curse of dimensionality: [Dudley, 1969, Fournier and Guillin, 2015] show that their sample complexity grows exponentially with the dimension, meaning that a huge amount of data is needed as soon as the data is high-dimensional. Although these results can be mitigated when the data lives on a low-dimensional manifold [Weed and Bach, 2019], this assumption is not always met nor sufficient in practice. A line of work initiated by Cuturi [2013] advocates adding a small entropic penalty to the original OT problem, which results in improved sample complexity bounds [Genevay et al., 2019], but does not take into account the low-dimensional structures of the data. Another approach exploits the fact that computing Wasserstein distances between two distributions on the real line boils down to the direct comparison of their generalized quantile functions [Santambrogio, 2015, Theorem 2.9]. Computing quantile functions only requires sorting values, with a mere log-linear complexity. The *sliced Wasserstein distance* [Rabin et al., 2011] consists in computing the expected Wasserstein distance between the projections of the two distributions onto a line drawn uniformly at random on the sphere. Even simpler is the approach used in some applications of optimal transport to real data, *e.g.* in [Schiebinger et al., 2019]: practitioners first reduce the dimensionality of the data using *principal component analysis* (PCA) [Pearson, 1901] before applying optimal transport techniques.

Contributions The main contribution of this chapter is the introduction and initial study of new optimal transport distances that we call the *projection robust Wasserstein distances* and the *subspace robust Wasserstein distance*.

1. **Maximizing the 2-Wasserstein distance between the projections of the measures:** Given two probability distributions $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, we propose to project them onto a k-dimensional subspace E of $\mathbb{R}^d$ and compute the 2-Wasserstein distance between the projections. We then select the subspace E in an adversarial way:

$$\mathcal{P}_k(\mu, \nu) \stackrel{\text{def}}{=} \sup_{E \in \mathcal{G}_k} W_2\left(P_{E\sharp}\mu, P_{E\sharp}\nu\right)$$

   where $\mathcal{G}_k = \left\{E \text{ is a subspace of } \mathbb{R}^d, \dim(E) = k\right\}$ is the Grassmannian and P_E is the orthogonal projector onto $E \in \mathcal{G}_k$.

   a) We show that $\mathcal{P}_k$ defines a distance over $\mathscr{P}_2(\mathbb{R}^d)$.

   b) We show that an analog definition with the p-Wasserstein distance W_p, $p \geq 1$, instead of W_2 also defines a distance over $\mathscr{P}_p(\mathbb{R}^d)$.

   c) The maximization problem in the definition of $\mathcal{P}_k$ is not convex because of the non-convexity of the Grassmannian $\mathcal{G}_k$, so it seems difficult to compute or even approximate it.

2. **Convex relaxation of $\mathcal{P}_k$ and its theoretical study:** We therefore propose to relax the definition of $\mathcal{P}_k$. To do so, we swap the supremum (over the Grassmannian) and the infimum (over the transportation plans) in the definition of $\mathcal{P}_k$ to define the *subspace robust Wasserstein distance*:

$$\mathcal{S}_k(\mu,\nu) \stackrel{\text{def}}{=} \inf_{\pi\in\Pi(\mu,\nu)} \sup_{E\in\mathcal{G}_k} \left(\int \|P_E(x)-P_E(y)\|^2 \, d\pi(x,y)\right)^{1/2}.$$

   a) We show that $\mathcal{S}_k$ defines a geodesic distance over $\mathscr{P}_2(\mathbb{R}^d)$ and that it is strongly equivalent to W_2. We give tight constants for the metric equivalence inequalities, and we exhibit explicit geodesics.

   b) We show that $\mathcal{S}_k$ is indeed a convex relaxation of $\mathcal{P}_k$:

$$\mathcal{S}_k(\mu,\nu) = \max_{\substack{0\preceq\Omega\preceq I\\ \operatorname{trace}(\Omega)=k}} W_2\left(\Omega^{1/2}{}_\sharp\mu, \Omega^{1/2}{}_\sharp\nu\right).$$

   The set $\mathcal{R} = \left\{\Omega\in\mathbb{R}^{d\times d}, 0\preceq\Omega\preceq I\,,\, \operatorname{trace}(\Omega)=k\right\}$ is in fact the convex hull of the set of orthogonal projection matrices of rank k.

   c) We show that $\mathcal{S}_k$ admits an alternative formulation in terms of the second order moment matrix of the displacements, clarifying the link with PCA:

$$\mathcal{S}_k^2(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \sum_{l=1}^{k} \lambda_l(V_\pi)$$

   where $V_\pi \stackrel{\text{def}}{=} \iint (x-y)(x-y)^\top \, d\pi(x,y)$ and λ_l is the l^{th} largest eigenvalue.

   d) We prove that $\mathcal{S}_k$ can be reinterpreted as the maximization of the optimal transport cost $\mathscr{T}_c$ over the ground-cost function c:

$$\mathcal{S}_k^2(\mu,\nu) = \max_{\substack{0\preceq\Omega\preceq I\\ \operatorname{trace}(\Omega)=k}} \mathscr{T}_{d_\Omega^2}(\mu,\nu)$$

   where d_Ω^2 stands for the squared Mahalanobis distance:

$$d_\Omega^2(x,y) = (x-y)^\top\Omega(x-y).$$

3. **Numerical tools to compute $\mathcal{S}_k$:** When μ,ν are discrete measures, $\mathcal{S}_k$ defines a finite-dimensional convex optimization problem. We propose two algorithms to solve it:

   a) *Projected super-gradient ascent:* we consider the maximization over $\Omega\in\mathcal{R}$ of the concave function $f:\Omega\mapsto\mathscr{T}_{d_\Omega^2}(\mu,\nu)$ which admits the following superdifferential:

$$\partial f(\Omega) = \operatorname{conv}\left\{V_{\pi_\star}, \pi_\star\in\operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)}\langle\Omega, V_\pi\rangle\right\}.$$

   The algorithm follows the following iterations until convergence:

i. Compute $\pi_\star$ by solving the OT problem (at fixed Ω);
ii. Take a supergradient step: $\Omega \leftarrow \Omega + \varepsilon V_{\pi_\star}$ where $\varepsilon > 0$ is the step-size;
iii. Project Ω onto the constraint set $\mathcal{R}$.

b) *Frank-Wolfe algorithm with entropic regularization:* in order to speed the OT computations up, we add an entropic regularization to the objective:

$$f_\gamma : \Omega \mapsto \mathscr{S}^\gamma_{d^2_\Omega}(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int d^2_\Omega \, d\pi + \gamma \, \mathrm{KL}(\pi || \mu \otimes \nu),$$

where $\gamma > 0$ is the regularization strength. This also makes the objective f_γ smooth, allowing for Frank-Wolfe iterations instead of gradient ascent. Frank-Wolfe iterations do not need projections onto $\mathcal{R}$ and do not require adjusting a step-size. The algorithm goes as follow:

i. We solve, at fixed Ω, the entropic OT problem

$$\min_{\pi \in \Pi(\mu,\nu)} \int d^2_\Omega \, d\pi + \gamma \, \mathrm{KL}(\pi || \mu \otimes \nu)$$

using Sinkhorn algorithm;
ii. We solve, at fixed π, the problem

$$\max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega) = k}} \int d^2_\Omega \, d\pi$$

whose maximizer is given by $\widehat{\Omega} = U \, \mathrm{diag}\left([\mathbf{1}_k, \mathbf{0}_{d-k}]\right) U^\top$, using the eigendecomposition of $V_\pi = U \, \mathrm{diag}\left(\lambda_1, ..., \lambda_d\right) U^\top$;
iii. We run the Frank-Wolfe update: $\Omega \leftarrow (1 - \tau)\Omega + \tau\widehat{\Omega}$ where $\tau = 2/(2 + n)$ at the n^{th} iteration.

Chapter 3: Regularized Optimal Transport is Ground-Cost Adversarial

This chapter is based on [Paty and Cuturi, 2020].

Regularizing the optimal transport problem has proven crucial for the optimal transport theory to impact the field of machine learning. For instance, it is known that regularizing OT problems with entropy leads to faster computations and better differentiation using the Sinkhorn algorithm [Cuturi, 2013], as well as better sample complexity bounds than classic OT [Genevay et al., 2019]. In this chapter, we depart from this practical perspective and propose a new interpretation of regularization as a robust mechanism, and show using Fenchel duality that any convex regularization of optimal transport can be interpreted as ground-cost adversarial.

Related work The key to using optimal transport in applications lies in the different forms of regularization of the original optimal transport problem. Adding a small convex regularization to the classical linear cost not only helps on the algorithmic side, by convexifying the objective and allowing for faster solvers, but also introduces a regularity trade-off that prevents from overfitting on data measures. Although entropy-regularized OT is the most studied regularization of OT, due to its algorithmic advantages [Cuturi, 2013], several other convex regularizations of the transport plan have been proposed in the community: quadratically-regularized OT [Essid and Solomon, 2017], OT with capacity constraints [Korman and McCann, 2015], Group-Lasso regularized OT [Courty et al., 2016], OT with Laplacian regularization [Flamary et al., 2014], Tsallis Regularized OT [Muzellec et al., 2017], among others. On the other hand, regularizing the dual Kantorovich problem was shown in [Liero et al., 2018] to be equivalent to unbalanced OT, that is optimal transport with relaxed marginal constraints. The question of understanding why regularizing OT proves critical has triggered several approaches. A compelling reason is statistical: although classical OT suffers from the curse of dimensionality, as its empirical version converges at a rate of order $\mathcal{O}\left(n^{-1/d}\right)$ [Dudley, 1969, Fournier and Guillin, 2015], regularized OT and more precisely Sinkhorn divergences [Genevay et al., 2018] have a sample complexity of $O(1/\sqrt{n})$ [Genevay et al., 2019, Mena and Niles-Weed, 2019]. Entropy-regularized OT was also shown to perform maximum likelihood estimation in the Gaussian deconvolution model [Rigollet and Weed, 2018]. Taking another approach, Dessein et al. [2018], Blondel et al. [2018] have considered general classes of convex regularizations and characterized them from a more geometrical perspective. Recently, several papers [Genevay et al., 2018, Flamary et al., 2018, Deshpande et al., 2019, Kolouri et al., 2019, Niles-Weed and Rigollet, 2019, Paty and Cuturi, 2019] have proposed to maximize OT with respect to the ground-cost function, which can in turn be interpreted in light of ground metric learning [Cuturi and Avis, 2014]. This approach can also be viewed as an instance of robust optimization [Ben-Tal and Nemirovski, 1998, Ben-Tal et al., 2009, Bertsimas et al., 2011]: instead of considering a data-dependent, hence unstable minimization problem $\min_x f_{\hat{\theta}}(x)$ where $\hat{\theta}$ represents the data, the robust optimization literature adversarially chooses the parameters θ in a neighborhood of the data: $\max_{\theta\in\Theta} \min_x f_\theta(x)$.

Contributions Continuing along these lines, our main contribution in this chapter is to make a connection between *regularizing* and *maximizing* OT. Our main goal is to provide a novel interpretation of regularized optimal transport in terms of ground-cost robustness: regularizing OT amounts to maximizing **un**regularized OT with respect to the ground-cost function.

1. **We interpret convex regularizations of the transport plan as ground-cost robustness:** Here, $\mathcal{X}$ is a compact Hausdorff space. Let $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous (l.s.c) functional.

We show that:

$$\inf_{\pi\in\Pi(\mu,\nu)} F(\pi) = \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - F^*(c),$$

where $F^* : c \mapsto \sup_{\pi\in\mathscr{M}(\mathcal{X}^2)} \int c\, d\pi - F(\pi)$ is the Legendre-Fenchel conjugate of F. This means that minimizing a convex functional of the transport plan corresponds to the maximization of the classical linear optimal transport value over the ground-cost function, with a penalization on the ground-cost function. An important reformulation of this fact is the following: if $R : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R}\cup\{+\infty\}$ is l.s.c, $c_0 : \mathcal{X}^2 \to \mathbb{R}_+$ is continuous and $\varepsilon > 0$, then:

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c_0\, d\pi + \varepsilon R(\pi) = \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \varepsilon R^*\left(\frac{c-c_0}{\varepsilon}\right),$$

where R^* is the Legendre-Fenchel dual of R. In particular, this formulation implies that regularizing OT boils down to maximizing classic OT with respect to the ground-cost function, with a regularization that somewhat forces the adversarial ground-cost function to be "close" to the original one c_0, in the sense that $R^*\left(\frac{c-c_0}{\varepsilon}\right)$ cannot be too large. The link between the regularization on the transport plan and the regularization on the ground-cost function is given by the Legendre-Fenchel transform of the regularizers.

2. **We investigate the properties of optimal adversarial ground-cost functions:** We prove, under some technical assumption (that is verified *e.g.* for entropic or quadratic regularizations of OT), a duality theorem for "regularized" OT:

$$\inf_{\pi\in\Pi(\mu,\nu)} F(\pi) = \sup_{\phi,\psi\in\mathcal{C}(\mathcal{X})} \int \phi\, d\mu + \int \psi\, d\nu - F^*(\phi\oplus\psi)$$

where $\phi\oplus\psi : (x,y) \mapsto \phi(x)+\psi(y)$. We use this duality result to show that under the same technical assumption, there exists an optimal adversarial ground-cost function that is separable, *i.e.* that there exists functions $\phi,\psi : \mathcal{X} \to \mathbb{R}$ such that $\phi\oplus\psi$ is optimal in the maximization problem $\sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - F^*(c)$. In some sense, this is disappointing since a separable ground-cost function does not yield an interesting geometry on the ground space that could help analyze the data.

Chapter 4: Smooth and Strongly-Convex Nearest Brenier Potentials

This chapter is based on [Paty et al., 2020].

One of the greatest achievements of the OT literature in recent years lies in regularity theory: Caffarelli [2000] showed that the OT map between two well behaved measures is Lipschitz, or equivalently when considering 2-Wasserstein distances, that Brenier convex potentials (whose gradient yields an optimal map) are smooth. Instead of considering regularity as a property that can be proved under suitable assumptions, we consider in this chapter regularity as a condition that must be enforced when estimating OT: we propose algorithms that can recover optimal transport maps with small distortion that best approach the transport constraint, so that the associated Brenier potentials are strongly convex and smooth.

Related work Learning an optimal Monge map is crucial in some applications [Courty et al., 2016, Schiebinger et al., 2019]. To do so, two main approaches have been considered in the literature:

- learn an optimal transport map from an optimal transport plan, by considering its barycentric projection [Courty et al., 2016]: although the obtained map is indeed an optimal transport map, it is only defined on the data and cannot generalize to new points;

- learn a transport map that is defined everywhere, by considering a parametric family [Perrot et al., 2016, Seguy et al., 2018]: in this case, such maps cannot be guaranteed to be optimal transport maps.

In order to reconciliate those two approaches, we turn to the optimal transport literature, which provides a rich theory about Monge maps: the regularity theory. This theory gives properties of the optimal Monge map pushing forward a measure μ onto ν with a small average cost. When that cost is the quadratic Euclidean distance, the Monge map is necessarily the gradient ∇f of a convex function f. This major result, known as the Brenier [1987] theorem, states that the OT problem between μ and ν is solved as soon as there exists a convex function f such that $\nabla f_\sharp \mu = \nu$. Estimating a Monge map by forcing it to be the gradient of a convex potential will therefore guarantee its optimality. In that context, regularity in OT is usually understood as the property that the map ∇f is *Lipschitz*, a seminal result due to Caffarelli [2000] who proved that the Monge map can be guaranteed to be 1-Lipschitz when transporting a "fatter than Gaussian" measure $\mu \propto e^V \gamma_d$ towards a "thinner than Gaussian" measure $\nu \propto e^{-W} \gamma_d$ (here γ_d is the Gaussian measure on $\mathbb{R}^d$, $\gamma_d \propto e^{-\|\cdot\|^2}$, and V, W are two convex potentials). Equivalently, this result can be stated as the fact that the Monge map is the gradient of a 1-smooth Brenier [1987] potential. Adding such kind of regularity constraints on the estimated Monge map will allow for generalization outside the data, by means of the results obtained in the convex interpolation literature [Taylor et al., 2017, Section 3].

Contributions The main contribution of this chapter is to translate the idea that the OT map between sufficiently well-behaved distributions should be regular into an estimation procedure.

1. **We define *smooth and strongly convex nearest-Brenier* (SSNB) potentials:** Given two probability measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, a L-smoooth and ℓ-strongly convex function f such that $\nabla f_\sharp \mu = \nu$ may not always exist. We relax this equality and look instead for a smooth strongly convex function f that minimizes the Wasserstein distance between $\nabla f_\sharp \mu$ and ν:

$$\min_{f \text{ is } L\text{-smooth, } \ell\text{-strongly convex}} W_2\left(\nabla f_\sharp \mu, \nu\right). \tag{SSNB}$$

   We call such potential *nearest-Brenier* because they provide the way to transport μ to a measure as close as possible to ν using a smooth and strongly convex potential.

2. **We show that such potentials can be computed when measures are discrete:** Even when μ, ν are discrete probability measures, problem (SSNB) is infinite-dimensional because of the minimization over the set of smooth and strongly convex functions. We show that this problem can actually be rewritten as a finite-dimensional separately-convex optimization problem: at fixed potential f, the problem is a discrete OT problem; at fixed transport plan, the minimization over f is a convex quadratically-constrained quadratic program (QCQP), which can be solved using on-the-shelf solvers.

3. **We remark that the one-dimensional case is easier to compute:** In the univariate case, we show that computing the nearest-Brenier potential is equivalent to solving a variant of the isotonic regression problem in which the map must be strongly increasing and Lipschitz. A projected gradient descent approach can be used to solve this problem efficiently.

4. **We show how to compute the SSNB potential and map on the whole space:** We exploit the solutions to both these optimization problems to extend the Brenier potential and Monge map at any point. We show this can be achieved by solving a convex QCQP for each new point.

1.3 Grandes lignes et contributions de cette thèse

Le transport optimal date de la fin du XVIIIe siècle, lorsque le mathématicien français Monge [1781] proposa de résoudre le problème des déblais et des remblais. Cependant, la formulation mathématique de Monge atteignit rapidement ses limites en tant qu'il fut impossible de prouver l'existence d'une solution à son problème. Ce n'est qu'après cent cinquante ans que le transport optimal vécut une seconde jeunesse, lorsque le mathématicien russe Kantorovich [1942] comprit quel serait le cadre adéquat qui permettrait l'étude du problème de Monge, donnant ainsi naissance à des outils fondamentaux en mathématiques pures et appliquées. Ces dernières années, le transport optimal a trouvé de nouvelles applications au sein des statistiques et de l'apprentissage automatique, en tant qu'outil d'analyse des données : en apprentissage supervisé [Frogner et al., 2015, Abadeh et al., 2015, Courty et al., 2016], en informatique graphique [Solomon et al., 2015, Bonneel et al., 2016], en imagerie [Rabin and Papadakis, 2015, Cuturi and Peyré, 2016], pour des modèles génératifs [Arjovsky et al., 2017, Salimans et al., 2018, Genevay et al., 2018], en biologie [Hashimoto et al., 2016, Schiebinger et al., 2019, Huizing et al., 2021] ou en traitement automatique du langage [Grave et al., 2019, Alaux et al., 2019].

1.3.1 Le fléau de la dimension en transport optimal

J'ai commencé ma thèse en septembre 2018. À ce moment-là, l'un des problèmes majeurs dans l'estimation des distances de Wasserstein était le fléau de la dimension qu'elles subissent. Il est connu depuis [Dudley, 1969] qu'étant données une mesure de probabilité $\mu \in \mathscr{P}_p(\mathbb{R}^d)$ et sa version empirique $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ où $x_1, \dots, x_n \sim \mu$ sont des échantillons i.i.d. de μ, $\mathbb{E}[W_p(\hat{\mu}_n, \mu)] = \mathcal{O}\left(n^{-1/d}\right)$. Des bornes non-asymptotiques et des résultats de concentration ont aussi été prouvés par Fournier and Guillin [2015]. Dans [Weed and Bach, 2019], dont la première version a été publiée en 2017, les auteurs prouvent des bornes asymptotiques améliorées lorsque le support de la mesure μ est une variété de faible dimension : dans ce cas, $\mathbb{E}[W_p(\hat{\mu}_n, \mu)] = \mathcal{O}\left(n^{-1/k}\right)$ où k est la dimension de la variété $\mathrm{supp}(\mu)$. Bien que ce résultat semble résoudre la question du fléau de la dimension, on ne peut pas toujours supposer que la dimension de $\mathrm{supp}(\mu)$ est faible : même si elle est de dimension aussi faible que $k = 10$ dans un espace de dimension $d = 10^6$ (penser à des images), une vitesse de $\mathcal{O}\left(n^{-1/10}\right)$ reste très lente. Notons que l'inégalité triangulaire pour W_p nous permet de prouver des bornes similaires sur la quantité $\mathbb{E}\left[|W_p(\hat{\mu}_n, \hat{\nu}_n) - W_p(\mu, \nu)|\right]$ où $\nu \in \mathscr{P}_p(\mathbb{R}^d)$ et $\hat{\nu}_n$ est sa version empirique. Finalement, Genevay et al. [2019] ont montré au début de mes recherches doctorales que le transport optimal régularisé par l'entropie permettait effectivement de réduire le fléau de la dimension : pour une fonction de coût lipschitzienne c et à force de régularisation $\gamma > 0$ fixée, $\mathbb{E}\left[|\mathscr{S}_c^\gamma(\hat{\mu}_n, \hat{\nu}_n) - \mathscr{S}_c^\gamma(\mu, \nu)|\right] = \mathcal{O}\left(1/\sqrt{n}\right)$. Cependant, ce résultat se détériore lorsque $\gamma \to 0$, et est valable quelles que soient les mesures μ et ν,

tant qu'elles sont à supports compacts. Ainsi, les structures potentiellement faiblement dimensionnelles des mesures ne sont pas prises en compte par le transport optimal régularisé par l'entropie. La question de trouver des approches algorithmiques qui puissent réduire les effets du fléau de la dimension, tout en tirant partie des structures de faible dimension cachées dans les mesures, était donc partiellement ouverte au début de mon doctorat.

1.3.2 Obtention de robustesse grâce aux projections

Une approche, cependant, exploitait déjà les structures de faible dimension dans les mesures : la distance de Wasserstein par tranches (WT) introduite par Rabin et al. [2011]. La raison principale qui a mené à la définition de la distance WT est algorithmique, puisque calculer la distance de Wasserstein entre les projections unidimensionnelles des deux mesures possède une complexité log-linéaire, bien meilleure que la complexité cubique du problème de transport optimal originel. Bien que cette approche améliore les aspects computationnels, il m'a semblé que projeter des mesures aux supports de grande dimension sur des lignes unidimensionnelles causerait des pertes importantes d'information sur les mesures initiales, ou du moins que ce choix algorithmique des lignes était un peu *ad hoc*. Pire encore, lorsque l'on moyenne les valeurs $W_p\left(P_{\theta\sharp}\mu, P_{\theta\sharp}\nu\right)$ sur la sphère $\theta \in \mathbb{S}^{d-1}$, la plupart des projections ne discriminent que peu entre les mesures, polluant ainsi la moyenne globale. Il m'a donc paru clair que projeter les mesures, non plus sur des lignes mais sur des sous-espaces de faible dimension que nous choisirions de manière à optimiser un certain critère, créerait des opportunités de recherches novatrices. Dans l'analyse en composantes principales (ACP) [Pearson, 1901], les données sont projetées sur le sous-espace de faible dimension qui maximise la variance des points projetés. La traduction de ce fait dans le cadre des mesures de probabilité et du transport optimal s'est avérée être la clef pour définir des variantes robustes des distances de Wasserstein qui exploitent les structures de faible dimension dans les mesures.

Ce problème se réécrit naturellement comme la maximisation de la distance de Wasserstein par rapport à la fonction de coût. Le choix d'une fonction de coût lorsque l'on calcule le transport optimal est un problème important : comme remarqué par Genevay et al. [2018, Section 3], apprendre une fonction de coût de manière adverse s'avère crucial lorsque l'on calcule le transport optimal dans des espaces de grande dimension. Mais comment interpréter un tel choix ? Dans notre cas, le lien avec les sous-espaces de faible dimension rend l'interprétation aisée : le problème peut se réinterpréter naturellement comme un problème de transport optimal dans lequel l'objectif linéaire est remplacé par un objectif spectral convexe, à la manière de l'ACP où le sous-espace optimal est calculé à partir de la décomposition en sous-espaces propres de la matrice de covariance des données. Nous avons ainsi mis au jour un lien, quoique apparemment frêle, entre la maximisation des distances de Wasserstein par rapport à la fonction de

coût d'une part, et la convexification de l'objectif classique du transport optimal d'autre part. J'ai ainsi décidé de poursuivre dans cette voie, et d'éclaircir ce lien, ce qui s'est avéré productif : toute convexification du coût traditionnel de transport optimal peut se réécrire grâce à la dualité de Legendre comme une maximisation du problème par rapport à la fonction de coût. Réciproquement, cela signifie donc que toute régularisation convexe du transport optimal, et en particulier le transport optimal régularisé par l'entropie, est robuste à la fonction de coût. Ce travail a apporté une nouvelle lumière sur les raisons qui font que la régularisation fonctionne en pratique : la régularisation implique un choix automatique d'une fonction de coût robuste, et inversement, maximiser par rapport à la fonction de coût comme dans [Genevay et al., 2018] correspond à une régularisation implicite du problème de transport optimal.

1.3.3 Sur la difficulté de l'estimation des cartes de Monge

Alors que le fléau de la dimension et le choix robuste d'une fonction de coût concernent la *valeur* du transport optimal, certaines applications utilisent la *carte* de transport optimal. Par exemple, Courty et al. [2016] alignent deux jeux de données en transportant l'un sur l'autre de manière optimale ; dans [Schiebinger et al., 2019], les auteurs utilisent les cartes de transport optimal pour suivre l'évolution d'une colonie de cellules lorsque seulement des instantanés sont disponibles. Dans ces applications, avoir accès à une carte, et non un plan, de transport est absolument capital : un point doit être envoyé vers une unique destination. Il est encore plus important de pouvoir définir ces cartes sur l'espace tout entier, afin de pouvoir généraliser le transport hors des seules données. Malheureusement, le problème de Monge n'admet pas systématiquement de minimiseur, et à vrai dire, aucune carte ne peut transporter une mesure discrète sur une autre dont le support est de cardinal strictement plus grand. Une première approche proposée par Courty et al. [2016] consiste à calculer d'abord un plan de transport optimal avant d'en déduire une carte en considérant sa projection barycentrique, dont on sait qu'elle fournit une carte optimale dans le cas quadratique [Ambrosio et al., 2006, Théorème 12.4.4]. Dans ce cas, la carte obtenue est définie sur les données passées et ne peut pas se généraliser à de nouvelles données. Afin de définir une carte sur tout l'espace, Perrot et al. [2016] ont proposé de l'apprendre au sein d'une famille paramétrique (fonctions linéaires, ou non-linéaires au moyen de noyaux). Mais comment assurer alors l'optimalité du transport induit par une telle carte ? Les auteurs proposent d'apprendre conjointement un plan et une carte, de telle manière à ce que le plan minimise approximativement le coût de transport et que la carte soit approximativement égale à la projection barycentrique de ce plan. Cette carte n'induit donc pas un transport optimal. Deux ans plus tard, Seguy et al. [2018] ont continué cette approche en proposant de modéliser la carte de transport par un réseau de neurones profond. Encore une fois, l'optimalité du transport induit par une telle carte ne peut être garantie.

1.3.4 Cartes de Monge tirant partie de la convexité et de la régularité

Comment donc réconcilier ces deux spécifications : optimalité du transport et généralisation à tout l'espace ? Puisque aucune réponse n'avait été proposée lorsque j'ai commencé ma thèse, j'ai décidé de poursuivre cette piste de recherche. Au lieu de définir une carte à partir d'un plan de transport en utilisant la projection barycentrique, j'ai commencé à chercher une carte qui soit optimale *par définition*. Dans le cas quadratique, le théorème de Brenier [1987] donne une condition nécessaire et suffisante d'optimalité : une carte $T : \mathbb{R}^d \to \mathbb{R}^d$ est optimale si, et seulement si, elle est le gradient d'un potentiel convexe. En reformulant les procédures d'estimation en termes de potentiels convexes de Brenier, on garantit ainsi l'optimalité du transport induit par la carte apprise. En allant un peu plus loin, la théorie de la régularité du transport optimal nous assure que nous pouvons espérer au moins un peu de régularité sur la carte de Monge. Ajouter ainsi de tels *a priori* de régularité lors de l'estimation devrait aider à la généralisation : en contraignant la carte à être bi-lipschitzienne (c'est-à-dire que le potentiel de Brenier associé doit être fortement convexe et lisse), deux points proches doivent être transportés vers des destinations qui restent proches. Reste la question de savoir comment de tels potentiels fortement convexes et lisses, et leurs gradients, peuvent être généralisés à l'espace tout entier. Pour ce faire, la littérature portant sur les problèmes d'interpolation convexe (voir [Taylor et al., 2017, Section 3] pour un aperçu sur le sujet) nous vient en aide : elle nous fournit des méthodes algorithmiques qui sont capables de calculer une fonction fortement convexe et lisse $u : \mathbb{R}^d \to \mathbb{R}$ dont les valeurs $u(x_i)$ et/ou les valeurs de son gradient $\nabla u(x_i)$ sont prescrites sur un nombre fini de points $x_1, \ldots, x_n \in \mathbb{R}^d$. En combinant l'approche à la Brenier et ces méthodes d'interpolation convexe, nous avons pu proposer une procédure d'estimation qui retourne une carte de Monge bi-lipschitzienne, optimale, et définie sur l'espace tout entier.

1.3.5 Contributions

L'objectif principal de cette thèse est de contribuer à la compréhension de la régularisation du transport optimal, ainsi que de proposer de nouvelles variantes de régularisation des distances de transport optimal. Cette thèse est divisée en deux parties thématiques :

- **Partie I : Robustesse face à la fonction de coût**
  Le transport optimal est grandement dépendant du choix de la fonction de coût [Genevay et al., 2018, Section 3]. En pratique, on choisit souvent la distance Euclidienne, ou la distance Euclidienne au carré, menant ainsi aux distances de Wasserstein d'ordre 1 et 2, respectivement. Mais les distances de Wasserstein subissent le fléau de la dimension [Dudley, 1969, Weed and Bach, 2019] et doivent donc être régularisées. Dans cette

partie, nous proposons de maximiser le transport optimal par rapport à la fonction de coût, contrainte à appartenir à une classe de fonction donnée. Dans le chapitre 2, la classe de fonctions de coût admissibles est celle des distances de Mahalanobis, ce qui donne lieu à une variante robuste de la distance de Wasserstein d'ordre 2 et dont on montre qu'elle a des liens avec la réduction de dimension. Dans le chapitre 3, nous prouvons que maximiser le transport optimal par rapport à la fonction de coût correspond en fait à régulariser le transport optimal classique, généralisant ainsi un résultat obtenu au chapitre 2.

- **Partie II : Cartes de transport contraintes par la régularité**
  Dans diverses applications, avoir accès à la carte optimale de transport envoyant une mesure sur une autre est un enjeu crucial [Courty et al., 2016, Schiebinger et al., 2019]. En pratique, cependant, lorsqu'un nombre fini seulement de données est disponible, de telles cartes n'existent pas nécessairement ou ne sont définies que sur les points des données. De manière à étendre ces cartes à l'espace tout entier, nous proposons d'apprendre une carte qui est par nature optimale et qui est contrainte à être bi-lipschitzienne, permettant ainsi la généralisation en dehors des données.

Nous donnons à présent davantage de détails à propos des chapitres de cette thèse : pour chacun d'entre eux, nous esquissons les travaux qui y sont liés et nous présentons nos contributions originales.

Chapitre 2 : Distances de Wasserstein Robustes aux Sous-Espaces

Ce chapitre s'appuie sur la publication [Paty and Cuturi, 2019].

Le transport optimal souffre du fléau de la dimension [Dudley, 1969, Weed and Bach, 2019]. Les approches classiques pour résoudre ce problème utilisent des projections sur des lignes [Rabin et al., 2011], ou appliquent une réduction de dimension préalable aux mesures [Schiebinger et al., 2019]. Dans ce chapitre, nous proposons d'effectuer la réduction de dimension et le calcul du transport optimal en même temps. Cela mène à une variante robuste de la distance de Wasserstein d'ordre 2, que nous baptisons *distance de Wasserstein robuste aux sous-espaces*.

Travaux liés Les distances de Wasserstein souffrent du fléau de la dimension : Dudley [1969], Fournier and Guillin [2015] montrent que leur complexité statistique croît exponentiellement vite avec la dimension, si bien qu'un nombre exagérément grand de données est nécessaire dès lors qu'elles vivent en grande dimension. Bien que ces résultats soient en partie atténués lorsque les données

vivent sur une variété de faible dimension [Weed and Bach, 2019], cette hypothèse n'est pas toujours vérifiée ou suffisante en pratique. Une approche initiée par Cuturi [2013] propose d'ajouter une pénalisation entropique au problème de transport originel, ce qui résulte en des bornes statistiques améliorées [Genevay et al., 2019], mais ne prend pas en compte les structures de faible dimension présentes dans les données. Une approche concurrente exploite le fait que le calcul des distances de Wasserstein entre des distributions unidimensionelles ne demande que la comparaison de leurs fonctions quantiles [Santambrogio, 2015, Theorem 2.9]. Calculer les fonctions quantiles ne requiert que le tri des valeurs, ce qui possède une complexité log-linéaire. La *distance de Wassertein par tranches* [Rabin et al., 2011] consiste ainsi à calculer l'espérance de la distance de Wasserstein entre les projections des deux distributions sur une même ligne tirée aléatoirement uniformément sur la sphère. Une approche encore plus simple et utilisée dans certaines applications sur des données réelles, par exemple dans [Schiebinger et al., 2019], consiste à réduire la dimension des données en utilisant une *analyse en composantes principales* (ACP) [Pearson, 1901] avant d'appliquer les techniques de transport optimal.

Contributions La principale contribution de ce chapitre consiste en l'introduction et l'étude initiale de nouvelles distances de transport optimal que nous baptisons *distances de Wasserstein robustes aux projections* et *distance de Wasserstein robuste aux sous-espaces*.

1. **Maximisation de la distance de Wasserstein quadratique entre les projections des mesures :** Étant données deux distributions de probabilité $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, nous proposons de les projeter sur un sous-espace E de $\mathbb{R}^d$ de dimension k, puis de calculer la distance de Wasserstein d'ordre 2 entre ces projections. Nous choisissons ensuite le sous-espace E de manière adverse :

$$\mathcal{P}_k(\mu, \nu) \stackrel{\text{def}}{=} \sup_{E \in \mathcal{G}_k} W_2\left(P_{E\sharp}\mu, P_{E\sharp}\nu\right)$$

   où $\mathcal{G}_k = \left\{E \text{ est un sous-espace de } \mathbb{R}^d, \dim(E) = k\right\}$ est la Grassmannienne et P_E est le projecteur orthogonal sur $E \in \mathcal{G}_k$.

   a) Nous montrons que $\mathcal{P}_k$ munit $\mathscr{P}_2(\mathbb{R}^d)$ d'une structure métrique.

   b) Nous montrons qu'une définition analogue avec la distance de Wasserstein W_p, $p \geq 1$, au lieu de W_2, définit aussi une distance sur $\mathscr{P}_p(\mathbb{R}^d)$.

   c) Le problème de maximisation dans la définition de $\mathcal{P}_k$ n'est pas convexe à cause de la non convexité de la Grassmannienne $\mathcal{G}_k$, ce qui rend coûteux le calcul ou l'approximation de $\mathcal{P}_k$.

2. **Relaxation convexe de $\mathcal{P}_k$ et son étude théorique :** Nous proposons donc de relâcher la définition de $\mathcal{P}_k$. Pour ce faire, nous échangeons le supremum (sur la Grassmannienne) et l'infimum (sur les plans de transport) dans la définition de $\mathcal{P}_k$ pour définir la *distance de Wasserstein robuste aux sous-espaces* :

$$\mathcal{S}_k(\mu,\nu) \stackrel{\text{def}}{=} \inf_{\pi\in\Pi(\mu,\nu)} \sup_{E\in\mathcal{G}_k} \left(\int \|P_E(x)-P_E(y)\|^2 d\pi(x,y)\right)^{1/2}.$$

   a) Nous montrons que $\mathcal{S}_k$ définit une distance géodésique sur $\mathscr{P}_2(\mathbb{R}^d)$ et qui est fortement équivalente à W_2. Nous donnons les constantes exactes dans les inégalités d'équivalence métrique, et nous exhibons explicitement des géodésiques.

   b) Nous montrons que $\mathcal{S}_k$ est effectivement une relaxation convexe de $\mathcal{P}_k$:

$$\mathcal{S}_k(\mu,\nu) = \max_{\substack{0\preceq\Omega\preceq I\\ \operatorname{trace}(\Omega)=k}} W_2\left(\Omega^{1/2}{}_\sharp\mu, \Omega^{1/2}{}_\sharp\nu\right).$$

   L'ensemble $\mathcal{R} = \left\{\Omega\in\mathbb{R}^{d\times d}, 0\preceq\Omega\preceq I\,,\, \operatorname{trace}(\Omega)=k\right\}$ est en effet l'enveloppe convexe de l'ensemble des matrices de projection orthogonale de rang k.

   c) Nous montrons que $\mathcal{S}_k$ admet une formulation alternative en termes de la matrice des seconds moments des déplacements, clarifiant ainsi les liens avec l'ACP :

$$\mathcal{S}_k^2(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \sum_{l=1}^{k} \lambda_l(V_\pi)$$

   où $V_\pi \stackrel{\text{def}}{=} \iint (x-y)(x-y)^\top d\pi(x,y)$ et λ_l et la $l^{\text{ème}}$ plus grande valeur propre.

   d) Nous prouvons que $\mathcal{S}_k$ peut se réinterpréter comme la maximisation du coût de transport optimal $\mathscr{T}_c$ par rapport à la fonction de coût :

$$\mathcal{S}_k^2(\mu,\nu) = \max_{\substack{0\preceq\Omega\preceq I\\ \operatorname{trace}(\Omega)=k}} \mathscr{T}_{d_\Omega^2}(\mu,\nu)$$

   où d_Ω^2 désigne le carré de la distance de Mahalanobis :

$$d_\Omega^2(x,y) = (x-y)^\top\Omega(x-y).$$

3. **Outils numériques pour calculer $\mathcal{S}_k$:** Quand μ,ν sont des mesures discrètes, $\mathcal{S}_k$ définit un problème d'optimisation convexe de dimension finie. Nous proposons deux algorithmes pour le résoudre :

a) *Montée de surgradients projetés :* nous considérons la maximisation sur $\Omega \in \mathcal{R}$ de la fonction concave $f : \Omega \mapsto \mathscr{T}_{d^2_\Omega}(\mu, \nu)$ qui admet pour surdifférentielle l'ensemble :

$$\partial f(\Omega) = \operatorname{conv}\left\{ V_{\pi_\star}, \pi_\star \in \underset{\pi \in \Pi(\mu,\nu)}{\arg\min} \langle \Omega, V_\pi \rangle \right\}.$$

L'algorithme suit alors les itérations suivantes jusqu'à convergence :

i. Calculer $\pi_\star$ en résolvant le problème de transport (à Ω fixé) ;
ii. Faire un pas de surgradient : $\Omega \leftarrow \Omega + \varepsilon V_{\pi_\star}$ où ε est la taille du pas ;
iii. Projeter Ω sur l'ensemble des contraintes $\mathcal{R}$.

b) *Algorithme de Frank-Wolfe avec régularisation entropique :* afin d'accélérer les calculs de transport optimal, nous ajoutons une régularisation entropique à l'objectif :

$$f_\gamma : \Omega \mapsto \mathscr{S}^\gamma_{d^2_\Omega}(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int d^2_\Omega \, d\pi + \gamma \, \mathrm{KL}(\pi || \mu \otimes \nu),$$

où $\gamma > 0$ est la force de la régularisation. Cela rend aussi l'objectif f_γ lisse, permettant ainsi d'utiliser les itérations de Frank-Wolfe plutôt qu'une montée de gradients. Les itérations de Frank-Wolfe ne nécessitent pas de projections sur $\mathcal{R}$, ni d'ajuster une taille de pas. L'algorithme est le suivant :

i. À Ω fixé, on résout le problème de transport entropique

$$\min_{\pi \in \Pi(\mu,\nu)} \int d^2_\Omega \, d\pi + \gamma \, \mathrm{KL}(\pi || \mu \otimes \nu)$$

en utilisant l'algorithme de Sinkhorn ;

ii. À π fixé, on résout le problème

$$\max_{\substack{0 \preceq \Omega \preceq I \\ \operatorname{trace}(\Omega) = k}} \int d^2_\Omega \, d\pi$$

dont le maximiseur est donné par $\widehat{\Omega} = U \operatorname{diag}([\mathbf{1}_k, \mathbf{0}_{d-k}]) U^\top$, qui utilise la décomposition en valeurs propres de

$$V_\pi = U \operatorname{diag}(\lambda_1, ..., \lambda_d) U^\top ;$$

iii. On met à jour selon Frank-Wolfe : $\Omega \leftarrow (1 - \tau)\Omega + \tau \widehat{\Omega}$ où $\tau = 2/(2 + n)$ lors de la $n^{\text{ème}}$ itération.

Chapitre 3 : Le Transport Optimal Régularisé est Robuste à la Fonction de Coût

Ce chapitre s'appuie sur la publication [Paty and Cuturi, 2020].

Régulariser le problème de transport optimal s'est avéré crucial pour que la théorie du transport puisse avoir des répercussions sur le domaine de l'apprentissage automatique. Par exemple, il est bien connu que régulariser le transport optimal avec l'entropie permet d'accélérer les calculs numériques et de pouvoir différencier en utilisant l'algorithme de Sinkhorn [Cuturi, 2013], ainsi que de profiter de meilleures bornes statistiques [Genevay et al., 2019]. Dans ce chapitre, nous partons de ce point de vue pratique et nous proposons une nouvelle interprétation de la régularisation comme mécanisme de robustesse. Nous montrons au moyen de la dualité de Legendre que toute régularisation convexe du transport optimal peut se réinterpréter comme étant robuste à la fonction de coût.

Travaux liés La clef pour utiliser le transport optimal dans les applications tient aux différentes formes de régularisation du problème originel. Ajouter une petite régularisation convexe au coût linéaire classique aide non seulement du point de vue algorithmique, en convexifiant l'objectif à minimiser et en ouvrant la voie à des algorithmes plus rapides, mais introduit aussi un compromis sur la régularité qui empêche le surapprentissage sur les données. Bien que la régularisation entropique soit la plus étudiée en transport optimal, grâce à ses avantages algorithmiques [Cuturi, 2013], plusieurs autres régularisations convexes du plan de transport ont été proposées dans la littérature : régularisation quadratique [Essid and Solomon, 2017], transport avec contraintes de capacités [Korman and McCann, 2015], régularisation groupe-lasso [Courty et al., 2016], régularisation Laplacienne [Flamary et al., 2014], régularisation de Tsallis [Muzellec et al., 2017], parmi bien d'autres. D'un autre côté, Liero et al. [2018] ont montré que régulariser le problème dual de Kantorovitch correspondait au transport déséquilibré, c'est-à-dire au transport avec des contraintes marginales relaxées. Plusieurs approches se sont intéressées aux raisons qui font que la régularisation du transport optimal s'avère être cruciale en pratique. Une première raison est statistique : quoique le transport optimal classique subisse le fléau de la dimension, puisque sa version empirique converge à la vitesse de $\mathcal{O}\left(n^{-1/d}\right)$ [Dudley, 1969, Fournier and Guillin, 2015, Weed and Bach, 2019], le transport régularisé par l'entropie, et plus précisément les divergences de Sinkhorn, a une complexité statistique en $O(1/\sqrt{n})$ [Genevay et al., 2019, Mena and Niles-Weed, 2019]. Il a aussi été montré dans [Rigollet and Weed, 2018] que le transport entropique revient à calculer l'estimateur du maximum de vraisemblance dans le modèle de déconvolution gaussienne. Dans une autre approche, Dessein et al. [2018], Blondel et al. [2018] ont considéré des classes

générales de régularisation convexe et les ont caractérisées d'un point de vue géométrique. Récemment, plusieurs articles [Genevay et al., 2018, Flamary et al., 2018, Deshpande et al., 2019, Kolouri et al., 2019, Niles-Weed and Rigollet, 2019, Paty and Cuturi, 2019] ont proposé de maximiser le transport optimal par rapport à la fonction de coût, ce qui peut s'interpréter à la lumière de l'apprentissage de métrique sous-jacente [Cuturi and Avis, 2014]. Cette approche peut aussi s'interpréter sous l'angle de l'optimisation robuste [Ben-Tal and Nemirovski, 1998, Ben-Tal et al., 2009, Bertsimas et al., 2011] : plutôt que de considérer un problème de minimisation dépendant des données qui serait instable $\min_x f_{\hat{\theta}}(x)$, où $\hat{\theta}$ représente les données, la littérature de l'optimisation robuste choisit le paramètre θ de manière robuste dans un voisinage des données : $\max_{\theta\in\Theta} \min_x f_\theta(x)$.

Contributions Dans la continuité de ces travaux, notre contribution principale dans ce chapitre consiste à éclairer le lien entre la *régularisation* et la *maximisation* du transport optimal. Notre but principal est de fournir une nouvelle interprétation du transport optimal régularisé en termes de robustesse à la fonction de coût : régulariser le transport revient à maximiser le transport **non** régularisé par rapport à la fonction de coût.

1. **Nous interprétons les régularisations convexes du transport optimal en termes de robustesse à la fonction de coût :** Ici, $\mathcal{X}$ est un espace séparé et compact. Soit $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R}\cup\{+\infty\}$ une fonctionnelle semi-continue inférieurement (s.c.i) et convexe. Nous montrons que :

$$\inf_{\pi\in\Pi(\mu,\nu)} F(\pi) = \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - F^*(c),$$

où $F^* : c \mapsto \sup_{\pi\in\mathscr{M}(\mathcal{X}^2)} \int c\,d\pi - F(\pi)$ est la transformée de Legendre de F. Cela signifie que minimiser une fonctionnelle convexe du plan de transport correspond à la maximisation du transport optimal classique par rapport à la fonction de coût, avec une pénalisation sur la fonction de coût. Une reformulation importante de ce fait est la suivante : si $R : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R}\cup\{+\infty\}$ est s.c.i, $c_0 : \mathcal{X}^2 \to \mathbb{R}_+$ est continue et $\varepsilon > 0$, alors :

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c_0\,d\pi + \varepsilon R(\pi) = \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \varepsilon R^*\left(\frac{c-c_0}{\varepsilon}\right),$$

où R^* est la transformée de Legendre de R. Cette formulation montre que régulariser le transport optimal revient donc à maximiser le transport classique par rapport à la fonction de coût, avec une régularisation qui force le coût adverse à être proche du coût originel c_0, dans le sens où $R^*\left(\frac{c-c_0}{\varepsilon}\right)$ ne peut pas prendre une valeur trop élevée. Le lien entre la régularisation sur le plan de transport et la pénalisation sur la fonction de coût adverse est donnée par la transformée de Legendre des régularisations.

2. **Nous investigons les propriétés des fonctions de coût adverses optimales :** Nous prouvons, sous une hypothèse technique (qui est vérifiée par exemple pour les régularisations entropique et quadratique), un théorème de dualité pour le transport régularisé :

$$\inf_{\pi \in \Pi(\mu,\nu)} F(\pi) = \sup_{\phi,\psi \in \mathcal{C}(\mathcal{X})} \int \phi \, d\mu + \int \psi \, d\nu - F^*(\phi \oplus \psi)$$

où $\phi \oplus \psi : (x, y) \mapsto \phi(x) + \psi(y)$. Nous utilisons ce résultat de dualité pour montrer, sous la même hypothèse technique, qu'il existe une fonction de coût adverse optimale qui soit séparable, c'est-à-dire qu'il existe des fonctions $\phi, \psi : \mathcal{X} \to \mathbb{R}$ telles que $\phi \oplus \psi$ soit optimale dans le problème de maximisation $\sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu, \nu) - F^*(c)$. En un sens, ce résultat est décevant puisqu'une fonction de coût qui est séparable ne munit pas l'espace sous-jacent $\mathcal{X}$ d'une géométrie intéressante et à même d'analyser les données.

Chapitre 4 : Potentiels de Brenier Approchants Lisses et Fortement Convexes

Ce chapitre s'appuie sur la publication [Paty et al., 2020].

Une des plus grandes réussites de la littérature de transport optimal ces dernières années concerne la théorie de la régularité : Caffarelli [2000] a montré que la carte de transport entre deux mesures vérifiant certaines propriétés est lipschitzienne, ou de manière équivalente lorsque l'on considère la distance de Wasserstein d'ordre 2, que le potentiel convexe de Brenier (dont le gradient donne une carte optimale) est lisse. Plutôt que de considérer la régularité comme une propriété qui peut être prouvée sous certaines hypothèses, nous considérons dans ce chapitre la régularité comme une condition qui doit être imposée lorsque l'on estime le transport optimal : nous proposons des algorithmes qui peuvent produire des cartes de transport qui soient optimales, qui approchent au mieux la contrainte de transport, et qui aient une distorsion faible, c'est-à-dire que les potentiels de Brenier associés sont fortement convexes et lisses.

Travaux liés Apprendre une carte de Monge est un point crucial dans certaines applications du transport optimal [Courty et al., 2016, Schiebinger et al., 2019]. Pour ce faire, deux approches ont principalement été considérées dans la littérature :

- apprendre une carte de transport à partir d'un plan de transport optimal, en considérant sa projection barycentrique [Courty et al., 2016] : quoique la carte obtenue soit effectivement une carte de Monge optimale, elle n'est définie que sur les données et ne peut se généraliser à de nouveaux points ;

- apprendre une carte de transport qui soit définie partout, en considérant une famille paramétrique [Perrot et al., 2016, Seguy et al., 2018] : dans ce cas, on ne peut pas garantir que les cartes apprises sont des cartes de Monge optimales.

Afin de réconcilier ces deux approches, nous nous tournons vers la littérature du transport optimal, qui nous fournit une riche théorie sur les cartes de Monge : la théorie de la régularité. Cette théorie donne des propriétés sur la carte de Monge qui envoie une mesure μ sur une autre mesure ν avec un faible coût moyen. Lorsque ce coût est la distance Euclidienne au carré, la carte de Monge est nécessairement le gradient ∇f d'une fonction convexe f. Ce résultat majeur, connu comme le théorème de Brenier [1987], affirme que le problème de transport optimal entre μ et ν est résolu dès lors qu'il existe une fonction convexe f telle que $\nabla f_\sharp \mu = \nu$. Estimer une carte de Monge en la forçant à être le gradient d'une fonction convexe garantira ainsi son optimalité. Dans ce contexte, la régularité se comprend comme la propriété de lipschitzianité de la carte ∇f, résultat fondamental dû à Caffarelli [2000] qui a prouvé que la carte de Monge est nécessairement 1-lipschitzienne lorsqu'elle transporte une mesure "plus grosse qu'une Gaussienne" $\mu \propto e^V \gamma_d$ vers une mesure "plus fine qu'une Gaussienne" $\nu \propto e^{-W} \gamma_d$ (ici, γ_d est la mesure Gaussienne centrée réduite sur $\mathbb{R}^d$ $\gamma_d \propto e^{-\|\cdot\|^2}$, et V, W sont deux potentiels convexes). De manière équivalente, ce résultat se réécrit en disant que la carte de Monge est le gradient d'un potentiel de Brenier f convexe et 1-lisse. Ajouter des contraintes de régularité de ce genre lors de l'estimation d'une carte de Monge permettra de la définir en tout point, et donc en dehors des données, au moyen des résultats obtenus dans la littérature d'interpolation convexe [Taylor et al., 2017, Section 3].

Contributions La contribution principale de ce chapitre consiste à traduire cette idée selon laquelle la carte de transport optimal entre des mesures doit être suffisamment régulière en une procédure d'estimation.

1. **Nous définissons les *potentiels de Brenier approchants lisses et fortement convexes* (potentiels BALFC) :** Étant données deux mesures de probabilité $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, une fonction f qui soit ℓ-fortement convexe et L-lisse et telle que $\nabla f_\sharp \mu = \nu$ n'existe pas nécessairement. Nous relaxons cette égalité, et cherchons plutôt une fonction f fortement convexe et lisse qui minimise la distance de Wasserstein d'ordre 2 entre $\nabla f_\sharp \mu$ et ν :

$$\min_{f \text{ est } L\text{-lisse, } \ell\text{-fortement convexe}} W_2 \left(\nabla f_\sharp \mu, \nu\right). \tag{BALFC}$$

   Nous appelons de tels potentiels des "potentiels de Brenier approchants lisses et fortement convexes" car ils sont les meilleurs moyens de transporter la mesure μ en une mesure qui soit aussi proche que possible de ν tout en utilisant un potentiel lisse et fortement convexe.

2. **Nous montrons que de tels potentiels peuvent être calculés lorsque les mesures sont discrètes :** Même lorsque μ, ν sont des mesures de probabilité discrètes, le problème (BALFC) est de dimension infinie à cause de la minimisation sur les fonctions lisses et fortement convexes. Nous montrons que ce problème peut en fait se réécrire comme un problème d'optimisation de dimension finie qui soit séparément convexe : à potentiel f fixé, le problème est un problème de transport discret ; à plan de transport fixé, la minimisation sur f est un problème convexe quadratique à contraintes quadratiques, pour lequel il existe des solveurs numériques.

3. **Nous remarquons que le cas unidimensionnel est plus simple à résoudre :** Dans le cas univarié, nous montrons que calculer les potentiels (BALFC) revient à résoudre une variante du problème de régression isotonique dans lequel la carte à estimer est contrainte à être strictement croissante et lipschitzienne. Une approche par descente de gradient projeté peut alors être utilisée pour résoudre efficacement ce problème.

4. **Nous montrons comment calculer les potentiels** (BALFC) **et la carte associée dans tout l'espace :** Nous utilisons les solutions de ces problèmes d'optimisation pour étendre le potentiel (BALFC) et la carte associée à n'importe quel point de l'espace. Nous montrons que cela peut se faire en résolvant un problème convexe quadratique à contraintes quadratiques pour chaque point de l'espace.

1.4 Notation

General notation and sets

$\iota(A)$	:	The indicator function $\iota(A) = 0$ if the assertion A is true and $\iota(A) = +\infty$ otherwise
$\operatorname{conv} S$	:	The convex hull of the set S
$\lvert S\rvert$	:	The cardinal of the set S
$[\![n]\!]$	:	The set $[\![n]\!] = \{1, ..., n\}$
$\mathfrak{S}_n$	:	The set of permutations of $[\![n]\!]$
$\mathcal{G}_k$	:	The Grassmannian, *i.e.* the set of all k-dimensional subspaces in $\mathbb{R}^d$
$\mathbb{S}^{d-1}$	:	The unit sphere $\mathbb{S}^{d-1} = \left\{x \in \mathbb{R}^d, \|x\| = 1\right\}$ in $\mathbb{R}^d$

Measures

$\mathscr{M}(\mathcal{X})$	:	The space of Borel finite signed measures over $\mathcal{X}$
$\mathscr{P}(\mathcal{X})$	:	The set of probability measures on $\mathcal{X}$
$\mathscr{P}_p(\mathbb{R}^d)$	:	The set of probability measures on $\mathbb{R}^d$ with finite p first moments
δ_x	:	The Dirac mass concentrated at point x
$\mathscr{L}^d$	:	The Lebesgue measure on $\mathbb{R}^d$
$T_\sharp \mu$	:	The pushforward measure of μ by T
$\Pi(\mu, \nu)$	:	The set of couplings with marginals μ and ν
$\operatorname{supp}(\mu)$	:	The support of μ
$\mu \otimes \nu$	:	The product measure of μ and ν
$\mu \ll \lambda$	:	The measure μ is absolutely continuous with respect to λ
$\mathrm{KL}(\mu\|\|\lambda)$	:	The Kullback-Leibler divergence of μ with respect to λ

Functions

$\mathcal{C}(\mathcal{X})$	:	The space of real-valued continuous functions on $\mathcal{X}$
$f \oplus g$	:	The function $f \oplus g : (x, y) \mapsto f(x) + g(y)$
f^*	:	The Legendre-Fenchel transform of f
$\partial f(x)$	:	The subdifferential (*resp.* superdifferential) of the convex (*resp.* concave) function f at point x

Vectors and matrices

$A^\top$	:	The transpose matrix of A
$\operatorname{trace}(A)$	:	The trace of A
$\langle A, B\rangle$	:	The Frobenius inner-product between the matrices A and B
$\lambda_k(A)$	:	The k^{th} *largest* eigenvalue of A
$A \preceq B$	:	The Loewner order, *i.e.* $\Leftrightarrow B - A$ is positive semi-definite
$\operatorname{diag}(a)$	:	The diagonal matrix with diagonal a
P_E	:	The orthogonal projection matrix onto the subspace E
$\mathbf{1}_k$	:	The vector of all ones $(1, \dots, 1)^\top \in \mathbb{R}^k$
$\mathbf{0}_k$	:	The vector of all zeros $(0, \dots, 0)^\top \in \mathbb{R}^k$
$\langle a, b\rangle$	:	The inner-product between the vectors a and b
$a \otimes b$	:	The elementwise multiplication of a per b
$a \oslash b$	:	The elementwise division of a per b

Part I

Ground-Cost Robustness

Chapter 2

Subspace Robust Wasserstein Distances

2.1 Introduction

When using optimal transport (OT) on high-dimensional data, practitioners are often confronted to the intrinsic instability of OT with respect to input measures. A well known result states for instance that the sample complexity of Wasserstein distances can grow exponentially in dimension [Dudley, 1969, Fournier and Guillin, 2015], which means that an irrealistic amount of samples from two continuous measures is needed to approximate faithfully the true distance between them. This result can be mitigated when data lives on lower dimensional manifolds as shown in [Weed and Bach, 2019], but sample complexity bounds remain pessimistic even in that case. From a computational point of view, that problem can be interpreted as that of a lack of robustness and instability of OT metrics with respect to their inputs. This fact was already a common concern of the community when these tools were first adopted, as can be seen in the use of ℓ_1 costs [Ling and Okada, 2007] or in the common practice of thresholding cost matrices [Pele and Werman, 2009].

Regularization The idea to trade off a little optimality in exchange for more regularity is by now considered a crucial ingredient to make OT work in data sciences. A line of work initiated in [Cuturi, 2013] advocates adding an entropic penalty to the original OT problem, which results in faster and differentiable quantities, as well as improved sample complexity bounds [Genevay et al., 2019]. Following this, other regularizations [Dessein et al., 2018], notably quadratic [Blondel et al., 2018], have also been investigated. Sticking to an entropic regularization, one can also interpret the recent proposal by Altschuler et al. [2018b] to approximate Gaussian kernel matrices appearing in the regularized OT problem with Nyström-type factorizations (or exact features using a Taylor expansion [Cotter et al., 2011] as in [Altschuler et al., 2018a]), as robust

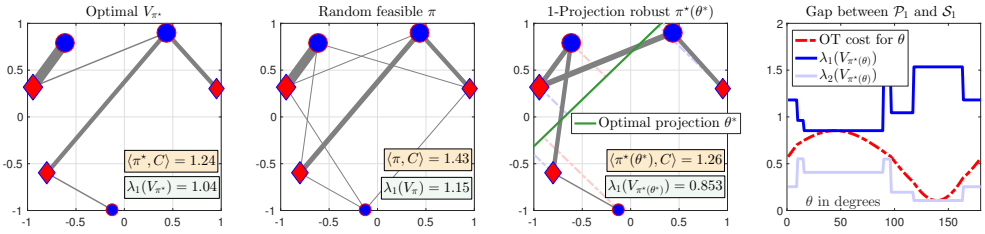


Figure 2.1: We consider two discrete measures (red and blue dots) on the plane. The left-most plot shows the optimal transport between these points, in which the width of the segment is proportional to the mass transported between two locations. The total cost is displayed in the lower right part of the plot as $\langle \pi^\star, C \rangle$, where C is the pairwise squared-Euclidean distance matrix. The largest eigenvalue of the corresponding second order moment matrix $V_{\pi^\star}$ of displacements, see (2.1), is given below. As can be expected and seen in the second plot, choosing a random transportation plan yields a higher cost. The third plot displays the most robust projection direction (green line), that upon which the OT cost of these point clouds is largest once projected. The maximal eigenvalue of the second order moment matrix (still in dimension 2) is smaller than that obtained with the initial OT plan. Finally, we plot as a function of the angle θ between $(0, 180)$ the OT cost (which, in agreement with the third plot, is largest for the angle corresponding to the green line of the third plot) as well as the corresponding maximal eigenvalue of the second order moment of the optimal plan corresponding to *each* of these angles θ. The maximum of the red curve, as well as the minimum reached by the dark blue one, correspond respectively to the values of the projection $\mathcal{P}_k$ and subspace $\mathcal{S}_k$ robust Wasserstein distances described in §3. They happen to coincide in this example, but one may find examples in which they do not, as can be seen in Figure 2.15. The smallest eigenvalue is given for illustrative purposes only.

approaches that are willing to trade-off yet a little more cost optimality in exchange for faster Sinkhorn iterations. In a different line of work, quantizing first the measures to be compared before carrying out OT on the resulting distributions of centroids is a fruitful alternative [Canas and Rosasco, 2012] which has been recently revisited in [Forrow et al., 2019]. Another approach exploits the fact that the OT problem between two distributions on the real line boils down to the direct comparison of their generalized quantile functions [Santambrogio, 2015, §2]. Computing quantile functions only requires sorting values, with a mere log-linear complexity. The *sliced* approximation of OT Rabin et al. [2011] consists in projecting two probability distributions on a given line, compute the optimal transport cost between these projected values, and then repeat this procedure to average these distances over several random lines. This approach can be used to define kernels [Kolouri et al., 2016], compute barycenters [Bonneel et al., 2015] but also to train generative

models [Kolouri et al., 2018, Deshpande et al., 2018]. Beyond its practical applicability, this approach is based on a perhaps surprising point-of-view: OT on the real line may be sufficient to extract geometric information from two high-dimensional distributions. Our work builds upon this idea, and more candidly asks what can be extracted from a little more than a real line, namely a subspace of dimension $k \geq 2$. Rather than project two measures on several lines, we consider in this chapter projecting them on a k-dimensional subspace that maximizes their transport cost. This results in optimizing the Wasserstein distance over the ground metric, which was already considered for supervised learning [Cuturi and Avis, 2014, Flamary et al., 2018].

Contributions This optimal projection translates into a "max-min" robust OT problem with desirable features. Although that formulation cannot be solved with convex solvers, we show that the corresponding "min-max" problem admits on the contrary a tight convex relaxation and also has an intuitive interpretation. To see that, one can first notice that the usual 2-Wasserstein distance can be described as the minimization of the *trace* of the second order moment matrix of the displacements associated with a transport plan. We show that computing a maximally discriminating optimal k-dimensional subspace in this "min-max" formulation can be carried out by minimizing the sum of the k largest eigenvalues (instead of the entire trace) of that second order moment matrix. A simple example summarizing the link between these two "min-max" and "max-min" quantities is given in Figure 2.1. That figure considers a toy example where points in dimension $d = 2$ are projected on lines $k = 1$, our idea is designed to work for larger k and d, as shown in section 2.5.

Chapter structure We define in section 2.2 our "max-min" and "min-max" formulations for, respectively, projection (PRW) and subspace (SRW) robust Wasserstein distances. We study the geodesic structure induced by the SRW distance on the space of probability measures in section 2.3, as well as its dependence on the dimension parameter k. We provide computational tools to evaluate SRW using entropic regularization in section 2.4. We conclude this chapter with experiments in section 2.5 to validate and illustrate our claims, on both simulated and real datasets.

2.2 Subspace Robust Wasserstein Distances

We consider here different robust formulations of the Wasserstein distance. Consider first for $k \in [\![d]\!]$, the Grassmannian of k-dimensional subspaces of $\mathbb{R}^d$:

$$\mathcal{G}_k = \left\{ E \subset \mathbb{R}^d, \dim(E) = k \right\}.$$

For $E \in \mathcal{G}_k$, we note P_E the orthogonal projector onto E. Given two measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, a first attempt at computing a robust version of $W_2(\mu, \nu)$

is to consider the worst possible OT cost over all possible low dimensional projections:

Definition 2.1. *For $k \in [\![d]\!]$, the k-dimensional projection robust 2-Wasserstein (PRW) distance between μ and ν is*

$$\mathcal{P}_k(\mu,\nu) = \sup_{E\in\mathcal{G}_k} W_2\left(P_{E\sharp}\mu, P_{E\sharp}\nu\right).$$

As we show in section 2.6, this quantity is well posed and itself worthy of interest, yet difficult to compute. In this chapter, we focus our attention on the corresponding "min-max" problem, to define the k-dimensional subspace robust 2-Wasserstein (SRW) distance:

Definition 2.2. *For $k \in [\![d]\!]$, the k-dimensional subspace robust 2-Wasserstein distance between μ and ν is*

$$\mathcal{S}_k(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \sup_{E\in\mathcal{G}_k} \left(\int \|P_E(x-y)\|^2 \, d\pi(x,y)\right)^{1/2}.$$

Remark 2.1. *Both quantities $\mathcal{S}_k$ and $\mathcal{P}_k$ can be interpreted as robust variants of the W_2 distance. By a simple application of weak duality we have that $\mathcal{P}_k(\mu,\nu) \leq \mathcal{S}_k(\mu,\nu)$.*

Lemma 2.1. *Optimal solutions for $\mathcal{S}_k$ exist,* i.e.

$$\mathcal{S}_k(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \max_{E\in\mathcal{G}_k} \left(\int \|P_E(x-y)\|^2 \, d\pi(x,y)\right)^{1/2}.$$

Proof. For $\pi \in \Pi(\mu,\nu)$, the application $E \mapsto \int \|P_E(x) - P_E(y)\|^2 \, d\pi(x,y)$ is continuous and $\mathcal{G}_k$ is compact, so the supremum is a maximum. Moreover, the application $\pi \mapsto \max_{E\in\mathcal{G}_k} \int \|P_E(x) - P_E(y)\|^2 \, d\pi(x,y)$ is lower semicontinuous as the maximum of lower semicontinuous functions. Since $\Pi(\mu,\nu)$ is compact (for any sequence in $\Pi(\mu,\nu)$ is tight), the infimum is a minimum. □

W_2 as Trace-minimization For any measures $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$ and any coupling $\pi \in \Pi(\mu,\nu)$, we define the $d \times d$ second-order displacement matrix:

$$V_\pi \stackrel{\text{def}}{=} \int (x-y)(x-y)^\top d\pi(x,y). \tag{2.1}$$

Notice that when a coupling π corresponds to a Monge map, namely $\pi = (\mathrm{Id}, T)_\sharp\mu$, then one can interpret even more naturally V_π as the second order moment of all displacement $(x - T(x))(x - T(x))^\top$ weighted by μ. With this convention, we remark that the total cost of a coupling π is equal to the trace of V_π, using the simple identity $\mathrm{trace}(x-y)(x-y)^\top = \|x-y\|^2$ and the linearity of the integral sum. Computing the W_2 distance can therefore be interpreted

as minimizing the trace of V_π.

We show next that the SRW variant $\mathcal{S}_k$ can be elegantly reformulated as a function of the eigendecomposition of the displacement second-order moment matrix V_π defined in (2.1):

Lemma 2.2. *For $k \in [\![d]\!]$ and $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, one has*

$$\mathcal{S}_k^2(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \max_{\substack{U\in\mathbb{R}^{k\times d}\\ UU^\top = I_k}} \int \|Ux-Uy\|^2 \, d\pi(x,y) = \min_{\pi\in\Pi(\mu,\nu)} \sum_{l=1}^{k} \lambda_l(V_\pi).$$

Proof. A classical variational result by Fan [1949] states that

$$\sum_{l=1}^{k} \lambda_l(V_\pi) = \max_{\substack{U\in\mathbb{R}^{k\times d}\\ UU^\top = I_k}} \operatorname{trace}\left(UV_\pi U^\top\right).$$

Then using the linearity of the trace:

$$\begin{aligned}
\sum_{l=1}^{k} \lambda_l(V_\pi) &= \max_{\substack{U\in\mathbb{R}^{k\times d}\\ UU^\top = I_k}} \int \operatorname{trace}\left[U(x-y)(x-y)^\top U^\top\right] \, d\pi(x,y) \\
&= \max_{\substack{U\in\mathbb{R}^{k\times d}\\ UU^\top = I_k}} \int \|U(x-y)\|^2 \, d\pi(x,y) \\
&= \max_{E\in\mathcal{G}_k} \int \|P_E(x)-P_E(y)\|^2 \, d\pi(x,y).
\end{aligned}$$

Taking the minimum over $\pi \in \Pi(\mu,\nu)$ yields the result. □

This characterization as a sum of eigenvalues will be crucial to study theoretical properties of $\mathcal{S}_k$. Subspace robust Wasserstein distances can in fact be interpreted as a convex relaxation of projection robust Wasserstein distances: they can be computed as the maximum of a concave function over a convex set, which will make computations tractable.

Theorem 2.1. *For $k \in [\![d]\!]$ and $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$,*

$$\mathcal{S}_k^2(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \max_{\substack{0\preceq\Omega\preceq I\\ \operatorname{trace}(\Omega)=k}} \int d_\Omega^2 \, d\pi \tag{2.2}$$

$$= \max_{\substack{0\preceq\Omega\preceq I\\ \operatorname{trace}(\Omega)=k}} \min_{\pi\in\Pi(\mu,\nu)} \int d_\Omega^2 \, d\pi \tag{2.3}$$

$$= \max_{\substack{0\preceq\Omega\preceq I\\ \operatorname{trace}(\Omega)=k}} W_2^2\left(\Omega^{1/2}{}_\sharp\mu, \Omega^{1/2}{}_\sharp\nu\right) \tag{2.4}$$

where d^2_Ω *stands for the squared Mahalanobis distance*

$$d^2_\Omega(x,y) = (x-y)^\top \Omega (x-y).$$

Proof. $\boxed{\mathcal{S}^2_k(\mu,\nu) = (2.2)}$ We fix $\pi \in \Pi(\mu,\nu)$ and focus on the inner maximization in (2.2):

$$\max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega)=k}} \int d^2_\Omega \, d\pi = \max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega)=k}} \langle \Omega, V_\pi \rangle .$$

A result by [Overton and Womersley, 1993] shows that this is equal to

$$\max_{\substack{U \in \mathbb{R}^{k \times d} \\ UU^\top = I_k}} \mathrm{trace}\left(U V_\pi U^\top\right)$$

which is nothing but the sum of the k largest eigenvalues of V_π by Fan [1949]'s result. By lemma 2.2, taking the minimum over $\pi \in \Pi(\mu,\nu)$ yields the result.

$\boxed{(2.2) = (2.3)}$ We will use Sion's minimax theorem to interchange the minimum and the maximum. Put $f(\Omega,\pi) = \int d^2_\Omega \, d\pi$ and

$$\mathcal{R} = \left\{ \Omega \in \mathbb{R}^{d \times d}, 0 \preceq \Omega \preceq I \, , \, \mathrm{trace}(\Omega) = k \right\}.$$

Note that $\mathcal{R}$ is convex and compact, and $\Pi(\mu,\nu)$ is convex (and actually compact, but we won't need it here). Moreover, f is bilinear and for any $\pi \in \Pi(\mu,\nu)$, $f(\cdot,\pi)$ is continuous. Let $\Omega \in \mathcal{R}$. Let us show that $f(\Omega,\cdot)$ is lower semicontinuous for the weak convergence. Let $(\phi_j)_{j\in\mathbb{N}}$ be an increasing sequence of bounded continuous functions, converging pointwise to d^2_Ω. Then $f(\Omega,\pi) = \sup_{j\in\mathbb{N}} \int \phi_j \, d\pi$. For $j \in \mathbb{N}$, ϕ_j is continuous and bounded, so $\pi \mapsto \int \phi_j \, d\pi$ is continuous for the weak convergence. Then $f(\Omega,\cdot)$ is lower semicontinuous as the supremum of continuous functions. Then by Sion's minimax theorem,

$$\min_{\pi \in \Pi(\mu,\nu)} \max_{\Omega \in \mathcal{R}} f(\Omega,\pi) = \max_{\Omega \in \mathcal{R}} \min_{\pi \in \Pi(\mu,\nu)} f(\Omega,\pi)$$

which is exactly (2.2) = (2.3).

$\boxed{(2.3) = (2.4)}$ Fix $\Omega \in \mathcal{R}$. One has:

$$\begin{aligned}
\min_{\pi \in \Pi(\mu,\nu)} \int d^2_\Omega \, d\pi &= \min_{\pi \in \Pi(\mu,\nu)} \int \|\Omega^{1/2}(x-y)\|^2 \, d\pi(x,y) \\
&= \min_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|^2 \, d\left[\Omega^{1/2} \otimes \Omega^{1/2}\right]_\sharp \pi(x,y) \\
&= \min_{\rho \in \Pi(\Omega^{1/2}{}_\sharp \mu, \, \Omega^{1/2}{}_\sharp \nu)} \int \|x-y\|^2 \, d\rho(x,y) \\
&= W_2^2\left(\Omega^{1/2}{}_\sharp \mu, \Omega^{1/2}{}_\sharp \nu\right)
\end{aligned}$$

where we have used Lemma 2.6. Taking the maximum over $\Omega \in \mathcal{R}$ gives the result. □

We can now prove that *both* PRW and SRW variants are, indeed, distances over $\mathscr{P}_2(\mathbb{R}^d)$.

Proposition 2.1. *For $k \in [\![d]\!]$, both $\mathcal{P}_k$ and $\mathcal{S}_k$ are distances over $\mathscr{P}_2(\mathbb{R}^d)$.*

Proof. Symmetry is clear for both objects, and for $\mu \in \mathscr{P}_2(\mathbb{R}^d)$, $\mathcal{S}_k(\mu,\mu) = \mathcal{P}_k(\mu,\mu) = 0$. Let $\mu,\nu \in \mathscr{P}_2(\mathbb{R}^d)$ such that $\mathcal{S}_k(\mu,\nu) = 0$. Then $\mathcal{P}_k(\mu,\nu) = 0$ and for any $E \in \mathcal{G}_k$, $W_2(P_{E\sharp}\mu, P_{E\sharp}\nu) = 0$, *i.e.* $P_{E\sharp}\mu = P_{E\sharp}\nu$. Lemma 2.7 then shows that $\mu = \nu$. For the triangle inequalities, let $\mu_0,\mu_1,\mu_2 \in \mathscr{P}_2(\mathbb{R}^d)$. Let $\Omega_\star \in \{0 \preceq \Omega \preceq I, \operatorname{trace}(\Omega) = k\}$ be optimal between μ_0 and μ_2. Using the triangle inequalities for the Wasserstein distance,

$$\begin{aligned}
\mathcal{S}_k(\mu_0,\mu_2) &= W_2\left[{\Omega_\star^{1/2}}_\sharp\mu_0, {\Omega_\star^{1/2}}_\sharp\mu_2\right] \\
&\leq W_2\left[{\Omega_\star^{1/2}}_\sharp\mu_0, {\Omega_\star^{1/2}}_\sharp\mu_1\right] + W_2\left[{\Omega_\star^{1/2}}_\sharp\mu_1, {\Omega_\star^{1/2}}_\sharp\mu_2\right] \\
&\leq \sup_{\substack{0\preceq\Omega\preceq I \\ \operatorname{trace}(\Omega)=k}} W_2\left[{\Omega^{1/2}}_\sharp\mu_0, {\Omega^{1/2}}_\sharp\mu_1\right] \\
&+ \sup_{\substack{0\preceq\Omega\preceq I \\ \operatorname{trace}(\Omega)=k}} W_2\left[{\Omega^{1/2}}_\sharp\mu_1, {\Omega^{1/2}}_\sharp\mu_2\right] \\
&= \mathcal{S}_k(\mu_0,\mu_1) + \mathcal{S}_k(\mu_1,\mu_2).
\end{aligned}$$

The same argument, used this time with projections, yields the triangle inequalities for $\mathcal{P}_k$. □

2.3 Geometry of Subspace Robust Distances

We prove in this section that SRW distances share several fundamental geometric properties with the Wasserstein distance. The first one states that distances between Diracs match the ground metric:

Lemma 2.3. *For $x,y \in \mathbb{R}^d$ and $k \in [\![d]\!]$,*

$$\mathcal{S}_k(\delta_x,\delta_y) = \|x-y\|.$$

Proof. We use the fact that the pushforward by f of a Dirac at x is the Dirac at $f(x)$, and that the W_2 distance between two Diracs is the Euclidean distance between the points:

$$\mathcal{S}_k(\delta_x,\delta_y) = \max_{\substack{0\preceq\Omega\preceq I \\ \operatorname{trace}(\Omega)=k}} W_2\left({\Omega^{1/2}}_\sharp\delta_x, {\Omega^{1/2}}_\sharp\delta_y\right) = \max_{\substack{0\preceq\Omega\preceq I \\ \operatorname{trace}(\Omega)=k}} \|\Omega^{1/2}(x-y)\|.$$

Since $\|\Omega^{1/2}(x-y)\| \leq \|x-y\|$ with equality for any orthogonal projection matrix Ω onto a subspace $E \in \mathcal{G}_k$ such that $\operatorname{span}(y-x) \subset E$, the result follows. □

Metric Equivalence. Subspace robust Wasserstein distances $\mathcal{S}_k$ are equivalent to the Wasserstein distance W_2:

Proposition 2.2. *For $k \in [\![d]\!]$, $\mathcal{S}_k$ is equivalent to W_2. More precisely, for $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$,*

$$\sqrt{\frac{k}{d}} W_2(\mu,\nu) \leq \mathcal{S}_k(\mu,\nu) \leq W_2(\mu,\nu).$$

Moreover, the constants are tight since

$$\begin{aligned} \mathcal{S}_k(\delta_x,\delta_y) &= W_2(\delta_x,\delta_y) \\ \mathcal{S}_k(\delta_0,\sigma) &= \sqrt{\tfrac{k}{d}} W_2(\delta_0,\sigma) \end{aligned}$$

where $\delta_x, \delta_y, \delta_0$ are Dirac masses at points $x, y, 0 \in \mathbb{R}^d$ and σ is the uniform probability distribution over the centered unit sphere in $\mathbb{R}^d$.

Proof. Let $k \in [\![d]\!]$ and $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$. Let us prove the upper bound on $\mathcal{S}_k$. Using the change of variable formula and the fact that for any $\Omega \in \left\{\Omega \in \mathbb{R}^{d\times d}, 0 \preceq \Omega \preceq I \,;\, \mathrm{trace}(\Omega) = k\right\}$, $\Omega^{1/2}$ is 1-Lipschitz,

$$\begin{aligned} \mathcal{S}_k^2(\mu,\nu) &= \max_{\substack{0\preceq\Omega\preceq I \\ \mathrm{trace}(\Omega)=k}} \min_{\pi\in\Pi(\mu,\nu)} \int \|\Omega^{1/2}(x-y)\|^2 \, d\pi(x,y) \\ &\leq \max_{\substack{0\preceq\Omega\preceq I \\ \mathrm{trace}(\Omega)=k}} \min_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|^2 \, d\pi(x,y) \\ &= W_2^2(\mu,\nu) \end{aligned}$$

which gives the upper bound. For the lower bound, we define $\mathcal{B}_k \subset \mathcal{G}_k$ the (finite) set of k-dimensional subspaces of $\mathbb{R}^d$ spanned by k vectors of the canonical basis of $\mathbb{R}^d$:

$$\mathcal{B}_k = \left\{\mathrm{span}(e_{\sigma(1)}, ..., e_{\sigma(k)}), \sigma \in \mathfrak{S}_d\right\}.$$

Let us now bound $\mathcal{S}_k$ from below:

$$\begin{aligned} \mathcal{S}_k^2(\mu,\nu) &= \min_{\pi\in\Pi(\mu,\nu)} \max_{\substack{0\preceq\Omega\preceq I \\ \mathrm{trace}(\Omega)=k}} \int \|\Omega^{1/2}(x-y)\|^2 \, d\pi(x,y) \\ &\geq \min_{\pi\in\Pi(\mu,\nu)} \max_{E\in\mathcal{B}_k} \int \|P_E(x) - P_E(y)\|^2 \, d\pi(x,y) \\ &= \min_{\pi\in\Pi(\mu,\nu)} \max_{\substack{A\subset[\![d]\!] \\ |A|=k}} \int \sum_{i\in A} (x_i - y_i)^2 \, d\pi(x,y) \\ &= \min_{\pi\in\Pi(\mu,\nu)} \max_{\substack{A\subset[\![d]\!] \\ |A|=k}} \sum_{i\in A} \int (x_i - y_i)^2 \, d\pi(x,y). \end{aligned}$$

For $\pi \in \Pi(\mu, \nu)$,

$$\max_{\substack{A \subset [\![d]\!] \\ |A|=k}} \sum_{i \in A} \int (x_i - y_i)^2 \, d\pi(x, y)$$

is the sum of the k largest elements of $I = \left\{ \int (x_i - y_i)^2 \, d\pi(x, y), i \in [\![d]\!] \right\}$, so it is greater than $\frac{k}{d}$ times the sum of all the elements in I:

$$\mathcal{S}_k^2(\mu, \nu) \geq \frac{k}{d} \min_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \, d\pi(x, y) = \frac{k}{d} W_2^2(\mu, \nu).$$

Note that in the case of $\mu = \delta_0$ and $\nu = \sigma$, the two inequalities in the proof of the lower bound are equalities, hence the tight lower bound constant. □

Dependence on the dimension. We fix $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ and we ask the following question : how does $\mathcal{S}_k(\mu, \nu)$ depend on the dimension $k \in [\![d]\!]$? The following lemma gives a result in terms of the eigenvalues of V_{π_k}, where $\pi_k \in \Pi(\mu, \nu)$ is optimal for some dimension k, then we translate in Proposition 2.3 this result in terms of $\mathcal{S}_k$.

Lemma 2.4. *Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$. For any $k \in [\![d-1]\!]$,*

$$\lambda_{k+1}(V_{\pi_{k+1}}) \leq \mathcal{S}_{k+1}^2(\mu, \nu) - \mathcal{S}_k^2(\mu, \nu) \leq \lambda_{k+1}(V_{\pi_k})$$

where for $L \in [\![d]\!]$, $\pi_L \in \underset{\pi \in \Pi(\mu,\nu)}{\arg\min} \sum_{l=1}^{L} \lambda_l(V_\pi)$.

Proof. Let us first prove the lower bound:

$$\begin{aligned} \mathcal{S}_{k+1}^2(\mu, \nu) &= \sum_{l=1}^{k} \lambda_l(V_{\pi_{k+1}}) + \lambda_{k+1}(V_{\pi_{k+1}}) \\ &\geq \sum_{l=1}^{k} \lambda_l(V_{\pi_k}) + \lambda_{k+1}(V_{\pi_{k+1}}) \\ &= \mathcal{S}_k^2(\mu, \nu) + \lambda_{k+1}(V_{\pi_{k+1}}). \end{aligned}$$

Let us now prove the upper bound:

$$\begin{aligned} \mathcal{S}_{k+1}^2(\mu, \nu) &= \min_{\pi \in \Pi(\mu,\nu)} \sum_{l=1}^{k+1} \lambda_l(V_\pi) \\ &\leq \sum_{l=1}^{k+1} \lambda_l(V_{\pi_k}) \\ &= \mathcal{S}_k^2(\mu, \nu) + \lambda_{k+1}(V_{\pi_k}). \end{aligned}$$

□

Proposition 2.3. *Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$. The sequence $k \mapsto \mathcal{S}_k^2(\mu,\nu)$ is increasing and concave. In particular, for $k \in [\![d-1]\!]$,*

$$\mathcal{S}_{k+1}^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu) \geq \frac{W_2^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu)}{d-k}.$$

Moreover, for any $k \in [\![d-1]\!]$,

$$\mathcal{S}_k(\mu,\nu) \leq \mathcal{S}_{k+1}(\mu,\nu) \leq \sqrt{\frac{k+1}{k}}\mathcal{S}_k(\mu,\nu).$$

Proof. The fact that $\mathcal{S}_k^2(\mu,\nu)$ is an increasing sequence of k is direct using lemma 2.4, since for any $\pi \in \Pi(\mu,\nu)$, V_π has only nonnegative eigenvalues. Let $k \in [\![d-2]\!]$. Then using twice lemma 2.4,

$$\mathcal{S}_{k+2}^2(\mu,\nu) - \mathcal{S}_{k+1}^2(\mu,\nu) \leq \lambda_{k+2}(V_{\pi_{k+1}}) \leq \lambda_{k+1}(V_{\pi_{k+1}}) \leq \mathcal{S}_{k+1}^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu),$$

which shows that $k \mapsto \mathcal{S}_k^2(\mu,\nu)$ is concave.

Let $k \in [\![d-1]\!]$. Although the minoration of $\mathcal{S}_{k+1}^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu)$ is a direct consequence of concavity, we give a direct computation using lemma 2.4:

$$\begin{aligned}
\mathcal{S}_{k+1}^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu) &\geq \lambda_{k+1}(V_{\pi_{k+1}}) \\
&\geq \frac{1}{d-k-1}\sum_{l=k+2}^{d} \lambda_l(V_{\pi_{k+1}}) \\
&= \frac{1}{d-k-1}\left[\sum_{l=1}^{d} \lambda_l(V_{\pi_{k+1}}) - \sum_{l=1}^{k+1} \lambda_l(V_{\pi_{k+1}})\right] \\
&\geq \frac{1}{d-k-1}\left[W_2^2(\mu,\nu) - \mathcal{S}_{k+1}^2(\mu,\nu)\right],
\end{aligned}$$

which implies that

$$(d-k)\left[\mathcal{S}_{k+1}^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu)\right] \geq W_2^2(\mu,\nu) - \mathcal{S}_k^2(\mu,\nu).$$

Finally, the majoration of $\mathcal{S}_k(\mu,\nu)$ is a direct consequence of lemma 2.4:

$$\mathcal{S}_{k+1}^2(\mu,\nu) \leq \mathcal{S}_k^2(\mu,\nu) + \lambda_{k+1}(V_{\pi_k}) \leq \mathcal{S}_k^2(\mu,\nu) + \frac{1}{k}\sum_{l=1}^{k} \lambda_l(V_{\pi_k}) = \frac{k+1}{k}\mathcal{S}_k^2(\mu,\nu).$$

□

Geodesics We have shown in Proposition 2.2 that for any $k \in [\![d]\!]$, $\left(\mathscr{P}_2(\mathbb{R}^d), \mathcal{S}_k\right)$ is a metric space with the same topology as that of the Wasserstein space $\left(\mathscr{P}_2(\mathbb{R}^d), W_2\right)$. We conclude this section by showing that $\left(\mathscr{P}_2(\mathbb{R}^d), \mathcal{S}_k\right)$ is in fact a geodesic length space, and exhibits explicit constant speed geodesics. This can be used to interpolate between measures in $\mathcal{S}_k$ sense.

Proposition 2.4. *Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ and $k \in [\![d]\!]$. Take*

$$\pi_\star \in \operatorname*{arg\,min}_{\pi \in \Pi(\mu,\nu)} \sum_{l=1}^{k} \lambda_l(V_\pi)$$

and let $f_t(x,y) = (1-t)x + ty$. Then the curve

$$t \mapsto \mu_t := {f_t}_\sharp \pi_\star$$

is a constant speed geodesic in $\left(\mathscr{P}_2(\mathbb{R}^d), \mathcal{S}_k\right)$ connecting μ and ν. Consequently, $\left(\mathscr{P}_2(\mathbb{R}^d), \mathcal{S}_k\right)$ is a geodesic space.

Proof. For $s,t \in [0,1]$, put $\pi(s,t) = (f_s, f_t)_\sharp \pi_\star \in \Pi(\mu_s, \mu_t)$, which is our candidate for an optimal transport plan between μ_s and μ_t. Then

$$\begin{aligned}
\mathcal{S}_k^2(\mu_s,\mu_t) &\leq \sum_{l=1}^{k} \lambda_l(V_{\pi(s,t)}) \\
&= \sum_{l=1}^{k} \lambda_l \left\{ \int [f_s(x,y) - f_t(x,y)] \, [f_s(x,y) - f_t(x,y)]^\top \, d\pi_\star(x,y) \right\} \\
&= \sum_{l=1}^{k} \lambda_l \left((t-s)^2 V_{\pi_\star} \right) \\
&= (t-s)^2 \mathcal{S}_k^2(\mu,\nu)
\end{aligned}$$

where we have used

$$\begin{aligned}
f_s(x,y) - f_t(x,y) &= (1-s)x + sy - (1-t)x - ty \\
&= (t-s)(x-y).
\end{aligned}$$

Then for $0 \leq s < t \leq 1$, using the triangular inequality,

$$\begin{aligned}
\mathcal{S}_k(\mu,\nu) &\leq \mathcal{S}_k(\mu,\mu_s) + \mathcal{S}_k(\mu_s,\mu_t) + \mathcal{S}_k(\mu_t,\nu) \\
&\leq (s + (t-s) + (1-t)) \, \mathcal{S}_k(\mu,\nu) = \mathcal{S}_k(\mu,\nu)
\end{aligned}$$

which implies equality everywhere, and in particular optimality for $\pi(s,t)$. Then for all $s,t \in [0,1]$,

$$\mathcal{S}_k(\mu_s,\mu_t) = |t-s| \mathcal{S}_k(\mu,\nu),$$

which shows that the curve (μ_t) has constant speed

$$|\mu_t'| = \lim_{\varepsilon \to 0} \frac{\mathcal{S}_k(\mu_{t+\varepsilon}, \mu_t)}{|\varepsilon|} = \mathcal{S}_k(\mu,\nu),$$

and that the length of the curve (μ_t) is

$$\sup \left\{ \sum_{i=0}^{n-1} \mathcal{S}_k(\mu_{t_i}, \mu_{t_{i+1}}) \,\middle|\, \begin{array}{c} n \geq 1 \\ 0 = t_0 < \ldots < t_n = 1 \end{array} \right\} = \mathcal{S}_k(\mu,\nu),$$

i.e. that (μ_t) is a geodesic connecting μ and ν. □

2.4 Computation

We provide in this section algorithms to compute the saddle point solution of $\mathcal{S}_k$. μ, ν are now discrete with respectively n and m points and weights a and b : $\mu := \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu := \sum_{j=1}^m b_i \delta_{y_j}$. For $k \in [\![d]\!]$, three different objects are of interest: *(i)* the value of $\mathcal{S}_k(\mu, \nu)$, *(ii)* an optimal subspace $E_\star$ obtained through the relaxation for SRW, *(iii)* an optimal transport plan solving SRW. A subspace can be used for dimensionality reduction, whereas an optimal transport plan can be used to compute a geodesic, *i.e.* to interpolate between μ and ν.

2.4.1 Computational challenges to approximate $\mathcal{S}_k$

We observe that solving $\min_{\pi \in \Pi(\mu,\nu)} \sum_{l=1}^k \lambda_l(V_\pi)$ is challenging. Considering a direct projection onto the transportation polytope

$$\Pi(\mu, \nu) = \left\{ \pi \in \mathbb{R}^{n \times m}, \pi \mathbf{1}_m = a,\ \pi^\top \mathbf{1}_n = b \right\}$$

would result in a costly quadratic network flow problem. The Frank-Wolfe algorithm, which does not require such projections, cannot be used directly because the application $\pi \mapsto \sum_{l=1}^k \lambda_l(V_\pi)$ is not smooth.

On the other hand, thanks to Theorem 2.1, solving the maximization problem is easier. Indeed, we can project onto the set of constraints $\mathcal{R} = \{\Omega \in \mathbb{R}^{d \times d}, 0 \preceq \Omega \preceq I\,;\, \mathrm{trace}(\Omega) = k\}$ using Dykstra's projection algorithm [Boyle and Dykstra, 1986]. In this case, we will only get the value of $\mathcal{S}_k(\mu, \nu)$ and an optimal subspace, but not necessarily the actual optimal transport plan due to the lack of uniqueness for OT plans in general.

Smoothing It is well known that saddle points are hard to compute for a bilinear objective [Hammond, 1984]. Computations are greatly facilitated by adding smoothness, which allows the use of saddle point Frank-Wolfe algorithms [Gidel et al., 2017]. Out of the two problems, the maximization problem is seemingly easier. Indeed, we can leverage the framework of regularized OT [Cuturi, 2013] to output, using Sinkhorn's algorithm, a unique optimal transport plan $\pi_\star$ at each inner loop of the maximization. To save time, we remark that initial iterations can be solved with a low accuracy by limiting the number of iterations, and benefit from warm starts, using the scalings computed at the previous iteration, see [Peyré and Cuturi, 2019, §4].

2.4.2 Projected Supergradient Method for SRW

In order to compute SRW and an optimal subspace, we can solve equation (2.3) by maximizing the concave function

$$f : \Omega \mapsto \min_{\pi \in \Pi(\mu,\nu)} \sum_{i,j} d^2_\Omega(x_i, y_j)\, \pi_{i,j} = \min_{\pi \in \Pi(\mu,\nu)} \langle \Omega, V_\pi \rangle$$

Algorithm 1 Projected supergradient method for SRW

Input: Measures (x_i, a_i) and (y_j, b_j), dimension k, learning rate τ_0
$\pi \leftarrow \mathrm{OT}((x,a),(y,b), cost = \|\cdot\|^2)$
$U \leftarrow$ top k eigenvectors of V_π
Initialize $\Omega = UU^\top \in \mathbb{R}^{d\times d}$
for $t = 0$ **to** max_iter **do**
  $\pi \leftarrow \mathrm{OT}((x,a),(y,b), cost = d_\Omega^2)$
  $\tau = \tau_0/(t+1)$
  $\Omega \leftarrow \mathrm{Proj}_{\mathcal{R}}\left[\Omega + \tau V_\pi\right]$
end for
Output: $\Omega, \langle \Omega, V_\pi \rangle$

over the convex set $\mathcal{R}$. Since f is not differentiable, but only superdifferentiable, we can only use a projected supergradient method. This algorithm is outlined in Algorithm 1. Note that by Danskin's theorem [Bertsekas, 1971, Proposition A.22], for any $\Omega \in \mathcal{R}$,

$$\partial f(\Omega) = \mathrm{conv}\left\{ V_{\pi_\star}, \pi_\star \in \underset{\pi\in\Pi(\mu,\nu)}{\arg\min} \langle \Omega, V_\pi \rangle \right\}.$$

2.4.3 Frank-Wolfe using Entropy Regularization

Entropy-regularized optimal transport can be used to compute a unique optimal plan given a subspace. Let $\gamma > 0$ be the regularization strength. In this case, we want to maximize the concave function

$$f_\gamma : \Omega \mapsto \min_{\pi\in\Pi(\mu,\nu)} \langle \Omega, V_\pi \rangle + \gamma \sum_{i,j} \pi_{i,j} \left[\log(\pi_{i,j}) - 1\right]$$

over the convex set $\mathcal{R}$. Since for all $\Omega \in \mathcal{R}$, there is a unique $\pi_\star$ minimizing $\pi \mapsto \langle \Omega, V_\pi \rangle + \gamma \sum_{i,j} \pi_{i,j} \left[\log(\pi_{i,j}) - 1\right]$, f_γ is differentiable. Instead of running a projected gradient ascent on $\Omega \in \mathcal{R}$, we propose to use the Frank-Wolfe algorithm when the regularization strength is positive. Indeed, there is no need to tune a learning rate in Frank-Wolfe, making it easier to use. We only need to compute, for fixed $\pi \in \Pi(\mu,\nu)$, the maximum over $\mathcal{R}$ of $\Omega \mapsto \langle \Omega, V_\pi \rangle$:

Lemma 2.5. *For $\pi \in \Pi(\mu,\nu)$, compute the eigendecomposition of $V_\pi = U \,\mathrm{diag}\,(\lambda_1, ..., \lambda_d)\, U^\top$ with $\lambda_1 \geq ... \geq \lambda_d$. Then for $k \in [\![d]\!]$, a solution to*

$$\max_{\substack{0\preceq\Omega\preceq I \\ \mathrm{trace}(\Omega)=k}} \int d_\Omega^2 \, d\pi$$

is given by $\widehat{\Omega} = U \,\mathrm{diag}\,([\mathbf{1}_k, \mathbf{0}_{d-k}])\, U^\top$.

This algorithm is outlined in algorithm 2.

Proof. Although this is a direct consequence of [Overton and Womersley, 1993], we give an explicit proof. Fix $\pi \in \Pi(\mu, \nu)$. Using the linearity of the trace,

$$\max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega)=k}} \int d_\Omega^2 \, d\pi = \max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega)=k}} \mathrm{trace}(\Omega V_\pi),$$

which is a SDP. Its dual writes

$$\min_{\substack{s \in \mathbb{R}, Z \in \mathbb{R}^{d \times d} \\ Z \succeq 0 \\ Z + sI \succeq V_\pi}} \mathrm{trace}(Z) + ks.$$

Let us write the eigendecomposition of $V_\pi = U \mathrm{diag}(\lambda_1, ..., \lambda_d) U^\top$ with $\lambda_1 \geq ... \geq \lambda_d$. Put $\widehat{\Omega} = U \, \mathrm{diag}\left([\mathbf{1}_k, \mathbf{0}_{d-k}]\right) U^\top$, $\widehat{Z} = U \mathrm{diag}((\lambda_1 - \lambda_k)_+, ..., (\lambda_d - \lambda_k)_+) U^\top$ and $\widehat{s} = \lambda_k$. Then $0 \preceq \widehat{\Omega} \preceq I$, $\mathrm{trace}(\widehat{\Omega}) = k$ and $(\widehat{s}, \widehat{Z})$ is admissible for the dual problem, with corresponding primal and dual values

$$\mathrm{trace}(\widehat{\Omega} V_\pi) = \sum_{l=1}^{k} \lambda_l,$$

$$\mathrm{trace}(\widehat{Z}) + k\widehat{s} = \sum_{l=1}^{k} (\lambda_l - \lambda_k) + k\lambda_k = \sum_{l=1}^{k} \lambda_l.$$

We found primal and dual admissible variables that give the same value, so these variables are optimal. In particular, $\widehat{\Omega}$ is solution to

$$\max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega)=k}} \int d_\Omega^2 \, d\pi.$$

□

2.4.4 Initialization and Stopping Criteria

We propose to initialize Algorithms 1 and 2 with $\Omega_0 = U U^\top$ where $U \in \mathbb{R}^{d \times k}$ is the matrix of the top k eigenvectors (*i.e.* the eigenvectors associated with the top k eigenvalues) of $V_{\pi_\star}$ and $\pi_\star$ is an optimal transport plan between μ and ν. In other words, Ω_0 is the projection matrix onto the k first principal components of the transport-weighted displacement vectors. Note that Ω_0 would be optimal is $\pi_\star$ were optimal for the min-max problem, and that this initialization only costs the equivalent of one iteration.

When entropic regularization is used, Sinkhorn algorithm is run at each iteration of Algorithms 1 and 2. We propose to initialize the potentials in Sinkhorn

Algorithm 2 Frank-Wolfe algorithm for regularized SRW

Input: Measures (x_i, a_i) and (y_j, b_j), dimension k, regularization strength $\gamma > 0$, precision $\varepsilon > 0$
$\pi \leftarrow \text{reg\_OT}((x, a), (y, b), reg = \gamma, cost = \|\cdot\|^2)$
$U \leftarrow$ top k eigenvectors of V_π
Initialize $\Omega = UU^\top \in \mathbb{R}^{d\times d}$
for $t = 0$ **to** max_iter **do**
  $\pi \leftarrow \text{reg\_OT}((x, a), (y, b), reg = \gamma, cost = d_\Omega^2)$
  $U \leftarrow$ top k eigenvectors of V_π
  if $\sum_{l=1}^{k} \lambda_l(V_\pi) - \langle \Omega, V_\pi \rangle \leq \varepsilon \langle \Omega, V_\pi \rangle$ **then**
    break
  end if
  $\widehat{\Omega} \leftarrow U \operatorname{diag}([\mathbf{1}_k, \mathbf{0}_{d-k}])\, U^\top$
  $\tau = 2/(2 + t)$
  $\Omega \leftarrow (1 - \tau)\Omega + \tau\widehat{\Omega}$
end for
Output: $\Omega, \pi, \langle \Omega, V_\pi \rangle$

algorithm with the latest computed potentials, so that the number of iterations in Sinkhorn algorithm should be small after a few iterations of Algorithms 1 or 2.

We sometimes need to compute $\mathcal{S}_k(\mu, \nu)$ for all $k \in [\![d]\!]$, for example to choose the optimal k with an "elbow" rule. To speed the computations up, we propose to compute this sequence iteratively from $k = d$ to $k = 1$. At each iteration, *i.e.* for each dimension k, we initialize the algorithm with $\Omega_0 = UU^\top$, where $U \in \mathbb{R}^{d\times k}$ is the matrix of the top k eigenvectors of $V_{\pi_{k+1}}$ and π_{k+1} is the optimal transport plan for dimension $k + 1$. We also initialize the Sinkhorn algorithm with the latest computed potentials.

Instead of running a fixed number of iterations in Algorithm 2, we propose to stop the algorithm when the computation error is smaller than a fixed threshold ε. The computation error at iteration t is:

$$\frac{|\mathcal{S}_k(\mu, \nu) - \widehat{\mathcal{S}_k}(t)|}{\mathcal{S}_k(\mu, \nu)} \leq \frac{\Delta(t)}{\widehat{\mathcal{S}_k}(t)}$$

where $\widehat{\mathcal{S}_k}(t)$ is the computed "max-min" value and $\Delta(t)$ is the duality gap at iteration t. We stop as soon as $\Delta(t)/\widehat{\mathcal{S}_k}(t) \leq \varepsilon$.

2.5 Experiments

We first compare SRW with the experimental setup used to evaluate FactoredOT [Forrow et al., 2019]. We then study the ability of SRW distances

to capture the dimension of sampled measures by looking at their value for increasing dimensions k, as well as their robustness to noise.

2.5.1 Fragmented Hypercube

We first consider $\mu = \mathcal{U}([-1,1])^d$ to be uniform over an hypercube, and $\nu = T_\sharp \mu$ the pushforward of μ under the map $T(x) = x + 2\,\mathrm{sign}(x) \odot (\sum_{k=1}^{k^*} e_k)$, where sign is taken elementwise, $k^* \in [\![d]\!]$ and $(e_1, ..., e_d)$ is the canonical basis of $\mathbb{R}^d$. The map T splits the hypercube into four different hyperrectangles. T is a subgradient of a convex function, so by Brenier's theorem [1991] it is an optimal transport map between μ and $\nu = T_\sharp \mu$ and

$$W_2^2(\mu, \nu) = \int \|x - T(x)\|^2 \, d\mu(x) = 4k^*.$$

Note that for any x, the displacement vector $T(x) - x$ lies in the k^*-dimensional subspace $\mathrm{span}\{e_1, ..., e_{k^*}\} \in \mathcal{G}_{k^*}$, which is optimal. This means that for $k \geq k^*$, $\mathcal{S}_k^2(\mu, \nu)$ is constant equal to $4k^*$. We show the interest of plotting, based on two empirical distributions $\hat{\mu}$ from μ and $\hat{\nu}$ from ν, the sequence $k \mapsto \mathcal{S}_k^2(\hat{\mu}, \hat{\nu})$, for different values of k^*. That sequence is increasing concave by proposition 2.3, and increases more slowly after $k = k^*$, as can be seen on Figure 2.2. This is the case because the last $d - k^*$ dimensions only represent noise, but is recovered in our plot.

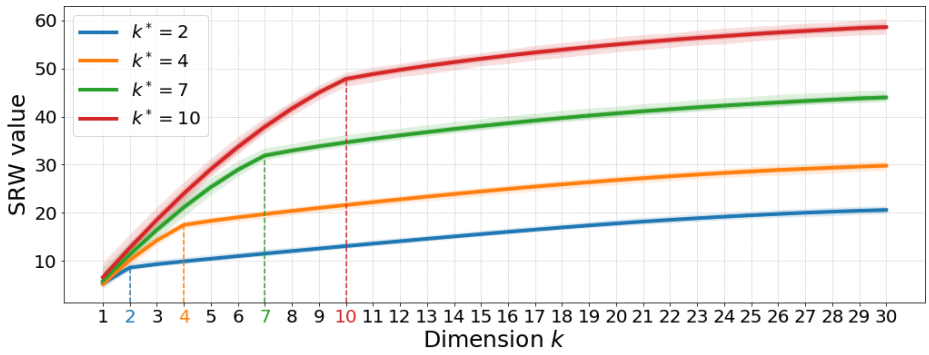


Figure 2.2: $\mathcal{S}_k^2(\hat{\mu}, \hat{\nu})$ depending on the dimension $k \in [\![d]\!]$, for $k^* \in \{2, 4, 7, 10\}$, where $\hat{\mu}, \hat{\nu}$ are empirical measures from μ and ν respectively with 100 points each. Each curve is the mean over 100 samples, and shaded area show the min and max values.

We consider next $k^* = 2$, and choose from the result of Figure 2.2, $k = 2$. We look at the estimation error $|W_2^2(\mu, \nu) - \mathcal{S}_k^2(\hat{\mu}, \hat{\nu})|$ where $\hat{\mu}, \hat{\nu}$ are empirical measures from μ and ν respectively with n points each. In Figure 2.3, we plot this estimation error depending on the number of points n. In Figure 2.4, we plot the subspace estimation error $\|\Omega^* - \widehat{\Omega}\|$ depending on n, where Ω^* is the optimal projection matrix onto $\mathrm{span}\{e_1, e_2\}$.

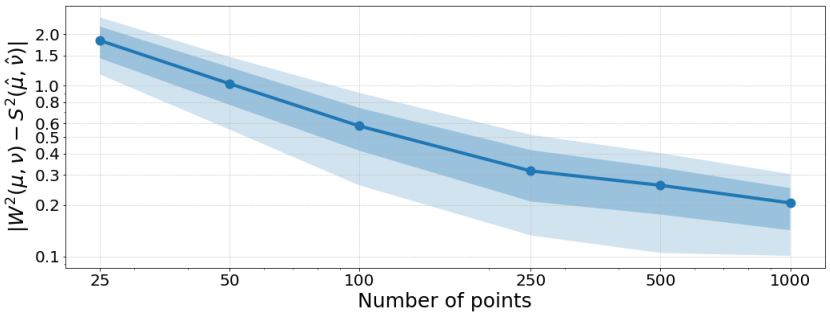


Figure 2.3: Mean estimation error over 500 random samples for n points, $n \in \{25, 50, 100, 250, 500, 1000\}$. The shaded areas represent the 10%-90% and 25%-75% quantiles over the 500 samples.

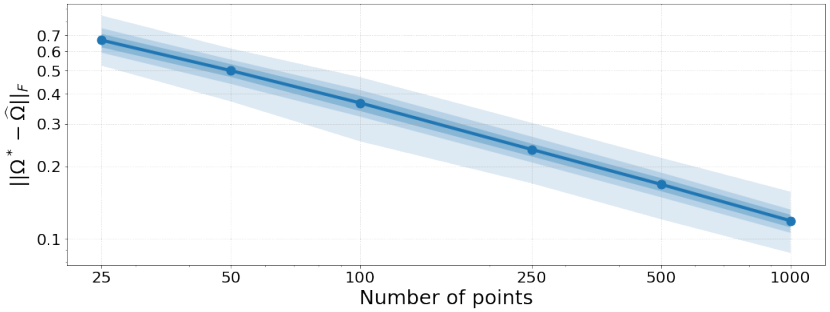


Figure 2.4: Mean estimation of the subspace estimation error over 500 samples, depending on $n \in \{25, 50, 100, 250, 500, 1000\}$. The shaded areas represent the 10%-90% and 25%-75% quantiles over the 500 samples.

We also plot the optimal transport plan (in the sense of W_2, Figure 2.5 left) and the optimal transport plan (in the sense of $\mathcal{S}_2$) between $\hat{\mu}$ and $\hat{\nu}$ (with $n = 250$ points each, Figure 2.5 right).

2.5.2 Disk to Annulus

Let $k^* \in [\![d]\!]$. We now consider μ the uniform distribution over the k^*-dimensional disk embedded in $\mathbb{R}^d$,

$$\mu = \mathcal{U}(\{x \in \mathbb{R}^d, \|(x_1, ..., x_{k^*})\| \leq 1, x_i \in [0, 1] \text{ for } i = (k^* + 1), ..., d\})$$

and ν the uniform distribution over a k^*-dimensional annulus (cylinder) embedded in $\mathbb{R}^d$,

$$\nu = \mathcal{U}(\{x \in \mathbb{R}^d, 2 \leq \|(x_1, ..., x_{k^*})\| \leq 3, x_i \in [0, 1] \text{ for } i = (k^* + 1), ..., d\}).$$

We do the same experiments as for the fragmented hypercube. Based on two empirical distributions $\hat{\mu}$ from μ and $\hat{\nu}$ from ν, we plot in Figure 2.6 the

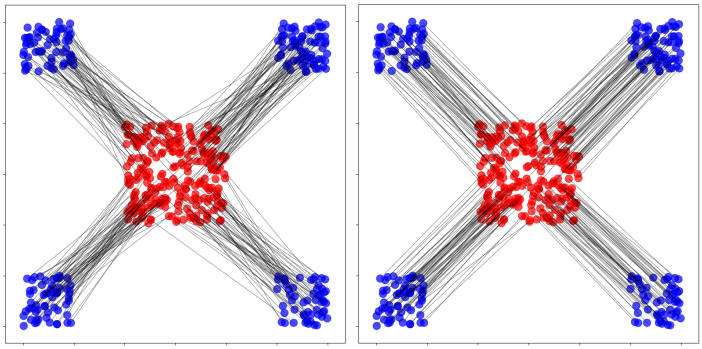


Figure 2.5: Fragmented hypercube, $n = 250$, $d = 30$. Optimal mapping in the Wasserstein space (left) and in the SRW space (right). Geodesics in the SRW space are robust to statistical noise.

sequence $k \mapsto \mathcal{S}_k^2(\hat{\mu}, \hat{\nu})$, for different values of k^*. An "elbow" shows at $k = k^*$, because the last $d - k^*$ dimensions only represent noise, which is recovered in our plot.

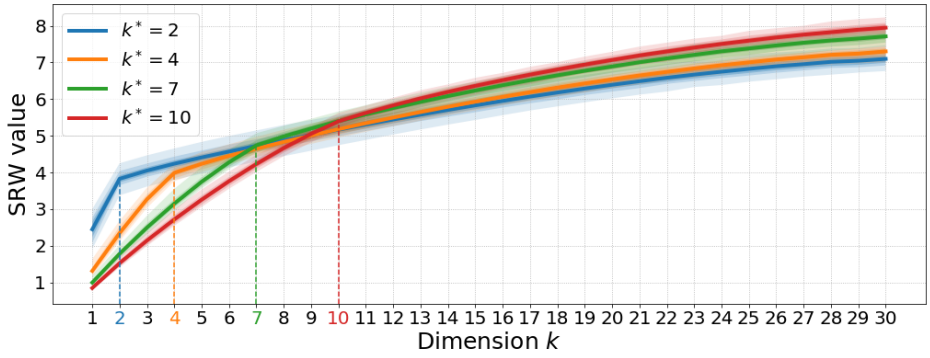


Figure 2.6: $\mathcal{S}_k^2(\hat{\mu}, \hat{\nu})$ depending on the dimension $k \in [\![d]\!]$, for $k^* \in \{2, 4, 7, 10\}$, where $\hat{\mu}, \hat{\nu}$ are empirical measures from μ and ν respectively with 100 points each. Each curve is the mean over 100 samples, and shaded area show the minimum and maximum values.

We now consider $k^* = 2$, and choose $k = 2$. We will need the value of $W_2^2(\mu, \nu)$, which we compute by exhibiting the optimal Monge map using the Brenier theorem:

$$W_2^2(\mu, \nu) = \frac{14}{5} + \frac{8}{5\sqrt{5}} \log\left(\frac{3+\sqrt{5}}{2}\right) \approx 3.48865.$$

We plot in Figure 2.7 the estimation error $|W_2^2(\mu, \nu) - \mathcal{S}_k^2(\hat{\mu}, \hat{\nu})|$ depending on the number of points n in the empirical measures $\hat{\mu}, \hat{\nu}$ from μ and ν. In Figure 2.8, we plot the subspace estimation error $\|\Omega^* - \widehat{\Omega}\|$ depending on n, where Ω^* is the optimal projection matrix onto $\mathrm{span}\{e_1, e_2\}$.

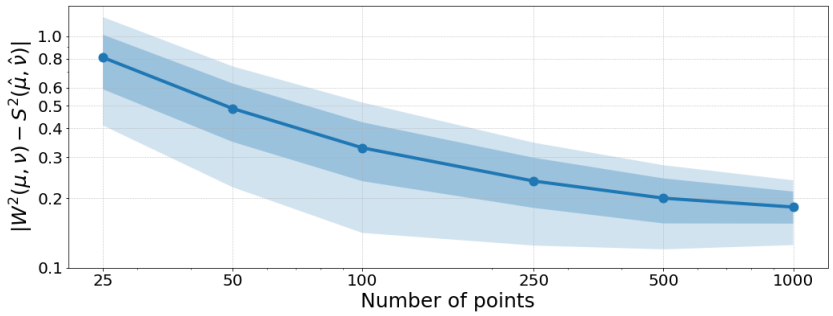


Figure 2.7: Mean estimation error over 500 random samples for $n \in \{25, 50, 100, 250, 500, 1000\}$. The shaded areas represent the 10%-90% and 25%-75% quantiles over the 500 samples.

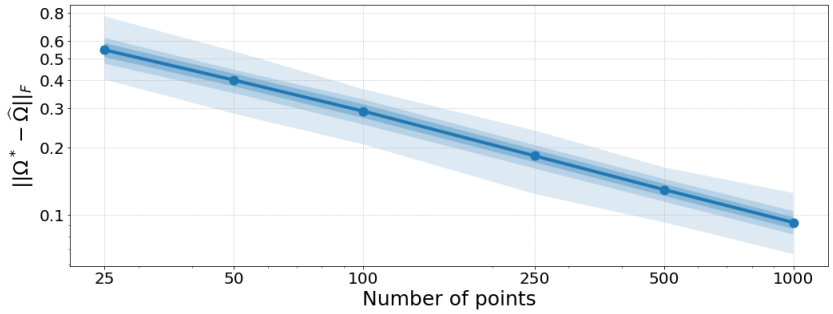


Figure 2.8: Mean estimation of the subspace estimation error over 500 samples, depending on $n \in \{25, 50, 100, 250, 500, 1000\}$. The shaded areas represent the 10%-90% and 25%-75% quantiles over the 500 samples.

We plot in Figure 2.9 the optimal transport plan (in the sense of W_2, Figure 2.9 left) and the optimal transport plan (in the sense of $\mathcal{S}_2$, Figure 2.9 right) between $\hat{\mu}$ and $\hat{\nu}$ (with $n = 250$ points each).

2.5.3 Robustness, with 20-dimensional Gaussians

We consider $\mu = \mathcal{N}(0, \Sigma_1)$ and $\nu = \mathcal{N}(0, \Sigma_2)$, with $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ semidefinite positive of rank k. It means that the supports of μ and ν are k-dimensional subspaces of $\mathbb{R}^d$. Although those two subspaces are k-dimensional, they may be different. Since the union of two k-dimensional subspaces is included in a $2k$-dimensional subspace, for any $l \geq 2k$, $\mathcal{S}_l^2(\mu, \nu) = W_2^2(\mu, \nu)$.

For our experiment, we simulated 100 independent couples of covariance matrices Σ_1, Σ_2 in dimension $d = 20$, each having independently a Wishart distribution with $k = 5$ degrees of freedom. For each couple of matrices, we draw $n = 100$ points from $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$ and considered $\hat{\mu}$ and $\hat{\nu}$ the

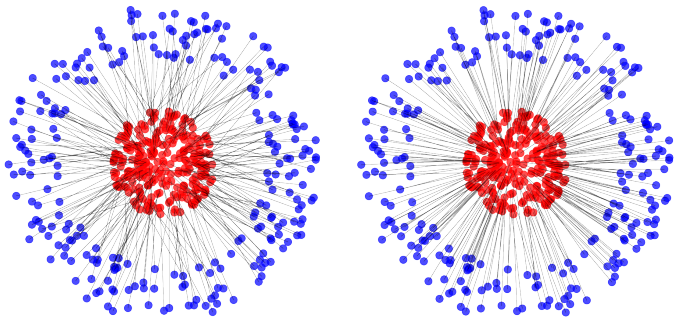


Figure 2.9: Disk to annulus, $n = 250$, $d = 30$. Optimal mapping in the Wasserstein space (left) and in the SRW space (right). Geodesics in the SRW space are robust to statistical noise.

empirical measures on those points. In Figure 2.10, we plot the mean (over the 100 samples) of $l \mapsto \mathcal{S}_l^2(\hat{\mu}, \hat{\nu})/W_2^2(\hat{\mu}, \hat{\nu})$. We plot the same curve for noisy data, where each point was added a $\mathcal{N}(0, I)$ random vector. With moderate noise, the data is only approximately on the two $k = 5$-dimensional subspaces, but the SRW does not vary too much.

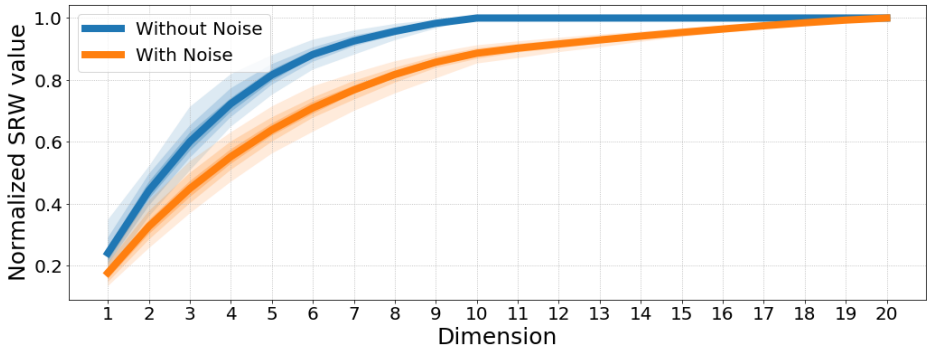


Figure 2.10: Mean normalized SRW distance, depending on the dimension. The shaded area show the 10%-90% and 25%-75% quantiles and the minimum and maximum values over the 100 samples.

2.5.4 $\mathcal{S}_k$ is Robust to Noise

As in experiment 2.5.3, we consider 100 independent samples of couples $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$, each following independently a Wishart distribution with $k = 5$ degrees of freedom. For each couple, we draw $n = 100$ points from $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$ and consider the empirical measures $\hat{\mu}$ and $\hat{\nu}$ on those points. We then gradually add Gaussian noise $\sigma\mathcal{N}(0, I)$ to the points, giving measures $\hat{\mu}_\sigma$, $\hat{\nu}_\sigma$. In Figure 2.11, we plot the mean (over the 100 samples) of the relative

errors

$$\sigma \mapsto \frac{|\mathcal{S}_5^2(\hat{\mu}_\sigma, \hat{\nu}_\sigma) - \mathcal{S}_5^2(\hat{\mu}_0, \hat{\nu}_0)|}{\mathcal{S}_5^2(\hat{\mu}_0, \hat{\nu}_0)}$$

and

$$\sigma \mapsto \frac{|W_2^2(\hat{\mu}_\sigma, \hat{\nu}_\sigma) - W_2^2(\hat{\mu}_0, \hat{\nu}_0)|}{W_2^2(\hat{\mu}_0, \hat{\nu}_0)}.$$

Note that for small noise level, the imprecision in the computation of the SRW distance adds up to the error caused by the added noise. SRW distances seem more robust to noise than the Wasserstein distance when the noise has moderate to high variance.

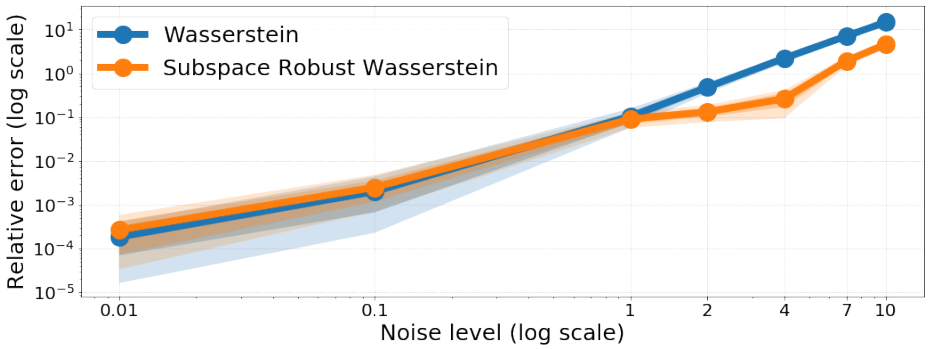


Figure 2.11: Mean SRW distance over 100 samples, depending on the noise level. Shaded areas show the min-max values and the 10%-90% quantiles.

2.5.5 Computation time

We consider the Fragmented Hypercube experiment, with increasing dimension d and fixed $k^* = 2$. Using $k = 2$ and Algorithm 2 with $\gamma = 0.1$ and stopping threshold $\varepsilon = 0.05$, we plot in Figure 2.12 the mean computation time of both SRW and Wasserstein distances on GPU, over 100 random samplings with $n = 100$. It shows that SRW computation is quadratic in dimension d, because of the eigendecomposition of matrix V_π in Algorithm 2.

2.5.6 Real Data Experiment

We consider the scripts of seven movies. Each script is transformed into a list of words, and using word2vec [Mikolov et al., 2018], into a measure over $\mathbb{R}^{300}$ where the weights correspond to the frequency of the words. The complete vocabulary used consists of the 20000 most common words in English, except for the 2000 most common words, hence a total size of 18000 words. All the words in a movie script that whether do not belong to the vocabulary list, are digits or begin with a capital letter, are deleted. The remaining words form a

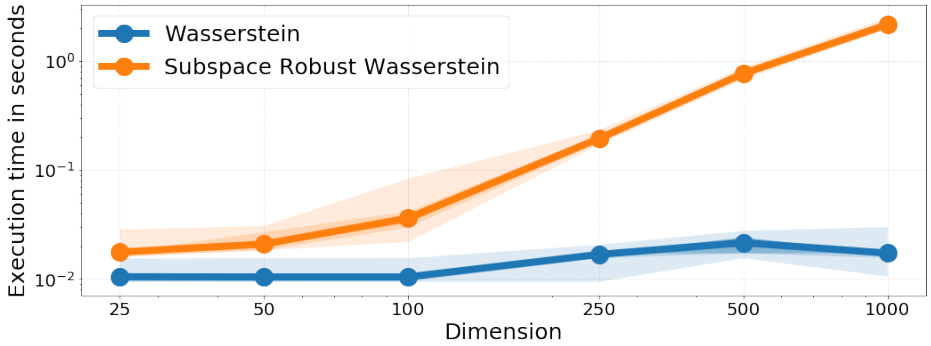


Figure 2.12: Mean computation times on GPU (log-log scale). The shaded areas show the minimum and maximum values over the 100 experiments.

discrete measure in $\mathbb{R}^{300}$, with the weights proportional to their frequency in the movie script.

We then compute the SRW distance between all pairs of films: see Figure 2.13 for the SRW values. Movies with a same genre or thematic tend to be closer to each other: this can be visualized by running a two-dimensional metric multidimensional scaling (mMDS) on the SRW distances, as shown in Figure 2.14 (left).

	D	*G*	*I*	*KB1*	*KB2*	*TM*	*T*
D	0	0.186	0.186	0.195	0.203	0.186	**0.171**
G	0.186	0	**0.173**	0.197	0.204	0.176	0.185
I	0.186	0.173	0	0.196	0.203	**0.171**	0.181
KB1	0.195	0.197	0.196	0	**0.165**	0.190	0.180
KB2	0.203	0.204	0.203	**0.165**	0	0.194	0.180
TM	0.186	0.176	**0.171**	0.190	0.194	0	0.183
T	**0.171**	0.185	0.181	0.180	0.180	0.183	0

Figure 2.13: $\mathcal{S}_k^2$ distances between different movie scripts. Bold values correspond to the minimum of each line. *D*=Dunkirk, *G*=Gravity, *I*=Interstellar, *KB1*=Kill Bill Vol.1, *KB2*=Kill Bill Vol.2, *TM*=The Martian, *T*=Titanic.

In Figure 2.14 (right), we display the projection of the two measures associated with films *Kill Bill Vol.1* and *Interstellar* onto their optimal subspace. We compute the first (weighted) principal component of each projected measure, and find among the whole dictionary their 5 nearest neighbors in terms of cosine similarity. For *Kill Bill Vol.1* , these are: 'swords', 'hull', 'sword', 'ice', 'blade'. For *Interstellar*, they are: 'spacecraft', 'planets', 'satellites', 'asteroids', 'planet'. The optimal subspace recovers the semantic dissimilarities between the two films.

Figure 2.14: *Left:* Metric MDS projection for the distances of Figure 2.13. *Right:* Optimal 2-dimensional projection between *Kill Bill Vol.1* (red) and *Interstellar* (blue). Words appearing in both scripts are displayed in violet. For clarity, only the 30 most frequent words of each script are displayed.

Conclusion

In this chapter, we have proposed a new family of optimal transport distances with robust properties. These distances take a particular interest when used with a squared-Euclidean cost, in which case they have several properties, both theoretical and computational. These distances share important properties with the 2-Wasserstein distance, yet seem far more robust to random perturbation of the data and able to capture better signal. We have provided algorithmic tools to compute these SRW distances. They come at a relatively modest overhead, given that they require using regularized OT as the inner loop of a Frank-Wolfe type algorithm.

2.6 Supplementary Results about Projection Robust Wasserstein Distances

In this section, we prove some basic properties of projection robust Wasserstein distances $\mathcal{P}_k$. First note that the definition of $\mathcal{P}_k$ makes sense, since for any $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, $k \in [\![d]\!]$ and $E \in \mathcal{G}_k$, $P_{E\sharp}\mu$ and $P_{E\sharp}\nu$ have a second moment (for orthogonal projections are 1-Lipschitz).

$\mathcal{P}_k$ is also well posed, since one can prove the existence of a maximizing subspace. To prove this, we will need the following lemma stating that the admissible set of couplings between the projected measures are exactly the projections of the admissible couplings between the original measures:

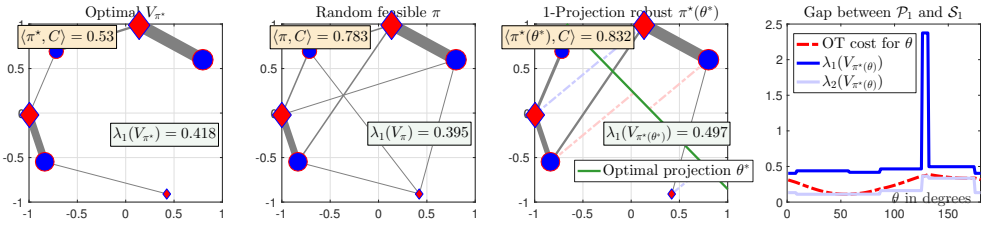


Figure 2.15: This figure should be compared to Figure 2.1. We also present an example for which the explicit computation of projection $\mathcal{P}_k$ and subspace $\mathcal{S}_k$ robust Wasserstein distances can be carried out explicitly, by simple enumeration. Unlike in Figure 2.1, and as can be seen in the rightmost plot, these two quantities do not coincide here. That plot reveals that the minimum across all maximal eigenvalues of second order moment matrices computed on all optimal OT plans obtained by enumerating all lines (the subspace robust quantity) is strictly larger than the worst possible projection cost.

Lemma 2.6. *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ Borel and $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$. Then $\Pi(f_\sharp \mu, f_\sharp \nu) = \{(f \otimes f)_\sharp \pi, \pi \in \Pi(\mu, \nu)\}$.*

Proof. Let $E = \mathrm{Im}(f) \subset \mathbb{R}^d$ and $\pi \in \Pi(\mu, \nu)$. Then $(f \otimes f)_\sharp \pi$ is an admissible transport plan between $f_\sharp \mu$ and $f_\sharp \nu$. Indeed, for any Borel set $A \subset E$, $(f \otimes f)_\sharp \pi(A \times E) = \pi(f^{-1}(A) \times f^{-1}(E)) = \pi(f^{-1}(A) \times \mathbb{R}^d) = \mu(f^{-1}(A)) = f_\sharp \mu(A)$, so $(f \otimes f)_\sharp \pi$ has first marginal $f_\sharp \mu$, and likewise, has second marginal $f_\sharp \nu$, *i.e.* $(f \otimes f)_\sharp \pi \in \Pi(f_\sharp \mu, f_\sharp \nu)$.
Conversely, let $\rho \in \Pi(f_\sharp \mu, f_\sharp \nu)$. Let us construct $\pi \in \Pi(\mu, \nu)$ such that $(f \otimes f)_\sharp \pi = \rho$. For any Borel sets $A, B \subset \mathbb{R}^d$, put

$$\pi(A \times B) = \frac{\rho(f(A) \times f(B))\mu(A)\nu(B)}{f_\sharp \mu(f(A))\, f_\sharp \nu(f(B))}$$

if $f_\sharp \mu(f(A)) \neq 0$ and $f_\sharp \mu(f(B)) \neq 0$, and $\pi(A, B) = 0$ otherwise. Then $\pi \in \Pi(\mu, \nu)$. Indeed, for any Borel set $A \subset \mathbb{R}^d$, $\pi(A \times \mathbb{R}^d) = \rho(f(A) \times E)\frac{\mu(A)}{f_\sharp \mu(f(A))} = \mu(A)$ if $f_\sharp \mu(f(A)) \neq 0$ and $\pi(A \times \mathbb{R}^d) = 0$ if $f_\sharp \mu(f(A)) = 0$. But then, $\mu(A) \leq \mu(f^{-1}(f(A))) = f_\sharp \mu(f(A)) = 0$ so $\mu(A) = 0 = \pi(A \times \mathbb{R}^d)$. The same calculations give the result for the second marginal.
There remains to prove that $(f \otimes f)_\sharp \pi = \rho$. For any Borel sets $A, B \subset E$, noting that $f(f^{-1}(A)) = A$ and $f(f^{-1}(B)) = B$,

$$\begin{aligned}(f \otimes f)_\sharp \pi(A \times B) &= \pi(f^{-1}(A) \times f^{-1}(B)) \\ &= \rho(A \times B)\frac{\mu(f^{-1}(A))}{f_\sharp \mu(A)}\frac{\nu(f^{-1}(B))}{f_\sharp \mu(B)} \\ &= \rho(A \times B)\end{aligned}$$

if $f_\sharp \mu(A) \neq 0$ and $f_\sharp \nu(B) \neq 0$. Otherwise, $(f \otimes f)_\sharp \pi(A \times B) = 0$ and $\rho(A \times B) \leq \min\{\rho(A \times E), \rho(E \times B)\} = \min\{f_\sharp \mu(A), f_\sharp \nu(B)\} = 0$, so $\rho(A \times B) = (f \otimes f)_\sharp \pi(A \times B) = 0$. □

This can be used to get the following result:

Proposition 2.5. *For $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ and $k \in [\![d]\!]$, there exists a subspace $E_\star \in \mathcal{G}_k$ such that*

$$\mathcal{P}_k(\mu, \nu) = W_2\left(P_{E_\star \sharp}\mu, P_{E_\star \sharp}\nu\right).$$

Proof. We endow the Grassmannian $\mathcal{G}_k$ with the metric topology associated with metric $(E, F) \mapsto \|P_E - P_F\|$, where P_E and P_F are respectively the linear projectors onto E and F. Then it is well known that $\mathcal{G}_k$ is compact under this topology.

We only have to show that, for $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, the map $f : E \mapsto W_2\left(P_{E\sharp}\mu, P_{E\sharp}\nu\right)$ is upper semicontinuous. For any orthogonal projector P, using lemma 2.6,

$$\begin{aligned} W_2^2\left(P_\sharp\mu, P_\sharp\nu\right) &= \min_{\rho \in \Pi(P_\sharp\mu, P_\sharp\nu)} \int \|x-y\|^2 \, d\rho(x,y) \\ &= \min_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|^2 \, d(P\otimes P)_\sharp\pi(x,y) \\ &= \min_{\pi \in \Pi(\mu,\nu)} \int \|P(x-y)\|^2 \, d\pi(x,y). \end{aligned}$$

Since for $\pi \in \Pi(\mu,\nu)$, the application $P \mapsto \int \|P(x-y)\|^2 \, d\pi(x,y)$ is continuous, the application $g : P \mapsto W_2^2\left(P_\sharp\mu, P_\sharp\nu\right)$ is upper semicontinuous as the minimum of continuous functions. As the application $h : E \mapsto P_E$ is continuous, and $x \mapsto \sqrt{x}$ is nondecreasing, $f = \sqrt{g \circ h}$ is upper semicontinuous. □

Note that we could define projection robust Wasserstein distances for any $p \geq 1$ by:

$$\sup_{E \in \mathcal{G}_k} W_p\left(P_{E\sharp}\mu, P_{E\sharp}\nu\right).$$

Then there is still existence of optimal subspaces, and it defines a distance over $\mathscr{P}_p(\mathbb{R}^d)$.

To prove the identity of indiscernibles, we use the following Lemma due to [Rényi, 1952], generalizing a theorem by [Cramér and Wold, 1936]:

Lemma 2.7. *Let $(E_j)_{j\in J}$ be a family of subspaces of $\mathbb{R}^d$ such that $\bigcup_{j\in J} E_j = \mathbb{R}^d$. Let $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$ such that for all $j \in J$, $P_{E_j\sharp}\mu = P_{E_j\sharp}\nu$. Then $\mu = \nu$.*

Proof. Let $j \in J$. Since $P_{E_j\sharp}\mu = P_{E_j\sharp}\nu$, their characteristic functions are equal, *i.e.* for all $t \in \mathbb{R}^d$,

$$\begin{aligned} \int \exp i\langle t|x\rangle \, dP_{E_j\sharp}\mu(x) &= \int \exp i\langle t|x\rangle \, dP_{E_j\sharp}\nu(x) \\ \Leftrightarrow \int \exp i\langle t|P_{E_j}x\rangle \, d\mu(x) &= \int \exp i\langle t|P_{E_j}x\rangle \, d\nu(x) \\ \Leftrightarrow \int \exp i\langle P_{E_j}t|x\rangle \, d\mu(x) &= \int \exp i\langle P_{E_j}t|x\rangle \, d\nu(x) \end{aligned}$$

i.e. the characteristic functions of μ and ν coincide on E_j, for all $j \in J$. Since the subspaces $(E_j)_{j\in J}$ cover the whole space $\mathbb{R}^d$, μ and ν have the same characteristic functions on $\mathbb{R}^d$, hence $\mu = \nu$. □

Chapter 3

Regularized Optimal Transport is Ground-Cost Adversarial

3.1 Introduction

The key to using optimal transport (OT) in applications lies in the different forms of regularization of the original OT problem (KP). Adding a small convex regularization to the classical linear cost not only helps on the algorithmic side, by convexifying the objective and allowing for faster solvers, but also introduces a regularity trade-off that prevents from overfitting on data measures.

Regularizing OT Although entropy-regularized OT is the most studied regularization of OT, due to its algorithmic advantages [Cuturi, 2013], several other convex regularizations of the transport plan have been proposed in the community: quadratically-regularized OT [Essid and Solomon, 2017], OT with capacity constraints [Korman and McCann, 2015], Group-Lasso regularized OT [Courty et al., 2016], OT with Laplacian regularization [Flamary et al., 2014], Tsallis Regularized OT [Muzellec et al., 2017], among others. On the other hand, regularizing the dual Kantorovich problem was shown in [Liero et al., 2018] to be equivalent to unbalanced OT, that is optimal transport with relaxed marginal constraints.

Understanding why regularization helps The question of understanding why regularizing OT proves critical has triggered several approaches. A compelling reason is statistical: Although classical OT suffers from the curse of dimensionality, as its empirical version converges at a rate of order $\mathcal{O}\left(n^{-1/d}\right)$ [Dudley, 1969, Fournier and Guillin, 2015, Weed and Bach, 2019], regularized OT and more precisely Sinkhorn divergences have a sample complexity of $\mathcal{O}\left(1/\sqrt{n}\right)$ [Genevay et al., 2019, Mena and Niles-Weed, 2019]. Entropic OT was also shown to perform maximum likelihood estimation in the Gaussian deconvolution model [Rigollet and Weed, 2018]. Taking another approach, Des-

sein et al. [2018], Blondel et al. [2018] have considered general classes of convex regularizations and characterized them from a more geometrical perspective.

Robustness Recently, several papers [Genevay et al., 2018, Flamary et al., 2018, Deshpande et al., 2019, Kolouri et al., 2019, Niles-Weed and Rigollet, 2019, Paty and Cuturi, 2019] have proposed to maximize OT with respect to the ground-cost function, which can in turn be interpreted in light of ground metric learning [Cuturi and Avis, 2014]. This approach can also be viewed as an instance of robust optimization [Ben-Tal and Nemirovski, 1998, Ben-Tal et al., 2009, Bertsimas et al., 2011]: instead of considering a data-dependent, hence unstable minimization problem $\min_x f_{\hat{\theta}}(x)$ where $\hat{\theta}$ represents the data, the robust optimization literature adversarially chooses the parameters θ in a neighborhood of the data: $\min_x \max_{\theta \in \Theta} f_\theta(x)$. Continuing along these lines, we make a connection between *regularizing* and *maximizing* OT.

Contributions Our main goal is to provide a novel interpretation of regularized optimal transport in terms of ground cost robustness: regularizing OT amounts to maximizing **un**regularized OT with respect to the ground cost. Our contributions are:

1. We show that any convex regularization of the transport plan corresponds to ground-cost robustness (§ 3.3);
2. We reinterpret classical regularizations of OT in the ground-cost adversarial setting (§ 3.4);
3. We prove, under some technical assumption, a duality theorem for regularized OT, which we use to show that under the same assumption, there exists an optimal adversarial ground-cost that is separable (§ 3.5);
4. We extend ground-cost robustness to the case of more than two measures (§ 3.6);
5. We propose algorithms to solve the above-mentioned problems (§3.7) and illustrate them on data (§ 3.8).

3.2 Background and Notations

Let $\mathcal{X}$ be a compact Hausdorff space, and define $\mathscr{P}(\mathcal{X})$ the set of Borel probability measures over $\mathcal{X}$. We write $\mathcal{C}(\mathcal{X})$ for the set of continuous functions from $\mathcal{X}$ to $\mathbb{R}$, endowed with the supremum norm. For $\phi, \psi \in \mathcal{C}(\mathcal{X})$, we write $\phi \oplus \psi \in \mathcal{C}(\mathcal{X}^2)$ for the function $\phi \oplus \psi : (x, y) \mapsto \phi(x) + \psi(y)$.

In this chapter, all vectors will be denoted with **bold** symbols. For a Boolean assertion A, we write $\iota(A)$ for its indicator function $\iota(A) = 0$ if A is true and $\iota(A) = +\infty$ otherwise.

Space of Measures Since $\mathcal{X}$ is compact, the dual space of $\mathcal{C}(\mathcal{X}^2)$ is the set $\mathscr{M}(\mathcal{X}^2)$ of Borel finite signed measures over $\mathcal{X}^2$. For $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R}$, we recall that F is Fréchet-differentiable at π if there exists $\nabla F(\pi) \in \mathcal{C}(\mathcal{X}^2)$ such that for any $h \in \mathscr{M}(\mathcal{X}^2)$, as $t \to 0$

$$F(\pi + th) = F(\pi) + t \int \nabla F(\pi)\, dh + o(t).$$

Similarly, $G : \mathcal{C}(\mathcal{X}^2) \to \mathbb{R}$ is Fréchet-differentiable at c if there exists $\nabla G(c) \in \mathscr{M}(\mathcal{X}^2)$ such that for any $h \in \mathcal{C}(\mathcal{X}^2)$, as $t \to 0$

$$G(c + th) = G(c) + t \int h\, d\nabla G(c) + o(t).$$

Legendre–Fenchel Transformation For any functional $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$, we can define its convex conjugate $F^* : \mathcal{C}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$ and biconjugate $F^{**} : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$ as

$$F^*(c) := \sup_{\pi \in \mathscr{M}(\mathcal{X}^2)} \int c\, d\pi - F(\pi),$$
$$F^{**}(\pi) := \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \int c\, d\pi - F^*(c).$$

F^* is always lower semi-continuous (lsc) and convex as the supremum of continuous linear functions.

Specific notations For $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$, we write $\mathrm{dom}(F) = \left\{\pi \in \mathscr{M}(\mathcal{X}^2), F(\pi) < +\infty\right\}$ for its domain and will say that F is proper if $\mathrm{dom}(F) \neq \emptyset$.

We denote by $\mathscr{F}$ the set of proper lsc convex functions $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$, and for $\mu, \nu \in \mathscr{P}(\mathcal{X})$, we define the set $\mathscr{F}(\mu, \nu)$ of lsc convex functions that are proper on $\Pi(\mu, \nu)$:

$$\mathscr{F}(\mu, \nu) = \{F \in \mathscr{F}, \exists \pi \in \Pi(\mu, \nu), F(\pi) < +\infty\}.$$

3.3 Ground-Cost Adversarial Optimal Transport

3.3.1 Definition

Instead of considering the classical *linear* formulation of optimal transport (KP), we consider in this chapter the following more general *nonlinear* convex formulation:

Definition 3.1. *Let $F \in \mathscr{F}$. For $\mu, \nu \in \mathscr{P}(\mathcal{X})$, we define:*

$$\mathscr{W}_F(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} F(\pi). \tag{3.1}$$

When $F(\pi) = \int c\, d\pi$, problem (3.1) corresponds to the classical optimal transport problem defined in (KP) and $\mathscr{W}_F = \mathscr{T}_c$.

Lemma 3.1. *The infimum in (3.1) is attained. Moreover, if $F \in \mathscr{F}(\mu,\nu)$, $\mathscr{W}_F(\mu,\nu) < +\infty$.*

Proof. We can apply Weierstrass's theorem since $\Pi(\mu,\nu)$ is compact and F is lsc by definition. For $F \in \mathscr{F}(\mu,\nu)$, there exists $\pi_0 \in \Pi(\mu,\nu)$ such that $F(\pi_0) < +\infty$, so $\mathscr{W}_F(\mu,\nu) \leq F(\pi_0) < +\infty$. □

The main result of this chapter is the following interpretation of problem (3.1) as a ground-cost adversarial OT problem:

Theorem 3.1. *For $\mu, \nu \in \mathscr{P}(\mathcal{X})$ and $F \in \mathscr{F}(\mu,\nu)$, minimizing F over $\Pi(\mu,\nu)$ is equivalent to the following convex problem:*

$$\mathscr{W}_F(\mu,\nu) = \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - F^*(c). \tag{3.2}$$

Proof. Since F is proper, lsc and convex, Fenchel-Moreau theorem ensures that it is equal to its convex biconjugate F^{**}, so:

$$\begin{aligned}\min_{\pi \in \Pi(\mu,\nu)} F(\pi) &= \min_{\pi \in \Pi(\mu,\nu)} F^{**}(\pi) \\ &= \min_{\pi \in \Pi(\mu,\nu)} \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \int c\, d\pi - F^*(c).\end{aligned}$$

Define the objective $l(\pi, c) := \int c\, d\pi - F^*(c)$. Since F^* is lsc as the convex conjugate of F, for any $\pi \in \Pi(\mu,\nu)$, $l(\pi,\cdot)$ is usc. It is also concave as the sum of concave functions. Likewise, for any $c \in \mathcal{C}(\mathcal{X}^2)$, $l(\cdot, c)$ is continuous and convex (in fact linear). Since $\Pi(\mu,\nu)$ and $\mathcal{C}(\mathcal{X}^2)$ are convex, and $\Pi(\mu,\nu)$ is compact, we can use Sion's minimax theorem to swap the min and the sup:

$$\min_{\pi \in \Pi(\mu,\nu)} F(\pi) = \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \min_{\pi \in \Pi(\mu,\nu)} \int c\, d\pi - F^*(c).$$

Finally, $c \mapsto \mathscr{T}_c(\mu,\nu) - F^*(c)$ is concave since F^* is convex and $c \mapsto \mathscr{T}_c(\mu,\nu)$ is concave as the minimum of linear functionals. □

Remark 3.1. *Note that the inequality*

$$\mathscr{W}_F(\mu,\nu) \geq \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - F^*(c)$$

is in fact verified for any $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$ since $F \geq F^{**}$ is always verified.

The supremum in equation (3.2) is not necessarily attained. Under some regularity assumption on F, we show that the supremum is attained and relate the optimal couplings and the optimal ground costs:

Proposition 3.1. *Let $\mu, \nu \in \mathscr{P}(\mathcal{X})$ and $F \in \mathscr{F}(\mu, \nu)$. Suppose that F is Fréchet-differentiable on $\Pi(\mu, \nu)$. Then the supremum in (3.2) is attained at $c_\star = \nabla F(\pi_\star)$ where $\pi_\star$ is any minimizer of (3.1). Conversely, suppose F^* is Fréchet-differentiable everywhere. If $c_\star$ is the unique maximizer in (3.2), then $\pi_\star = \nabla F^*(c_\star)$ is a minimizer of (3.1).*

In section 3.5, we will further characterize $c_\star$ for a certain class of functions $F \in \mathscr{F}$.

Proof. Let $\pi_\star$ be a minimizer of (3.1). Then using the optimality condition for $\sup_{c \in \mathcal{C}(\mathcal{X}^2)} \int c \, d\pi - F^*(c)$, any c such that $\pi_\star \in \partial F^*(c)$ is a best response to $\pi_\star$. But by Fenchel-Young inequality, such c are exactly those in $\partial F(\pi_\star) = \{\nabla F(\pi_\star)\}$. Since $\nabla F(\pi_\star)$ is the unique best response to $\pi_\star$, it is necessarily optimal in (3.2). Conversely, if there is a unique maximizer $c_\star$, then as a result of the above, $c_\star = \nabla F(\pi_\star)$ for some minimizer $\pi_\star$ of the primal. Then $\nabla F^*(c_\star)$ is optimal in the primal. □

One interesting particular case of Theorem 3.1 is when the convex cost $\pi \mapsto F(\pi)$ is a convex regularization of the classical linear optimal transport:

Corollary 3.1. *Let $c_0 \in \mathcal{C}(\mathcal{X}^2)$, $\mu, \nu \in \mathscr{P}(\mathcal{X})$. Let $\varepsilon > 0$ and $R \in \mathscr{F}(\mu, \nu)$. Then:*

$$\begin{aligned} &\min_{\pi \in \Pi(\mu,\nu)} \int c_0 \, d\pi + \varepsilon R(\pi) \\ &= \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu, \nu) - \varepsilon R^* \left(\frac{c - c_0}{\varepsilon} \right). \end{aligned} \tag{3.3}$$

Proof. We apply theorem 3.1 with $F(\pi) = \int c_0 \, d\pi + \varepsilon R(\pi)$, for which we only need to compute the convex conjugate:

$$\begin{aligned} F^*(c) &= \sup_{\pi \in \mathscr{M}(\mathcal{X}^2)} \int c - c_0 \, d\pi - \varepsilon R(\pi) \\ &= \varepsilon \sup_{\pi \in \mathscr{M}(\mathcal{X}^2)} \int \frac{c - c_0}{\varepsilon} \, d\pi - R(\pi) \\ &= \varepsilon R^* \left(\frac{c - c_0}{\varepsilon} \right). \end{aligned}$$

□

Corollary 3.1 shows that the ground cost c_0 in regularized optimal transport acts as a prior on the adversarial ground cost. Indeed, in equation (3.3) the penalization term $\varepsilon R^* \left(\frac{c - c_0}{\varepsilon} \right)$ forces any optimal adversarial ground cost to be "close" to c_0, the closeness being measured in terms of the convex conjugate of the regularization: R^*.

Remark 3.2. *We can also consider the minimization of a proper usc concave function F over $\Pi(\mu,\nu)$. Since $-F \in \mathscr{F}$, by reusing the argument of the proof of Theorem 3.1:*

$$\begin{aligned}
\inf_{\pi\in\Pi(\mu,\nu)} F(\pi) &= \inf_{\pi\in\Pi(\mu,\nu)} -(-F)^{**}(\pi) \\
&= \inf_{\pi\in\Pi(\mu,\nu)} - \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \int c\,d\pi - (-F)^*(c) \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \inf_{c\in\mathcal{C}(\mathcal{X}^2)} \int -c\,d\pi + (-F)^*(c) \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \inf_{c\in\mathcal{C}(\mathcal{X}^2)} \int c\,d\pi + (-F)^*(-c) \\
&= \inf_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) + (-F)^*(-c).
\end{aligned}$$

Minimizing a concave function of the transport plan $\pi \in \Pi(\mu,\nu)$, or equivalently maximizing a convex function of π, amounts to finding a ground cost $c \in \mathcal{C}(\mathcal{X}^2)$ that minimizes the transport cost between μ and ν plus a convex penalization on c. Note that this is not a convex problem since the objective is the sum of a concave and a convex functions. When μ and ν are discrete measures, $\Pi(\mu,\nu)$ is a finite-dimensional compact polytope so one of its extreme points has to be a minimizer of F.

In the ground cost maximization problem, the maximization is carried out on any continuous function c on $\mathcal{X}^2$, and in particular we do not impose that c takes only nonnegative values. In other words, an optimal adversarial ground cost may take negative values, which prevents us from directly interpreting optimal adversarial ground costs as suitable dissimilarity measures over $\mathcal{X}$. In the following subsection, we impose that $c \geq 0$ in the adversarial problem when the space $\mathcal{X}$ is discrete and prove an analogue of Corollary 3.1.

3.3.2 Discrete Separable Case

In this subsection, we will focus on the discrete case where the space $\mathcal{X} = [\![n]\!]$ for some $n \in \mathbb{N}$. A probability measure $\mu \in \mathscr{P}(\mathcal{X})$ is then a histogram of size n that we will represent by a vector $\boldsymbol{\mu} \in \mathbb{R}^n_+$ such that $\sum_{i=1}^n \boldsymbol{\mu}_i = 1$. Cost functions $c \in \mathcal{C}(\mathcal{X}^2)$ and transport plans $\pi \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})$ are now matrices $\mathbf{c}, \boldsymbol{\pi} \in \mathbb{R}^{n\times n}$.

We focus on regularization functions R that are separable, *i.e.* of the form

$$R(\boldsymbol{\pi}) = \sum_{i=1}^n \sum_{j=1}^n R_{ij}(\boldsymbol{\pi}_{ij})$$

for some differentiable convex proper lsc $R_{ij} : \mathbb{R} \to \mathbb{R}$.

In applications, it is natural to constrain the adversarial ground cost $\mathbf{c} \in \mathbb{R}^{n\times n}$ to take nonnegative entries. Adding this constraint on the adversarial

cost corresponds to linearizing "at short range" the regularization R for "small transport values".

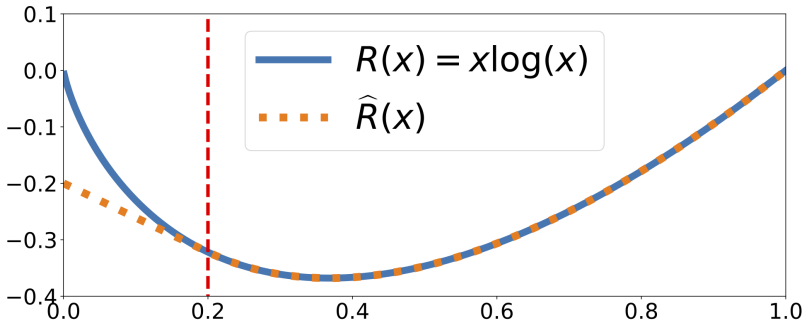


Figure 3.1: The entropy regularization $R(x) = x\log(x)$ and its linearized version $\widehat{R}(x)$ for small transport values.

Proposition 3.2. *Let $\varepsilon > 0$. For $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathscr{P}(\mathcal{X})$, it holds:*

$$\begin{aligned} &\sup_{\mathbf{c} \in \mathbb{R}_+^{n \times n}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \varepsilon \sum_{ij} R_{ij}^* \left(\frac{\mathbf{c}_{ij} - \mathbf{c}_{0ij}}{\varepsilon} \right) \\ &= \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon \sum_{ij} \widehat{R}_{ij}(\boldsymbol{\pi}_{ij}) \end{aligned} \tag{3.4}$$

where $\widehat{R}_{ij} : \mathbb{R} \to \mathbb{R}$ is the continuous convex function defined as

$$\widehat{R}_{ij}(x) := \begin{cases} R_{ij}(x) & \text{if } x \geq R_{ij}^{*\prime}\left(-\frac{\mathbf{c}_{0ij}}{\varepsilon}\right) \\ \frac{-\mathbf{c}_{0ij}}{\varepsilon} x - R_{ij}^*\left(-\frac{\mathbf{c}_{0ij}}{\varepsilon}\right) & \text{otherwise.} \end{cases}$$

Moreover, if R_{ij} is of class C^1, then $\widehat{R}_{ij}$ is also C^1.

Proof. As in the proof of Theorem 3.1, we use Sion's minimax theorem to get

$$\begin{aligned} &\sup_{\mathbf{c} \in \mathbb{R}_+^{n \times n}} \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{c}, \boldsymbol{\pi} \rangle - \varepsilon \sum_{ij} R_{ij}^* \left(\frac{\mathbf{c}_{ij} - \mathbf{c}_{0ij}}{\varepsilon} \right) \\ &= \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \sup_{\mathbf{c} \in \mathbb{R}_+^{n \times n}} \langle \mathbf{c}, \boldsymbol{\pi} \rangle - \varepsilon \sum_{ij} R_{ij}^* \left(\frac{\mathbf{c}_{ij} - \mathbf{c}_{0ij}}{\varepsilon} \right). \end{aligned}$$

Since the optimization in $\mathbf{c} \in \mathbb{R}_+^{n \times n}$ is separable, we only need to consider this optimization coordinate by coordinate, *i.e.* we only need to compute $\sup_{\mathbf{c}_{ij} \in \mathbb{R}_+} \boldsymbol{\pi}_{ij} \mathbf{c}_{ij} - f_{ij}^*(\mathbf{c}_{ij})$ for all $i, j \in [\![n]\!]$, where $f_{ij}^*(\mathbf{c}_{ij}) = \varepsilon R_{ij}^*\left(\frac{\mathbf{c}_{ij} - \mathbf{c}_{0ij}}{\varepsilon}\right)$. Fix $\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $i, j \in [\![n]\!]$, and define $g_{ij} : \mathbb{R} \ni \mathbf{c}_{ij} \mapsto \boldsymbol{\pi}_{ij} \mathbf{c}_{ij} - f_{ij}^*(\mathbf{c}_{ij})$. Suppose that $z_{ij} = f_{ij}'(\boldsymbol{\pi}_{ij}) \geq 0$. Then

$$f_{ij}(\boldsymbol{\pi}_{ij}) = f_{ij}^{**}(\boldsymbol{\pi}_{ij}) = g_{ij}(z_{ij}) = \sup_{\mathbf{c}_{ij} \in \mathbb{R}} g_{ij}(\mathbf{c}_{ij}),$$

and since $z_{ij} \geq 0$, $\sup_{\mathbf{c}_{ij}\in\mathbb{R}_+} g_{ij}(\mathbf{c}_{ij}) = f_{ij}(\boldsymbol{\pi}_{ij})$. This means that $\widehat{R}_{ij}(\boldsymbol{\pi}_{ij}) = R_{ij}(\boldsymbol{\pi}_{ij})$. Suppose now that $z_{ij} = f'_{ij}(\boldsymbol{\pi}_{ij}) < 0$. This means that

$$\sup_{\mathbf{c}_{ij}\in\mathbb{R}_+} g_{ij}(\mathbf{c}_{ij}) < \sup_{\mathbf{c}_{ij}\in\mathbb{R}} g_{ij}(\mathbf{c}_{ij}).$$

Since g_{ij} is concave, this shows that $\sup_{\mathbf{c}_{ij}\in\mathbb{R}_+} g_{ij}(\mathbf{c}_{ij}) = g_{ij}(0) = -f^*_{ij}(0)$, *i.e.* $\widehat{R}_{ij}(\boldsymbol{\pi}_{ij}) = \frac{-\mathbf{c}_{0ij}}{\varepsilon}\boldsymbol{\pi}_{ij} - R^*_{ij}\left(\frac{-\mathbf{c}_{0ij}}{\varepsilon}\right)$. Since R_{ij} is convex, R'_{ij} is increasing with pseudo-inverse ${R^*_{ij}}'$. Furthermore, the optimality condition in the convex conjugate problem gives, for any $\alpha \in \mathbb{R}$:

$$R^*_{ij}(\alpha) = \alpha \times {R^*_{ij}}'(\alpha) - R_{ij} \circ {R^*_{ij}}'(\alpha).$$

So if R_{ij} is of class C^1, taking $\alpha = \frac{-\mathbf{c}_{0ij}}{\varepsilon}$, as x increases to ${R^*_{ij}}'\left(-\frac{\mathbf{c}_{0ij}}{\varepsilon}\right)$:

$$\widehat{R}_{ij}(x) \longrightarrow R_{ij} \circ {R^*_{ij}}'\left(-\frac{\mathbf{c}_{0ij}}{\varepsilon}\right) = \widehat{R}_{ij} \circ {R^*_{ij}}'\left(-\frac{\mathbf{c}_{0ij}}{\varepsilon}\right),$$

meaning that $\widehat{R}_{ij}$ is of class C^1. □

3.4 Examples

3.4.1 Ground-Cost Adversarial Interpretation of Classical OT Regularizations

As presented in the introduction, several convex regularizations R have been proposed in the literature. We give the ground cost adversarial counterpart for some of them: two examples in the continuous setting, and four p-norm based regularizations in the discrete case.

Example 3.1 (Entropic Regularization)**.** *Let $\mu, \nu \in \mathscr{P}(\mathcal{X})$. For $\pi \in \Pi(\mu,\nu)$, we define its relative entropy as* $\mathrm{KL}(\pi\|\mu\otimes\nu) = \int \log \frac{d\pi}{d\mu\otimes\nu} d\pi$*. Then for $c_0 \in \mathcal{C}(\mathcal{X}^2)$ and $\varepsilon > 0$, it holds:*

$$\begin{aligned}&\min_{\pi\in\Pi(\mu,\nu)} \int c_0\, d\pi + \varepsilon\, \mathrm{KL}(\pi\|\mu\otimes\nu)\\ &= \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \varepsilon \int \exp\left(\frac{c-c_0}{\varepsilon}\right) d\mu\otimes\nu + \varepsilon.\end{aligned}$$

Proof. For $\pi \in \mathscr{M}(\mathcal{X}^2)$, let

$$R(\pi) = \begin{cases} \int \log \frac{d\pi}{d\mu\otimes\nu} d\pi - \int d\pi + 1 & \text{if } \pi \ll \mu\otimes\nu \\ +\infty & \text{otherwise.}\end{cases}$$

R is convex, and using proposition 7 in [Feydy et al., 2019],

$$R^*(c) = \int e^c - 1\, d\mu\otimes\nu.$$

Applying corollary 3.1 concludes the proof. □

Another case of interest is the so-called Subspace Robust Wasserstein distance that we defined in chapter 2. Here, the set of adversarial metrics is parameterized by a finite-dimensional parameter Ω, which allows to recover an adversarial metric defined on the whole space even when the measures are finitely supported.

Example 3.2 (Subspace Robust Wasserstein)**.** *Let $d \in \mathbb{N}$, $k \in [\![d]\!]$ and $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$. For $\pi \in \Pi(\mu,\nu)$, define $V_\pi = \int (x-y)(x-y)^\top d\pi(x,y)$ and $\lambda_1(V_\pi) \geq \ldots \geq \lambda_d(V_\pi)$ its ordered eigenvalues.*

Then $F : \pi \mapsto \sum_{l=1}^k \lambda_l(V_\pi)$ is convex, and

$$\mathcal{S}_k(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \sum_{l=1}^{k} \lambda_l(V_\pi) = \max_{\substack{0 \preceq \Omega \preceq I \\ \mathrm{trace}(\Omega)=k}} \mathscr{T}_{d^2_\Omega}(\mu,\nu)$$

where $d^2_\Omega(x,y) = (x-y)^\top \Omega (x-y)$ is the squared Mahalanobis distance.

Proof. This is a direct consequence of Theorem 2.1 and Lemma 2.2. Note that in this case, $\mathcal{X} = \mathbb{R}^d$ is not compact. This is not a problem since $F^* \equiv +\infty$ outside a compact set, *i.e.* the set on metrics on which the maximization takes place is compact. Indeed, one can show that:

$$F^*(c) = \iota(\exists 0 \preceq \Omega \preceq I \text{ with } \mathrm{trace}(\Omega) = k \text{ s.t. } c = d^2_\Omega). \qquad \square$$

Let us now consider p-norm based examples, which will subsume quadratically-regularized ($p=2$) OT studied in [Essid and Solomon, 2017, Lorenz et al., 2019], capacity-constrained ($p=+\infty$) OT proposed by Korman and McCann [2015] and Tsallis regularized ($p<0$) OT introduced by Muzellec et al. [2017].

For a matrix $\mathbf{w} \in \mathbb{R}_+^{n\times n}$ with $\sum_{ij} \mathbf{w}_{ij} = n^2$ and $\boldsymbol{\pi} \in \mathbb{R}^{n\times n}$, we denote by $\|\boldsymbol{\pi}\|^p_{\mathbf{w},p} = \sum_{ij} \mathbf{w}_{ij} |\boldsymbol{\pi}_{ij}|^p$ the $\mathbf{w}$-weighted (powered) p-norm of $\boldsymbol{\pi}$. We also write $1/\mathbf{w}$ for the matrix defined by $(1/\mathbf{w})_{ij} = 1/\mathbf{w}_{ij}$. In the following, except otherwise mentioned, we take $p, q \in [1,+\infty]$ such that $1/p + 1/q = 1$, $\mathbf{c}_0 \in \mathbb{R}^{n\times n}$, $\varepsilon > 0$.

Example 3.3 ($\|\cdot\|^p_{\mathbf{w},p}$ Regularization)**.**

$$\begin{aligned}\min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon \frac{1}{p} \|\boldsymbol{\pi}\|^p_{\mathbf{w},p} \\ = \sup_{\mathbf{c} \in \mathbb{R}^{n\times n}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon \frac{1}{q} \left\| \frac{\mathbf{c}-\mathbf{c}_0}{\varepsilon} \right\|^q_{1/\mathbf{w}^{q-1},q}.\end{aligned}$$

In particular when $p=2$ and $\mathbf{w}=1$, this corresponds to quadratically-regularized OT studied in [Essid and Solomon, 2017, Lorenz et al., 2019].

Proof. We denote by $\text{sign}(x)$ the set $\{+1\}$ if $x > 0$, $\{-1\}$ if $x < 0$ and $[-1, 1]$ if $x = 0$. We apply Corollary 3.1 with $R : \mathbb{R}^{n\times n} \to \mathbb{R}$ defined as $R(\boldsymbol{\pi}) = \frac{1}{p}\|\boldsymbol{\pi}\|^p_{\mathbf{w},p}$, for which we need to compute its convex conjugate:

$$R^*(\mathbf{c}) = \sup_{\boldsymbol{\pi}\in\mathbb{R}^{n\times n}} \langle \boldsymbol{\pi}, \mathbf{c}\rangle - \frac{1}{p}\sum_{ij} \mathbf{w}_{ij}|\boldsymbol{\pi}_{ij}|^p.$$

Subdifferentiating with respect to $\boldsymbol{\pi}_{ij}$:

$$\begin{aligned}\mathbf{c}_{ij} &\in \frac{1}{p}\mathbf{w}_{ij}\frac{\partial}{\partial \boldsymbol{\pi}_{ij}}|\boldsymbol{\pi}_{ij}|^p \\ &= \mathbf{w}_{ij}\,\text{sign}(\boldsymbol{\pi}_{ij})|\boldsymbol{\pi}_{ij}|^{p-1}\end{aligned}$$

This implies that $\text{sign}(\boldsymbol{\pi}_{ij}) = \text{sign}(\mathbf{c}_{ij})$, so:

$$\boldsymbol{\pi}_{ij} = \text{sign}(\mathbf{c}_{ij})\left|\frac{\mathbf{c}_{ij}}{\mathbf{w}_{ij}}\right|^{q-1}.$$

Finally,

$$\begin{aligned}R^*(\mathbf{c}) &= \sum_{ij} \mathbf{c}_{ij}\,\text{sign}(\mathbf{c}_{ij})\left|\frac{\mathbf{c}_{ij}}{\mathbf{w}_{ij}}\right|^{q-1} - \frac{1}{p}\mathbf{w}_{ij}\left|\frac{\mathbf{c}_{ij}}{\mathbf{w}_{ij}}\right|^{q} \\ &= \frac{1}{q}\sum_{ij}\frac{1}{\mathbf{w}_{ij}^{q-1}}|\mathbf{c}_{ij}|^q \\ &= \frac{1}{q}\|\mathbf{c}\|^q_{1/\mathbf{w}^{q-1},q}.\end{aligned}$$

□

Example 3.4 ($\|\cdot\|_{\mathbf{w},p}$ Penalization)**.**

$$\min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi}\rangle + \varepsilon\|\boldsymbol{\pi}\|_{\mathbf{w},p} = \sup_{\substack{\mathbf{c}\in\mathbb{R}^{n\times n} \\ \|\mathbf{c}-\mathbf{c}_0\|_{1/\mathbf{w},q}\leq\varepsilon}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}).$$

Proof. We apply Corollary 3.1 with $R : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ defined as $R(\boldsymbol{\pi}) = \|\boldsymbol{\pi}\|_{\mathbf{w},p}$, for which we need to compute its convex conjugate. We know that the dual of $\|\cdot\|_p$ is $\iota(\|\cdot\|_q \leq 1)$, and using classical results about convex conjugates, $\|\cdot\|^*_{\mathbf{w},p} = \iota(\|\cdot\|_{1/\mathbf{w},q} \leq 1)$. □

Example 3.5 ($\|\cdot\|_{\mathbf{w},p}$ Regularization)**.**

$$\min_{\substack{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu}) \\ \|\boldsymbol{\pi}\|_{\mathbf{w},p}\leq\varepsilon}} \langle \mathbf{c}_0, \boldsymbol{\pi}\rangle = \sup_{\mathbf{c}\in\mathbb{R}^{n\times n}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon\|\mathbf{c}-\mathbf{c}_0\|_{1/\mathbf{w},q}.$$

In particular when $p = +\infty$ and $\mathbf{w} = 1$, this coincides with capacity-constrained OT proposed by Korman and McCann [2015].

Proof. We apply Corollary 3.1 with $R : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ defined as $R(\boldsymbol{\pi}) = \iota(\|\boldsymbol{\pi}\|_{\mathbf{w},p} \leq 1)$, for which we need to compute its convex conjugate. We know that the dual of $\iota(\|\cdot\|_p \leq 1)$ is $\|\cdot\|_q$, and using classical results about convex conjugates, $\iota(\|\cdot\|_{\mathbf{w},p} \leq 1)^* = \|\cdot\|_{1/\mathbf{w},q}$. □

Example 3.6 (Tsallis Regularization)**.** *For $q \in (0,1)$, the Tsallis regularized OT problem [Muzellec et al., 2017]*

$$\min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi}\rangle - \varepsilon \frac{1}{1-q} \sum_{ij} \left(\boldsymbol{\pi}_{ij}^q - \boldsymbol{\pi}_{ij}\right)$$

is equivalent to

$$\sup_{\substack{\mathbf{c}\in\mathbb{R}^{n\times n}\\ \mathbf{c}\leq\mathbf{c}_0}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon^{\frac{1}{1-q}}(-p)^{-p} \left\| \frac{1}{\mathbf{c}_0 - \mathbf{c}} \right\|_{-p}^{-p} + \frac{\varepsilon}{1-q}$$

where $p < 0$ is such that $1/p + 1/q = 1$.

Proof. Since $\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})$, $\sum_{ij} \boldsymbol{\pi}_{ij} = 1$ so we can drop it for now and only consider the term $R(\boldsymbol{\pi}) = \frac{1}{q-1}\|\boldsymbol{\pi}\|_q^q$ which is separable in the coordinates of $\boldsymbol{\pi}$:

$$R(\boldsymbol{\pi}) = \sum_{ij} f(\boldsymbol{\pi}_{ij})$$

where we have defined the convex function

$$f(x) = \begin{cases} \frac{1}{q-1}x^q & \text{if } x \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

We compute its convex conjugate:

$$f^*(y) = \sup_{x\geq 0} \left\{ xy - \frac{1}{q-1}x^q \right\} = \begin{cases} \left(\frac{y}{p}\right)^p & \text{if } y \leq 0 \\ +\infty & \text{if } y > 0 \end{cases}$$

where $p = \frac{q}{q-1} \leq 0$ is such that $1/p + 1/q = 1$. Then $R^*(\mathbf{c}) = +\infty$ if $\mathbf{c}$ has a positive entry, and over $\mathbb{R}_-^{n\times n}$:

$$\begin{aligned} R^*(\mathbf{c}) &= \sum_{ij} f^*(\mathbf{c}_{ij}) = \sum_{ij} \left(\frac{\mathbf{c}_{ij}}{p}\right)^p \\ &= \sum_{ij} \left(\frac{-\mathbf{c}_{ij}}{-p}\right)^p \\ &= (-p)^{-p} \sum_{ij} \left(\frac{1}{-\mathbf{c}_{ij}}\right)^{-p}. \end{aligned}$$

Adding the term $\frac{\varepsilon}{1-q}$ we left aside to the result of Corollary 3.1, we find that Tsallis regularized OT corresponds to:

$$\begin{aligned}
&\sup_{\mathbf{c}\in\mathbb{R}^{n\times n}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon R^*\left(\frac{\mathbf{c}-\mathbf{c}_0}{\varepsilon}\right) + \frac{\varepsilon}{1-q} \\
&= \sup_{\substack{\mathbf{c}\in\mathbb{R}^{n\times n}\\ \mathbf{c}\leq\mathbf{c}_0}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon(-p)^{-p}\sum_{ij}\left[\frac{\varepsilon}{\mathbf{c}_{0ij}-\mathbf{c}_{ij}}\right]^{-p} + \frac{\varepsilon}{1-q} \\
&= \sup_{\substack{\mathbf{c}\in\mathbb{R}^{n\times n}\\ \mathbf{c}\leq\mathbf{c}_0}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon^{\frac{1}{1-q}}(-p)^{-p}\sum_{ij}\left[\frac{1}{\mathbf{c}_{0ij}-\mathbf{c}_{ij}}\right]^{-p} + \frac{\varepsilon}{1-q} \\
&= \sup_{\substack{\mathbf{c}\in\mathbb{R}^{n\times n}\\ \mathbf{c}\leq\mathbf{c}_0}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon^{\frac{1}{1-q}}(-p)^{-p}\left\|\frac{1}{\mathbf{c}_0-\mathbf{c}}\right\|_{-p}^{-p} + \frac{\varepsilon}{1-q}.
\end{aligned}$$

□

3.4.2 A Link With the Matching Literature in Economics

Maximizing the OT problem with respect to the ground cost has been proposed in the matching literature in economics as a way to recover a ground cost when only a matching is observed, see *e.g.* [Dupuy and Galichon, 2014, Galichon and Salanié, 2015, Dupuy et al., 2016]. In this subsection, we reinterpret their methods by showing that they are equivalent to some regularized OT problems. In other words, instead of interpreting a regularization problem as a robust OT problem as in subsection 3.4.1, we go the other way around and show that this practical OT maximization problem corresponds to a regularized OT problem.

Practitioners observe two probability measures $\mu,\nu\in\mathscr{P}(\mathcal{X})$ (*e.g.* features from a group of men and a group of women) and a matching $\pi_0\in\Pi(\mu,\nu)$ (*e.g.* dating or marriage data). Under the assumption that the matching is optimal for some criteria, we can determine these criteria by finding a ground cost $c_\star\in\mathcal{C}(\mathcal{X}^2)$ such that the matching π_0 is an optimal transport plan for the cost $c_\star$. Then $c_\star(x,y)$ can be interpreted as the unwillingness for two people with characteristics x and y to be matched.

As shown in Theorem 3 in [Galichon and Salanié, 2015],

$$\sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \int c\,d\pi_0 = \iota\left(\pi_0\in\Pi(\mu,\nu)\right) \tag{3.5}$$

and if $\pi_0\in\Pi(\mu,\nu)$, the supremum is attained at any $c_\star\in\mathcal{C}(\mathcal{X}^2)$ such that π_0 is an optimal transport plan for the cost $c_\star$. Indeed, the first order condition for the maximization problem and the envelope theorem give the result.

In practice, economists are more interested in discovering which features explain the most the observed matching π_0. To this end, they choose a parametric model for the cost c, for example a Mahalanobis model $c\in$

$\left\{d_\Omega^2 : (x,y) \mapsto (x-y)^\top \Omega (x-y), \Omega \succeq 0, \|\Omega\| \leq 1\right\}$. More generally, we can rewrite problem (3.5) as

$$\sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \int c \, d\pi_0 - R^*(c) \tag{3.6}$$

where $R \in \mathscr{F}$ is a lsc convex functional, *e.g.* $R^*(c) = \iota\left(\exists \Omega \succeq 0, \|\Omega\| \leq 1, c = d_\Omega^2\right)$ for the Mahalanobis model.

Using Theorem 3.1, we can then reinterpret problem (3.6):

$$\begin{aligned}&\sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \int c \, d\pi_0 - R^*(c)\\ &= \min_{\pi \in \Pi(\mu,\nu)} R(\pi - \pi_0)\end{aligned}$$

where we have used the fact that $R^{**} = R$. Solving equation (3.6) amounts to finding a matching $\pi \in \Pi(\mu,\nu)$ that is close to the observed matching π_0, as measured by R.

3.5 Characterization of the Adversarial Cost and Duality

Theorem 3.1 shows that regularizing OT is equivalent to maximizing unregularized OT with respect to the ground cost. This gives access to a robustly computed ground-cost $c_\star$. In this section, we first prove a duality theorem for problem (3.1) that we use to further characterize $c_\star$. We will first need a technical assumption on F:

Definition 3.2. *Let $F \in \mathscr{F}$. We will say that F is* separably $*$-increasing *if for any $\phi, \psi \in \mathcal{C}(\mathcal{X})$ and any $c \in \mathcal{C}(\mathcal{X}^2)$:*

$$\phi \oplus \psi \leq c \Rightarrow F^*(\phi \oplus \psi) \leq F^*(c). \tag{3.7}$$

In particular if F^* is increasing, F is separably $*$-increasing.

This definition, albeit not always verified *e.g.* in the discrete separable case of Proposition 3.2 and in the SRW case of Example 3.2, is indeed verified in various cases of interest, *e.g.* for the entropic or $\|\cdot\|_{\mathbf{w},p}^p$ regularizations:

Example 3.7. *For $\mu, \nu \in \mathscr{P}(\mathcal{X})$, $c_0 \in \mathcal{C}(\mathcal{X}^2)$ and $\varepsilon > 0$, the entropy-regularized OT function*

$$F : \pi \mapsto \int c_0 \, d\pi + \varepsilon \, \mathrm{KL}(\pi \| \mu \otimes \nu)$$

is separably $*$-increasing.

Proof. As in the proof of example 3.1,

$$F^*(c) = \varepsilon \int \exp\left(\frac{c - c_0}{\varepsilon}\right) - 1 \, d\mu \otimes \nu$$

which verifies condition (3.7) as an increasing functional. □

Example 3.8. *In the discrete setting $\mathcal{X} = [\![n]\!]$, let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathscr{P}(\mathcal{X})$, $\mathbf{c}_0 \in \mathbb{R}^{n \times n}$, $\mathbf{w} \in \mathbb{R}_+^{n \times n}$ summing to n^2. Take $p > 1$ and $\varepsilon > 0$. With $\varphi_p(x) = x^p$ if $x \geq 0$ and $\varphi_p(x) = +\infty$ if $x < 0$, the $\|\cdot\|_{\mathbf{w},p}^p$-regularized OT function*

$$F : \boldsymbol{\pi} \mapsto \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon \sum_{ij} \mathbf{w}_{ij} \varphi_p(\boldsymbol{\pi}_{ij})$$

is separably $*$-increasing.

Proof. Note that minimizing F over $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \subset \mathbb{R}_+^{n \times n}$ is equivalent to minimizing $\widetilde{F} : \boldsymbol{\pi} \mapsto \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon \sum_{ij} \mathbf{w}_{ij} |\boldsymbol{\pi}_{ij}|^p$. One can show that, with $q > 1$ such that $1/p + 1/q = 1$ and $(x)_+ := \max\{0, x\}$:

$$F^*(\mathbf{c}) = \varepsilon \frac{1}{q} \left\| \frac{(\mathbf{c} - \mathbf{c}_0)_+}{\varepsilon} \right\|_{1/\mathbf{w}^{q-1},q}^q$$

which clearly verifies condition (3.7). □

When F is separably $*$-increasing, we can easily prove a duality theorem for problem (3.1):

Theorem 3.2 ($\mathscr{W}_F$ duality)**.** *Let $\mu, \nu \in \mathscr{P}(\mathcal{X})$ and $F \in \mathscr{F}(\mu, \nu)$ a separably $*$-increasing function. Then:*

$$\mathscr{W}_F(\mu, \nu) = \max_{\phi, \psi \in \mathcal{C}(\mathcal{X})} \int \phi \, d\mu + \int \psi \, d\nu - F^*(\phi \oplus \psi). \tag{3.8}$$

Proof. Using Theorem 3.1 and Kantorovich duality (Dual-KP):

$$\begin{aligned}
\mathscr{W}_F(\mu, \nu) &= \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu, \nu) - F^*(c) \\
&= \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \max_{\substack{\phi, \psi \in \mathcal{C}(\mathcal{X}) \\ \phi \oplus \psi \leq c}} \int \phi \, d\mu + \int \psi \, d\nu - F^*(c) \\
&= \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \max_{\phi, \psi \in \mathcal{C}(\mathcal{X})} \int \phi \, d\mu + \int \psi \, d\nu - F^*(c) - \iota(\phi \oplus \psi \leq c) \\
&= \max_{\phi, \psi \in \mathcal{C}(\mathcal{X})} \int \phi \, d\mu + \int \psi \, d\nu + \sup_{c \in \mathcal{C}(\mathcal{X}^2)} -F^*(c) - \iota(\phi \oplus \psi \leq c) \\
&= \max_{\phi, \psi \in \mathcal{C}(\mathcal{X})} \int \phi \, d\mu + \int \psi \, d\nu - \inf_{\substack{c \in \mathcal{C}(\mathcal{X}^2) \\ \phi \oplus \psi \leq c}} F^*(c).
\end{aligned}$$

Since F is separably $*$-increasing, for any $\phi, \psi \in \mathcal{C}(\mathcal{X})$,

$$\inf_{\substack{c \in \mathcal{C}(\mathcal{X}^2) \\ \phi \oplus \psi \leq c}} F^*(c) = F^*(\phi \oplus \psi),$$

which shows the desired duality result. □

Theorem 3.2 subsumes the already known duality results for entropy-regularized OT and quadratically-regularized OT. It also enables us to characterize of the optimal adversarial ground cost when the convex objective $F \in \mathscr{F}$ is separably $*$-increasing:

Corollary 3.2. *If $\phi_\star, \psi_\star$ are optimal solutions in (3.8), the cost $\phi_\star \oplus \psi_\star \in \mathcal{C}(\mathcal{X}^2)$ is an optimal adversarial cost in (3.2).*

Proof. For $\phi, \psi \in \mathcal{C}(\mathcal{X})$, note that

$$\mathscr{T}_{\phi \oplus \psi}(\mu, \nu) = \int \phi \, d\mu + \int \psi \, d\nu.$$

Then using $\mathscr{W}_F$ duality:

$$\begin{aligned}
\mathscr{W}_F(\mu, \nu) &= \max_{\phi, \psi \in \mathcal{C}(\mathcal{X})} \int \phi \, d\mu + \int \psi \, d\nu - F^*(\phi \oplus \psi) \\
&= \max_{\phi, \psi \in \mathcal{C}(\mathcal{X})} \mathscr{T}_{\phi \oplus \psi}(\mu, \nu) - F^*(\phi \oplus \psi) \\
&\leq \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu, \nu) - F^*(c) \\
&= \mathscr{W}_F(\mu, \nu)
\end{aligned}$$

where we have used Theorem 3.1 in the last line. This shows that the inequality is in fact an equality, so if $\phi_\star, \psi_\star$ are optimal dual potentials in (3.8), $\phi_\star \oplus \psi_\star$ is an optimal adversarial cost in (3.2). □

Corollary 3.2 is quite striking. Indeed, in the regularized formulation of Corollary 3.1, any optimal ground cost $c_\star$ in equation (3.3) should be close (in R^* sense) to the prior cost c_0 because of the penalization term $\varepsilon R^*\left(\frac{c - c_0}{\varepsilon}\right)$. But under the assumption that F is separably $*$-increasing, we have just shown that regardless of c_0, there exists an optimal adversarial ground cost that is separable.

3.6 Adversarial Ground-Cost for Several Measures

For two measures $\mu, \nu \in \mathscr{P}(\mathcal{X})$ and a separably $*$-increasing function $F \in \mathscr{F}(\mu, \nu)$, corollary 3.2 shows that there exists an optimal adversarial ground cost $c_\star$ that is separable. This separability, which is verified *e.g.* in the entropic or quadratic case, means that the OT problem for $c_\star$ is degenerate in the sense

that any transport plan is optimal for the cost $c_\star$. From a metric learning point of view, $c_\star$ is not a suitable dissimilarity measure on $\mathcal{X}$. But why limit ourselves to two measures? If we observe $N \in \mathbb{N}$ measures $\mu_1, \dots, \mu_N \in \mathscr{P}(\mathcal{X})$, we could look for a ground cost $c \in \mathcal{C}(\mathcal{X}^2)$ that is adversarial to all the pairs:

$$\sup_{c \in \mathcal{C}(\mathcal{X}^2)} \sum_{i \neq j} \mathscr{T}_c(\mu_i, \mu_j) - F^*(c)$$

for some convex regularization $F^* : \mathcal{C}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$. We will specifically focus on the case where we observe a sequence of measures $\mu_{1:T} \stackrel{\text{def}}{=} \mu_1, \dots, \mu_T \in \mathscr{P}(\mathcal{X})$, $T \geq 2$. When we observe such time-dependent data, we can look for a sequence of adversarial costs $c_{1:T-1} \stackrel{\text{def}}{=} c_1, \dots, c_{T-1} \in \mathcal{C}(\mathcal{X}^2)$ which is globally adversarial:

Definition 3.3. *For $D : \mathcal{C}(\mathcal{X}^2) \times \mathcal{C}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$ and $F_t \in \mathscr{F}(\mu_t, \mu_{t+1})$, $t \in [\![T-1]\!]$, we define:*

$$\mathcal{W}_{D,F}(\mu_{1:T}) \stackrel{\text{def}}{=} \sup_{c_{1:T-1}} \sum_{t=1}^{T-1} \mathscr{T}_{c_t}(\mu_t, \mu_{t+1}) - D(c_t, c_{t+1}) - F_t^*(c_t) \tag{3.9}$$

with the convention $D(c_{T-1}, c_T) = 0$.

In problem (3.9), D acts as a time-regularization by forcing the adversarial sequence of ground-costs to vary "continuously" with time.

Taking inspiration from the Subspace Robust Wasserstein (SRW) distance, we propose as a particular case of definition 3.3 a generalization of SRW to the case of a sequence of measures $\mu_1, \dots, \mu_T$, $T \geq 2$:

Definition 3.4. *Let $d \in \mathbb{N}$ and $k \in [\![d]\!]$. Define*

$$\mathcal{R}_k \stackrel{\text{def}}{=} \left\{ \Omega \in \mathbb{R}^{d \times d}, 0 \preceq \Omega \preceq I, \operatorname{trace}(\Omega) = k \right\}.$$

We define the sequential SRW *between $\mu_1, \dots, \mu_T \in \mathscr{P}(\mathbb{R}^d)$ as:*

$$\mathcal{TS}_{k,\eta}(\mu_{1:T}) \stackrel{\text{def}}{=} \sup_{\Omega_1, \dots, \Omega_{T-1} \in \mathcal{R}_k} \sum_{t=1}^{T-1} \mathscr{T}_{d^2_{\Omega_t}}(\mu_t, \mu_{t+1}) - \eta \mathfrak{B}^2(\Omega_t, \Omega_{t+1}) \tag{3.10}$$

where $\mathfrak{B}^2(A, B) = \operatorname{trace}(A + B - 2(A^{\frac{1}{2}} B A^{\frac{1}{2}})^{\frac{1}{2}})$ is the squared Bures metric [Bures, 1969, Bhatia et al., 2018] on the SDP cone.

Note that problem (3.10) is convex. If $T = 2$, the sequential SRW is equal to the usual SRW distance: $\mathcal{TS}_{k,\eta}(\mu_1, \mu_2) = \mathcal{S}_k(\mu_1, \mu_2)$.

3.7 Algorithms

From now on, we only consider the discrete case $\mathcal{X} = [\![n]\!]$.

3.7.1 Projected (Sub)gradient Ascent Solves Nonnegative Adversarial Cost OT

In the setting of subsection 3.3.2, we propose to run a projected subgradient ascent on the ground cost $\mathbf{c} \in \mathbb{R}_+^{n\times n}$ to solve problem (3.4). Note that in this case, $\widehat{F}(\boldsymbol{\pi}) := \langle \mathbf{c}_0, \boldsymbol{\pi}\rangle + \varepsilon \sum_{ij} \widehat{R}^*_{ij}\left(\frac{\mathbf{c}_{ij}-\mathbf{c}_{0ij}}{\varepsilon}\right)$ is **not** separably $*$-increasing, so we can hope that the optimal adversarial ground cost will not be separable.

At each iteration of the ascent, we need to compute a subgradient of $g : \mathbf{c} \mapsto \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \varepsilon R^*\left(\frac{\mathbf{c}-\mathbf{c}_0}{\varepsilon}\right)$ given by Danskin's theorem:

$$\partial g(\mathbf{c}) = \operatorname{conv}\left\{ \boldsymbol{\pi}_\star - \nabla R^*\left(\frac{\mathbf{c}-\mathbf{c}_0}{\varepsilon}\right) \,\middle|\, \boldsymbol{\pi}_\star \in \operatorname*{arg\,min}_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{c}, \boldsymbol{\pi}\rangle \right\}.$$

Although projected subgradient ascent does converge, having access to gradients instead of subgradients, hence regularity, helps the convergence. We therefore propose to replace $\mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ by its entropy-regularized version

$$\mathscr{S}^\eta_{\mathbf{c}}(\mu, \nu) = \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{c}, \boldsymbol{\pi}\rangle + \eta \sum_{ij} \boldsymbol{\pi}_{ij}(\log \boldsymbol{\pi}_{ij} - 1)$$

in the definition of the obective g. Then g is differentiable, because there exists a unique solution $\boldsymbol{\pi}_\star$ in the entropic case (hence $\partial g(\mathbf{c})$ is a singleton). This will also speed up the computations of the gradient at each iteration using Sinkhorn's algorithm. We can interpret this addition of a small entropy term in the adversarial cost formulation as a further regularization of the primal:

Corollary 3.3. *Using the same notations as in Theorem 3.1, for $\eta \geq 0$:*

$$\sup_{\mathbf{c}\in\mathbb{R}^{n\times n}} \mathscr{S}^\eta_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\nu}) - F^*(\mathbf{c}) = \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} F(\boldsymbol{\pi}) + \eta \sum_{ij} \boldsymbol{\pi}_{ij}(\log \boldsymbol{\pi}_{ij} - 1).$$

Proof. Let $R(\boldsymbol{\pi}) := \sum_{ij} \boldsymbol{\pi}_{ij}(\log \boldsymbol{\pi}_{ij} - 1)$. Then:

$$\begin{aligned}
\sup_{\mathbf{c}\in\mathbb{R}^{n\times n}} \mathscr{S}^\eta_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\nu}) - F^*(\mathbf{c}) &= \sup_{\mathbf{c}\in\mathbb{R}^{n\times n}} \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \boldsymbol{\pi}, \mathbf{c}\rangle + \eta R(\boldsymbol{\pi}) - F^*(\mathbf{c}) \\
&= \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \eta R(\boldsymbol{\pi}) + \sup_{\mathbf{c}\in\mathbb{R}^{n\times n}} \langle \boldsymbol{\pi}, \mathbf{c}\rangle - F^*(\mathbf{c}) \\
&= \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \eta R(\boldsymbol{\pi}) + F(\boldsymbol{\pi})
\end{aligned}$$

where we have used Sion's minimax theorem as in the proof of Theorem 3.1 to swap the min and the sup, and used as well the fact that $F = F^{**}$ given by Fenchel-Moreau theorem. □

Algorithm 3 Projected *(sub)*Gradient Ascent for Nonnegative Adversarial Cost

Input: Histograms $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^n$, learning rate lr
Initialize $\mathbf{c} \in \mathbb{R}_+^{n \times n}$
for $i = 0$ **to** MAXITER **do**
  $\boldsymbol{\pi}_\star \leftarrow \mathrm{OT}(\boldsymbol{\mu}, \boldsymbol{\nu}, \mathrm{cost} = \mathbf{c})$
  $\mathbf{c} \leftarrow \mathrm{Proj}_{\mathbb{R}_+^{n \times n}} \left[\mathbf{c} + \mathrm{lr}\boldsymbol{\pi}_\star - \mathrm{lr}\nabla R^* \left(\frac{\mathbf{c}-\mathbf{c}_0}{\varepsilon}\right)\right]$
end for

3.7.2 Sinkhorn-like Algorithm for $*$-increasing $F \in \mathscr{F}$

If the function $F \in \mathscr{F}$ is separably $*$-increasing, we can directly write the optimality conditions for the concave dual problem (3.8):

$$\boldsymbol{\mu} = \nabla F^*(\boldsymbol{\phi}_\star \oplus \boldsymbol{\psi}_\star)\mathbf{1} \tag{3.11}$$

$$\boldsymbol{\nu} = \nabla F^*(\boldsymbol{\phi}_\star \oplus \boldsymbol{\psi}_\star)^\top \mathbf{1} \tag{3.12}$$

where $\mathbf{1}$ is the vector of all ones. We can then alternate between fixing $\boldsymbol{\psi}$ and solving for $\boldsymbol{\phi}$ in (3.11) and fixing $\boldsymbol{\phi}$ and solving for $\boldsymbol{\psi}$ in (3.12). In the case of entropy-regularized OT, this is equivalent to Sinkhorn's algorithm. In quadratically-regularized OT, this is equivalent to the alternate minimization proposed by Blondel et al. [2018]. We give the detailed derivation of these facts:

Proof. In the case of entropic OT,

$$F(\boldsymbol{\pi}) = \langle \boldsymbol{\pi}, \mathbf{c}_0 \rangle + \varepsilon \sum_{ij} \boldsymbol{\pi}_{ij} \left[\log \boldsymbol{\pi}_{ij} - 1\right],$$

so

$$F^*(\mathbf{c}) = \varepsilon \sum_{ij} \exp\left(\frac{\mathbf{c}_{ij} - \mathbf{c}_{0ij}}{\varepsilon}\right)$$

and

$$\nabla F^*(\mathbf{c}) = \left[\exp\left(\frac{\mathbf{c}_{ij} - \mathbf{c}_{0ij}}{\varepsilon}\right)\right]_{ij}.$$

Then the system of equations (3.11) (3.12) is:

$$\forall i,\ \boldsymbol{\mu}_i = \sum_j \exp\left(\frac{\boldsymbol{\phi}_{\star i} + \boldsymbol{\psi}_{\star j} - \mathbf{c}_{0ij}}{\varepsilon}\right) = \exp(\boldsymbol{\phi}_{\star i}/\varepsilon)\left[K \exp(\boldsymbol{\psi}_\star/\varepsilon)\right]_i$$

$$\forall j,\ \boldsymbol{\nu}_j = \sum_i \exp\left(\frac{\boldsymbol{\phi}_{\star i} + \boldsymbol{\psi}_{\star j} - \mathbf{c}_{0ij}}{\varepsilon}\right) = \exp(\boldsymbol{\psi}_{\star j}/\varepsilon)\left[K^\top \exp(\boldsymbol{\phi}_\star/\varepsilon)\right]_j$$

where $K = \exp(-\mathbf{c}_0/\varepsilon) \in \mathbb{R}^{n \times n}$ and exp is taken elementwise. Then solving alternatively for $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ is exactly Sinkhorn's algorithm.

In the case of quadratic OT, using the notations and results from example 3.8:

$$F(\boldsymbol{\pi}) = \langle \boldsymbol{\pi}, \mathbf{c}_0 \rangle + \varepsilon \varphi_2(\boldsymbol{\pi}_{ij}),$$

and

$$F^*(\mathbf{c}) = \frac{1}{2\varepsilon} \sum_{ij} \left[\left(\mathbf{c}_{ij} - \mathbf{c}_{0ij} \right)_+ \right]^2 .$$

Then:

$$\nabla F^*(\mathbf{c}) = \frac{1}{\varepsilon} \left(\mathbf{c} - \mathbf{c}_0 \right)_+ .$$

The system of equations (3.11) (3.12) is:

$$\forall i,\ \varepsilon \boldsymbol{\mu}_i = \sum_j \left(\boldsymbol{\phi}_{\star i} + \boldsymbol{\psi}_{\star j} - \mathbf{c}_{0ij} \right)_+$$
$$\forall j,\ \varepsilon \boldsymbol{\nu}_j = \sum_i \left(\boldsymbol{\phi}_{\star i} + \boldsymbol{\psi}_{\star j} - \mathbf{c}_{0ij} \right)_+$$

which is what Blondel et al. [2018] solve in their appendix B. □

3.7.3 Coordinate Ascent for Sequential SRW

Problem (3.10) is a globally convex problem of $\Omega_1, \dots, \Omega_{T-1}$. We propose to run a randomized coordinate ascent on the concave objective, *i.e.* to select $\tau \in [\![T-1]\!]$ randomly at each iteration and doing a gradient step for Ω_τ. We need to compute a subgradient of the objective $h : \Omega_\tau \mapsto \sum_{t=1}^{T-1} \mathscr{T}_{d^2_{\Omega_t}}(\mu_t, \mu_{t+1}) - \eta \mathfrak{B}^2(\Omega_t, \Omega_{t+1})$, given by:

$$\nabla h(\Omega_\tau) = V(\boldsymbol{\pi}_{\tau\star}) - \eta \partial_1 \mathfrak{B}^2(\Omega_\tau, \Omega_{\tau+1}) - \eta \partial_2 \mathfrak{B}^2(\Omega_{\tau-1}, \Omega_\tau) \tag{3.13}$$

where $\boldsymbol{\pi} \mapsto V(\boldsymbol{\pi})$ is defined in Example 3.2, $\boldsymbol{\pi}_{\tau\star} \in \mathbb{R}^{n \times n}$ is any optimal transport plan between $\mu_\tau, \mu_{\tau+1}$ for cost $d^2_{\Omega_\tau}$, and $\partial_1 \mathfrak{B}^2, \partial_2 \mathfrak{B}^2$ are the gradients of the squared Bures metric with respect to the first and second arguments, computed *e.g.* in [Muzellec and Cuturi, 2018].

Algorithm 4 Randomized (Block) Coordinate Ascent for sequential SRW

Input: Measures $\mu_1, \dots, \mu_T \in \mathscr{P}(\mathbb{R}^d)$, dimension k, learning rate lr
Initialize $\Omega_1, \dots, \Omega_{T-1} \in \mathbb{R}^{d \times d}$
for $i = 0$ **to** MAXITER **do**
  Draw $\tau \in [\![T-1]\!]$
  $\pi_{\tau\star} \leftarrow \mathrm{OT}(\mu_\tau, \mu_{\tau+1}, \mathrm{cost} = d^2_{\Omega_\tau})$
  $\Omega_\tau \leftarrow \mathrm{Proj}_{\mathcal{R}_k} \left[\Omega_\tau + \mathrm{lr} \nabla h(\Omega_\tau) \right]$ using (3.13)
end for

3.8 Experiments

3.8.1 Linearized Entropy-Regularized OT

We consider the entropy-regularized OT problem in the discrete setting:

$$\mathscr{S}^{\varepsilon}_{\mathbf{c}_0}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon R(\boldsymbol{\pi})$$

where $\mathbf{c}_0 \in \mathbb{R}^{n \times n}$ and $R : \boldsymbol{\pi} \mapsto \sum_{ij} \boldsymbol{\pi}_{ij}(\log \boldsymbol{\pi}_{ij} - 1)$. Since R is separable, we can constrain the associated adversarial cost to be nonnegative by linearizing the entropic regularization. By proposition 3.2, this amounts to solve

$$\begin{aligned} &\sup_{\mathbf{c} \in \mathbb{R}_+^{n \times n}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \varepsilon \sum_{ij} \exp\left(\frac{\mathbf{c}_{ij} - \mathbf{c}_{0ij}}{\varepsilon}\right) \\ &= \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon \sum_{ij} \widehat{R}_{ij}(\boldsymbol{\pi}_{ij}) \end{aligned} \tag{3.14}$$

where $\widehat{R}_{ij} : \mathbb{R} \to \mathbb{R}$ is defined as

$$\widehat{R}_{ij}(x) := \begin{cases} x(\log x - 1) & \text{if } x \geq \exp\left(-\frac{\mathbf{c}_{0ij}}{\varepsilon}\right) \\ \frac{-\mathbf{c}_{0ij}}{\varepsilon} x - \exp\left(-\frac{\mathbf{c}_{0ij}}{\varepsilon}\right) & \text{otherwise.} \end{cases}$$

We first consider $N = 100$ couples of measures $(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$ in dimension $d = 1000$, each measure being a uniform measure on $n = 100$ samples from a Gaussian distribution with covariance matrix drawn from a Wishart distribution with $k = d$ degrees of freedom. For each couple, we run Algorithm 3 to solve problem (3.14). This gives an adversarial cost $\mathbf{c}^{\varepsilon}{}_{\star}$. We plot in Figure 3.2 the mean value of $\left|\widehat{W}_{\varepsilon} - \mathscr{T}_{\|\cdot\|^2}(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)\right|$ depending on ε, for $\widehat{W}_{\varepsilon}$ equal to $\mathscr{T}_{\mathbf{c}^{\varepsilon}{}_{\star}}(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$, $\mathscr{S}^{\varepsilon}_{\|\cdot\|^2}(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$ and the value of (3.14). For small values of ε, all three values converge to the real Wasserstein distance. For large ε, Sinkhorn stabilizes to the MMD [Genevay et al., 2016] while the robust cost goes to 0 (for the adversarial cost goes to 0).

In Figure 3.3, we visualize the effect of the regularization ε on the ground cost $\mathbf{c}^{\varepsilon}{}_{\star}$ itself, for measures $\boldsymbol{\mu}, \boldsymbol{\nu}$ plotted in Figure 3.3a. We use multidimensional scaling on the adversarial cost matrix $\mathbf{c}^{\varepsilon}{}_{\star}$ (with distances between points from the same measures unchanged) to recover points in $\mathbb{R}^2$. For large values of ε, the adversarial cost goes to 0, which corresponds in the primal to a fully diffusive transport plan $\boldsymbol{\pi} = \boldsymbol{\mu}\boldsymbol{\nu}^{\top}$.

3.8.2 Learning a Metric on the Color Space

We consider 20 measures $(\boldsymbol{\mu}_i)_{i=1,\dots,10}, (\boldsymbol{\nu}_j)_{j=1,\dots,10}$ on the red-green-blue color space identified with $\mathcal{X} = [0, 1]^3$. Each measure is a point cloud corresponding to the colors used in a painting, divided into two types: ten portraits by

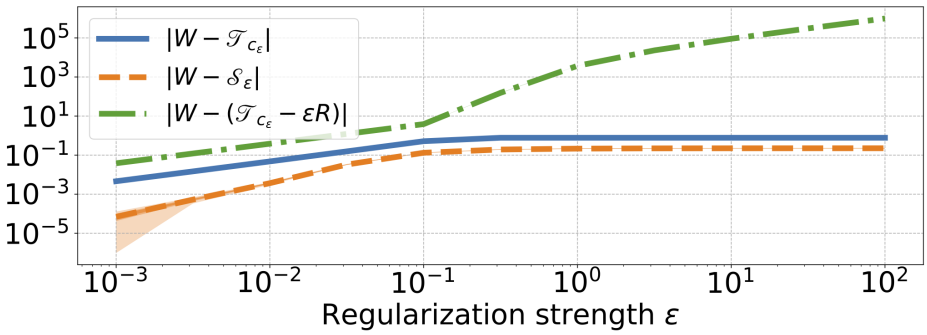
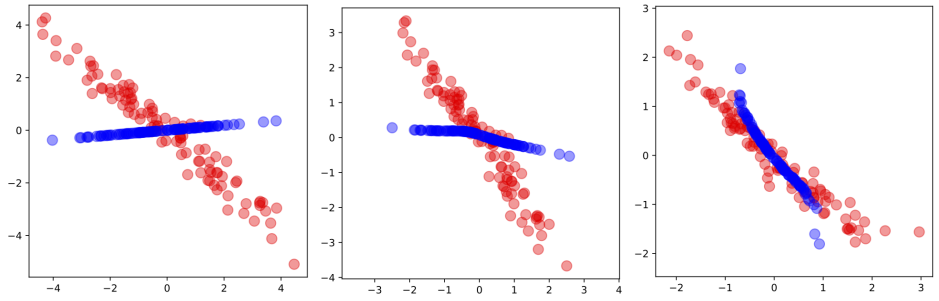


Figure 3.2: Mean value (over 100 runs) of the difference between the classical (2-Wasserstein) OT cost W and Sinkhorn $\mathscr{S}^\varepsilon$ (orange dashed), OT cost with adversarial nonnegative cost $\mathscr{T}_{c_\varepsilon}$ (blue line) and the value of problem (3.14) $\mathscr{T}_{c_\varepsilon} - \varepsilon R$ (green dot-dashed) depending on ε. The shaded areas represent the min-max, 10%-90% and 25%-75% percentiles, and appear negligeable except for numerical errors.



(a) Original Points (b) $\varepsilon = 0.01$ (c) $\varepsilon = 2$

Figure 3.3: Effect of the regularization strength on the metric: as ε grows, the associated adversarial cost shrinks the distances.

Modigliani $(\boldsymbol{\mu}_i, i \in M)$ and ten by Schiele $(\boldsymbol{\nu}_j, j \in S)$. As in SRW and sequential SRW formulations, we learn a metric $c_\Omega \in \mathcal{C}(\mathcal{X}^2)$ parameterized by a matrix $0 \preceq \Omega \preceq I$ such that $\operatorname{trace}\Omega = 1$ that best separates the Modiglianis and the Schieles:

$$\Omega_\star \in \operatorname*{arg\,max}_{\Omega \in \mathcal{R}_1} \sum_{i \in M} \sum_{j \in S} \mathscr{T}_{d^2_\Omega}(\boldsymbol{\mu}_i, \boldsymbol{\nu}_j).$$

We compute $\Omega_\star$ using projected SGD. We then use this "one-dimensional" metric $d^2_{\Omega_\star}$ as a ground metric for OT-based color transfer [Rabin et al., 2014]: an optimal transport plan $\boldsymbol{\pi}$ between two color palettes $\boldsymbol{\mu}_i, \boldsymbol{\nu}_j$ gives a way to transfer colors from one painting to the other. Visually, transferring the colors using the classical quadratic cost $\|\cdot\|^2$ or the adversarially-learnt one-dimensional metric $d^2_{\Omega_\star}$ makes no major difference, showing that when regularized, OT can extract sufficient information from lower dimensional representations.

(a) Modigliani (b) Schiele (c) Metric $d^2_{\Omega_\star}$ (d) Metric $\|\cdot\|^2$

Figure 3.4: Color transfer, best zoomed in. *(a)* and *(b)*: Original paintings. *(c)*: Schiele's painting with Modigliani's colors, using the learnt adversarial one-dimensional metric $d^2_{\Omega_\star}$. *(d)*: Schiele's painting with Modigliani's colors, using the Euclidean metric $\|\cdot\|^2$.

Conclusion

In this chapter, we have shown that any convex regularization of optimal transport can be recast as a ground-cost adversarial problem. This result extends the results of Theorem 2.1 and Lemma 2.2 that we proved in chapter 2. Under some technical assumption on the regularization, we proved a duality theorem for regularized OT, which we use to characterize the optimal ground-cost as a separate function of its two arguments. In order to overcome this degeneration, we proposed to constrain the robust ground-cost to take non-negative values. We also proposed a framework to learn an adversarial sequence of ground-cost functions which is adversarial to a time-varying sequence of measures.

Part II

Regularity-Constrained Maps

Chapter 4

Smooth and Strongly-Convex Nearest Brenier Potentials

4.1 Introduction

The optimal transport (OT) theory is useful in applications because it provides tools that can quantify the closeness between probability measures even when they do not have overlapping supports, and more generally because it defines tools to infer maps that can push-forward (or morph) one measure onto another. In applications, some form of regularization is used to ensure that computations are not only tractable but also meaningful, in the sense that the naive implementation of linear programs to solve OT on discrete histograms/measures are not only too costly but also suffer from the curse of dimensionality [Dudley, 1969, Panaretos and Zemel, 2019]. Regularization, defined explicitly or implicitly as an approximation algorithm, is therefore crucial to ensure that OT is meaningful and can work at scale.

Brenier Potentials and Regularity Theory In the OT literature, regularity has a different meaning, one that is usually associated with the properties of the optimal Monge map [Villani, 2009, §9-10] pushing forward a measure μ onto ν with a small average cost. When that cost is the quadratic Euclidean distance, the Monge map is necessarily the gradient ∇f of a convex function f. This major result, known as Brenier [1987] theorem, states that the OT problem between μ and ν is solved as soon as there exists a convex function f such that $\nabla f_\sharp \mu = \nu$. In that context, regularity in OT is usually understood as the property that the map ∇f is *Lipschitz*, a seminal result due to Caffarelli [2000] who proved that the Brenier map can be guaranteed to be 1-Lipschitz when transporting a "fatter than Gaussian" measure $\mu \propto e^V \gamma_d$ towards a "thinner than Gaussian" measure $\nu \propto e^{-W} \gamma_d$ (here γ_d is the Gaussian measure on $\mathbb{R}^d$, $\gamma_d \propto e^{-\|\cdot\|^2}$, and V, W are two convex potentials). Equivalently, this result can be stated as the fact that the Monge map is the gradient of a 1-smooth Brenier

[1987] potential.

Contributions Our goal in this work is to translate the idea that the OT map between sufficiently well-behaved distributions should be regular into an estimation procedure. Our contributions are:

1. Given two probability measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, a L-smoooth and ℓ-strongly convex function f such that $\nabla f_\sharp \mu = \nu$ may not always exist. We relax this equality and look instead for a smooth strongly convex function f that minimizes the Wasserstein distance between $\nabla f_\sharp \mu$ and ν. We call such potential nearest-Brenier because they provide the "nearest" way to transport μ to a measure like ν using a smooth and strongly convex potential.
2. When μ, ν are discrete probability measures, we show that nearest-Brenier potentials can be recovered as the solution of a QCQP/Wasserstein optimization problem. Our formulation builds upon recent advances in mathematical programming to quantify the worst-case performance of first order methods when used on smooth strongly convex functions [Taylor et al., 2017, Drori and Teboulle, 2014], yet results in simpler, convex problems.
3. In the univariate case, we show that computing the nearest-Brenier potential is equivalent to solving a variant of the isotonic regression problem in which the map must be strongly increasing and Lipschitz. A projected gradient descent approach can be used to solve this problem efficiently.
4. We exploit the solutions to both these optimization problems to extend the Brenier potential and Monge map at any point. We show this can be achieved by solving a QCQP for each new point.
5. We implement and test these algorithms on various tasks, in which smooth strongly convex potentials improve the statistical stability of the estimation of Wasserstein distances, and illustrate them on color transfer and domain adaptation tasks.

4.2 Regularity in Optimal Transport

Convexity and Transport: The Brenier Theorem Let $\mu \in \mathscr{P}_2(\mathbb{R}^d)$ and $f : \mathbb{R}^d \to \mathbb{R}$ convex and differentiable μ-a.e. Then ∇f, as a map from $\mathbb{R}^d$ to $\mathbb{R}^d$ is optimal for the Monge formulation of OT between the measures μ and $\nabla f_\sharp \mu$. The Brenier [1987] theorem shows that if $\mu = p\mathscr{L}^d$ (μ is absolutely continuous w.r.t. $\mathscr{L}^d$ with density p) and $\nu \in \mathscr{P}_2(\mathbb{R}^d)$, there always exists a convex f such that $\nabla f_\sharp \mu = \nu$, *i.e.* there exists an optimal Monge map sending μ to ν that is the gradient of a convex function f. Such a convex function f is called a Brenier potential between μ and ν. If moreover $\nu = q\mathscr{L}^d$, that is ν has density q, a change of variable formula shows that f should be solution to the Monge-Ampère [Villani, 2009, Eq.12.4] equation $\det(\nabla^2 f) = \frac{p}{q \circ \nabla f}$. The

study of the Monge-Ampère equation is the key to obtain regularity results on f and ∇f, see the recent survey by Figalli [2017].

Strong Convexity and Smoothness We recall that a differentiable convex function f is called L-smooth if its gradient function is L-Lipschitz, namely for all $x, y \in \mathbb{R}^d$ we have $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. It is called ℓ-strongly convex if $f - (\ell/2)\|\cdot\|^2$ is convex. Given a partition $\mathcal{E} = (E_1, \ldots, E_K)$ of $\mathbb{R}^d$, we will more generally say that f is $\mathcal{E}$-locally ℓ-strongly convex and L-smooth if the inequality above only holds for pairs (x, y) taken in the interior of any of the subsets E_k. We write $\mathcal{F}_{\ell,L,\mathcal{E}}$ for the set of such functions.

Regularity of OT maps Results on the regularity of the Brenier potential were first obtained by Caffarelli [2000]. For measures $\mu = e^V \gamma_d$ and $\nu = e^{-W}\gamma_d$, where V, W are convex functions and γ_d is the standard Gaussian measure on $\mathbb{R}^d$, Caffarelli's contraction theorem states that any Brenier potential $f_\star$ between μ and ν is 1-smooth. More general results have been proposed by Figalli [2010] who showed that local regularity holds in a general setting: loosely speaking, one can obtain "local Hölder regularity by parts" as soon as the measures have bounded densities and compact support.

4.3 Regularity as Regularization

Contrary to the viewpoint adopted in the OT literature [Caffarelli et al., 1999, Figalli, 2017], we consider here regularity (smoothness) and curvature (strong convexity), as *desiderata*, namely conditions that must be enforced when estimating OT, rather than properties that can be proved under suitable assumptions on μ, ν. Note that if a convex potential is ℓ-strongly convex and L-smooth, the map ∇f has distortion $\ell\|x-y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$. Therefore, setting $\ell = L = 1$ enforces that ∇f must be a translation, since f must be convex. If one were to lift the assumption that f is convex, one would recover the case where ∇f is an isometry, considered in [Cohen and Guibas, 1999, Alt and Guibas, 2000, Alaux et al., 2019]. Note that this distortion also plays a role when estimating the Gromov-Wasserstein distance between general metric spaces [Mémoli, 2011] and was notably used to enforce regularity when estimating the discrete OT problem [Flamary et al., 2014] in a Kantorovich setting. We enforce it here as a constraint in the space of convex functions.

Near-Brenier Smooth Strongly Convex Potentials We will seek functions f that are ℓ-strongly convex and L-smooth (or, alternatively, locally so) while at the same time such that $\nabla f_\sharp \mu$ is as *close as possible* to the target ν. If $\nabla f_\sharp \mu$ were to be exactly equal to ν, such a function would be called a Brenier potential. We quantify that nearness in terms of the Wasserstein distance between the pushfoward of μ and ν to define:

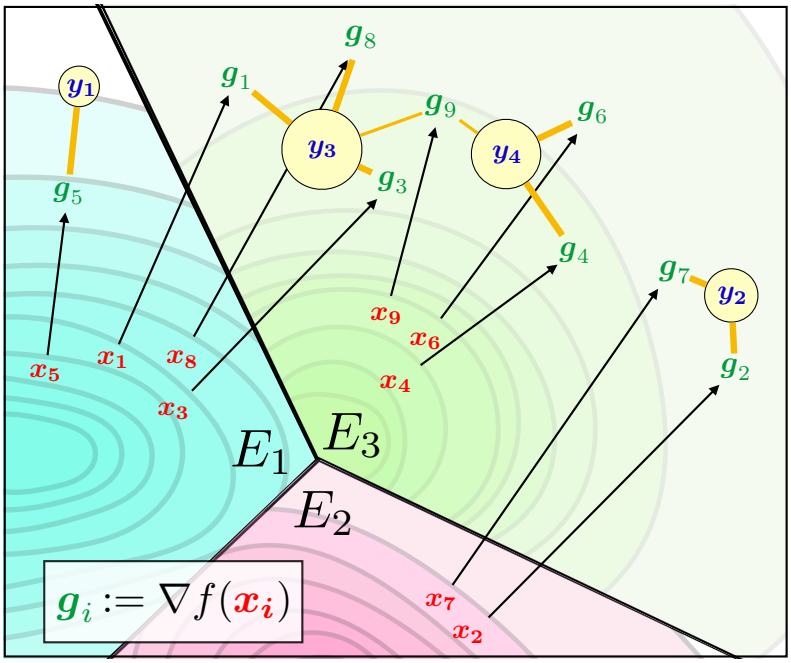


Figure 4.1: Points x_i mapped onto points $g_i := \nabla f(x_i)$ for a function f that is locally smooth strongly convex. SSNB potentials are such that the measure of endpoints g_i is as close as possible (in W_2 sense) to the measure supported on the y_j. Here this would be the sum of the squares of the length of these orange sticks.

Definition 4.1. *Let $\mathcal{E}$ be a partition of $\mathbb{R}^d$ and $0 \leq \ell \leq L$. For $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, we call $f_\star$ a ($\mathcal{E}$-locally) L-smooth ℓ-strongly convex nearest Brenier (SSNB) potential between μ and ν if*

$$f_\star \in \operatorname*{arg\,min}_{f \in \mathcal{F}_{\ell,L,\mathcal{E}}} W_2\left(\nabla f_\sharp \mu, \nu\right).$$

Remark 4.1. *When $\mathcal{E} = \{\mathbb{R}^d\}$, the gradient of any SSNB potential $f_\star$ defines an optimal transport map between μ and $\nabla f_{\star\sharp}\mu$. The associated transport value $W_2^2\left(\mu, \nabla f_{\star\sharp}\mu\right)$ does not define a metric between μ and ν because it is not symmetric, and $W_2(\mu, \nabla f_{\star\sharp}\mu) = 0 \not\Rightarrow \mu = \nu$ (take any ν that is not a Dirac and $\mu = \delta_{\mathbb{E}[\nu]}$). For more general partitions $\mathcal{E}$ one only has that property locally, and $f_\star$ can therefore be interpreted as a piecewise convex potential, giving rise to piecewise optimal transport maps, as illustrated in Figure 4.1.*

We now give a proof for the existence of SSNB potentials:

Proof. We write a proof of existence in the case where $\mathcal{E} = \{\mathbb{R}^d\}$. If $K > 1$, the proof can be applied independently on each set of the partition.

Let $(f_n)_{n\in\mathbb{N}}$ be such that $f_n(0) = 0$ for all $n \in \mathbb{N}$ and

$$W_2\left[(\nabla f_n)_\sharp \mu, \nu\right] \leq \frac{1}{n+1} + \inf_{f\in\mathcal{F}_{\ell,L}} W_2\left[(\nabla f)_\sharp \mu, \nu\right].$$

Let $x_0 \in \mathrm{supp}(\mu)$. Then there exists $C > 0$ such that for all $n \in \mathbb{N}$, $\|\nabla f_n(x_0)\| \leq C$. Indeed, suppose this is not true. Take $r > 0$ such that $V := \mu[B(x_0, r)] > 0$. By Prokhorov theorem, there exists $R > 0$ such that $\nu\left[B(0, R)\right] \geq 1 - \frac{V}{2}$. Then for $C > 0$ large enough, there exists an $n \in \mathbb{N}$ such that:

$$\begin{aligned}
W_2^2[(\nabla f_n)_\sharp \mu, \nu] &= \min_{\pi\in\Pi(\mu,\nu)} \int \|\nabla f_n(x) - y\|^2 \, d\pi(x, y) \\
&\geq \int \|\nabla f_n(x) - \mathrm{proj}_{B(0,R)}\left[\nabla f_n(x)\right]\|^2 \, d\mu(x) \\
&\geq \int_{B(x_0,r)\cap \mathrm{supp}(\mu)} \|\nabla f_n - \mathrm{proj}_{B(0,R)}\left[\nabla f_n\right]\|^2 \, d\mu \\
&\geq \frac{1}{2} V \min_{\substack{x\in B(x_0,r) \\ y\in B(0,R)}} \|\nabla f_n(x) - y\|^2 \\
&\geq \frac{1}{2} V (C - Lr - R)
\end{aligned}$$

which contradicts the definition of f_n when C is sufficiently large.
Then for $x \in \mathbb{R}^d$,

$$\|\nabla f_n(x)\| \leq L\|x - x_0\| + \|\nabla f_n(x_0)\| \leq L\|x - x_0\| + C.$$

Since $(\nabla f_n)_{n\in\mathbb{N}}$ is equi-Lipschitz, it converges uniformly (up to a subsequence) to some function g by Arzelà–Ascoli theorem. Note that g is L-Lipschitz.
Let $\varepsilon > 0$ and let $N \in \mathbb{N}$ such that $n \geq N \Rightarrow \|\nabla f_n - g\|_\infty \leq \varepsilon$. Then for $n \geq N$ and $x \in \mathbb{R}^d$,

$$|f_n(x)| = \left| \int_0^1 \langle \nabla f_n(tx), x \rangle \, dt \right| \leq \|x\|(\|g\|_\infty + \varepsilon)$$

so that $(f_n(x))$ converges up to a subsequence. Let ϕ, ψ be two extractions and α, β such that $f_{\phi(n)}(x) \to \alpha$ and $f_{\psi(n)}(x) \to \beta$. Then

$$\begin{aligned}
|\alpha - \beta| &= \lim_{n\to\infty} \left| \int_0^1 \langle \nabla f_{\phi(n)}(tx) - \nabla f_{\psi(n)}(tx), x \rangle \, dt \right| \\
&\leq \lim_{n\to\infty} \|x\| \|\nabla f_{\phi(n)} - \nabla f_{\psi(n)}\|_\infty = 0.
\end{aligned}$$

This shows that $(f_n)_{n\in N}$ converges pointwise to some function $f_\star$. In particular, $f_\star$ is convex. For $x \in \mathbb{R}^d$, using Lebesgue's dominated convergence theorem,

$$\begin{aligned}
f_\star(x) &= \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \int_0^1 \langle \nabla f_n(tx), x \rangle \, dt \\
&= \int_0^1 \left\langle \lim_{n\to\infty} \nabla f_n(tx), x \right\rangle dt = \int_0^1 \langle g(tx), x \rangle \, dt
\end{aligned}$$

so $f_\star$ is differentiable and $\nabla f_\star = g$. Using Lebesgue's dominated convergence theorem, uniform (hence pointwise) convergence of $(\nabla f_n)_{n\in\mathbb{N}}$ to $\nabla f_\star$ shows that $(\nabla f_n)_\sharp \mu \rightharpoonup (\nabla f_\star)_\sharp \mu$. Then classical optimal transport stability theorems *e.g.* [Villani, 2009, Theorem 5.19] show that

$$W_2\left[(\nabla f_\star)_\sharp \mu, \nu\right] = \lim_{n\to\infty} W_2\left[(\nabla f_n)_\sharp \mu, \nu\right] = \inf_{f\in\mathcal{F}_{\ell,L}} W_2\left[(\nabla f)_\sharp \mu, \nu\right],$$

i.e. $f_\star$ is a minimizer. □

Algorithmic Formulation as an Alternate QCQP/Wasserstein Problem We will work from now on with two discrete measures $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$, with supports defined as $x_1, \dots x_n \in \mathbb{R}^d$, $y_1, \dots y_m \in \mathbb{R}^d$, and $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_m)$ are probability weight vectors. We write $\mathcal{U}(\mathbf{a}, \mathbf{b})$ for the transportation polytope with marginals $\mathbf{a}$ and $\mathbf{b}$, namely the set of $n \times m$ matrices with nonnegative entries such that their row-sum and column-sum are respectively equal to $\mathbf{a}$ and $\mathbf{b}$. Set a desired smoothness $L > 0$ and strong-convexity parameter $\ell \leq L$, and choose a partition $\mathcal{E}$ of $\mathbb{R}^d$ (in our experiments $\mathcal{E}$ is either $\{\mathbb{R}^d\}$, or computed using a K-means partition of $\operatorname{supp}\mu$). For $k \in [\![K]\!]$, we write $I_k = \{i \in [\![n]\!] \text{ s.t. } x_i \in E_k\}$. The infinite dimensional optimization problem introduced in Definition 4.1 can be reduced to a QCQP that only focuses on the values and gradients of f at the points x_i. This result follows from the literature in the study of first order methods, which considers optimizing over the set of convex functions with prescribed smoothness and strong-convexity constants (see for instance [Taylor, 2017, Theorem 3.8 and Theorem 3.14]). We exploit such results to show that an SSNB f can not only be estimated at those points x_i, but also more generally recovered at any arbitrary point in $\mathbb{R}^d$.

Theorem 4.1. *The n values $u_i := f(x_i)$, and gradients $z_i := \nabla f(x_i)$ of a SSNB potential $f \in \mathcal{F}_{\ell,L,\mathcal{E}}$ can be recovered as:*

$$\min_{\substack{z_1,\dots z_n\in\mathbb{R}^d \\ u\in\mathbb{R}^n}} W_2^2\left(\sum_{i=1}^n a_i\delta_{z_i}, \nu\right) := \min_{P\in\mathcal{U}(a,b)} \sum_{i,j} P_{ij}\|z_i - y_j\|^2 \tag{4.1}$$

$$\begin{aligned} s.t.\ \forall k \leq K, \forall i,j \in I_k,& \\ u_i &\geq u_j + \langle z_j, x_i - x_j\rangle \\ &\quad + \frac{1}{2(1-\ell/L)}\left(\frac{1}{L}\|z_i - z_j\|^2 + \ell\|x_i - x_j\|^2\right. \\ &\qquad \left. -2\frac{\ell}{L}\langle z_j - z_i, x_j - x_i\rangle\right). \end{aligned}$$

Moreover, for $x \in E_k$, $v := f(x)$ and $g := \nabla f(x)$ can be recovered as:

$$\begin{aligned} &\min_{v\in\mathbb{R},\, g\in\mathbb{R}^d} v \\ &\text{s.t.}\, \forall i \in I_k, v \geq u_i + \langle z_i, x - x_i \rangle \\ &\qquad + \frac{1}{2(1-\ell/L)} \left(\frac{1}{L}\|g - z_i\|^2 \right. \\ &\qquad \left. + \ell\|x - x_i\|^2 - 2\frac{\ell}{L}\langle z_i - g, x_i - x \rangle \right). \end{aligned} \tag{4.2}$$

Proof. For $f \in \mathcal{F}_{\ell,L,\mathcal{E}}$, $\nabla f_\sharp \mu = \sum_{i=1}^n a_i \delta_{\nabla f(x_i)}$. Writing $z_i = \nabla f(x_i)$, we wish to minimize $W_2^2\left(\sum_{i=1}^n a_i \delta_{z_i}, \nu\right)$ over all the points $z_1, \dots, z_n \in \mathbb{R}^d$ such that there exists $f \in \mathcal{F}_{\ell,L,\mathcal{E}}$ with $\nabla f(x_i) = z_i$ for all $i \in [\![n]\!]$. Following [Taylor, 2017, Theorem 3.8], there exists such a f if, and only if, there exists $u \in \mathbb{R}^n$ such that for all $k \in [\![K]\!]$ and for all $i, j \in I_k$,

$$\begin{aligned} u_i \geq u_j + \langle z_j, x_i - x_j \rangle + \frac{1}{2(1-\ell/L)} \left(\frac{1}{L}\|z_i - z_j\|^2 \right. \\ \left. + \ell\|x_i - x_j\|^2 - 2\frac{\ell}{L}\langle z_j - z_i, x_j - x_i \rangle \right). \end{aligned}$$

Then minimizing over $f \in \mathcal{F}_{\ell,L,\mathcal{E}}$ is equivalent to minimizing over $(z_1, \dots, z_n, u)$ under these interpolation constraints.

The second part of the theorem is a direct application of [Taylor, 2017, Theorem 3.14]. □

We provide algorithms to compute a SSNB potential in dimension $d \geq 2$ when μ, ν are discrete measures. In order to solve Problem (4.1), we will alternate between minimizing over $(z_1, \dots, z_n, u)$ and computing a coupling P solving the OT problem. The OT computation can be efficiently carried out using Sinkhorn's algorithm [Cuturi, 2013]. The other minimization is a convex QCQP, separable in K smaller convex QCQP that can be solved efficiently. We use the barycentric projection (see Definition 4.2 below) of μ as an initialization for the points z.

4.4 One-dimensional Case and the Link with Constrained Isotonic Regression

We consider first SSNB potentials in arguably the simplest case, namely that of distributions on the real line. We use the definition of the "barycentric projection" of a coupling [Ambrosio et al., 2006, Def.5.4.2], which is the most geometrically meaningful way to recover a map from a coupling.

Definition 4.2 (Barycentric Projection)**.** *Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, and take π an optimal transport plan between μ and ν. The barycentric projection of π is defined as the map $\overline{\pi} : x \mapsto \mathbb{E}_{(X,Y)\sim\pi}[Y|X = x]$.*

Theorem 12.4.4 in [Ambrosio et al., 2006] shows that $\overline{\pi}$ is the gradient a convex function. It is then admissible for the SSNB optimization problem defined in Theorem 4.1 as soon as it verifies regularity (Lipschitzness) and curvature (strongly increasing). Although the barycentric projection map is not optimal in general, the following proposition shows that it is however optimal for univariate measures:

Proposition 4.1. *Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R})$ and $0 \leq \ell \leq L$. Suppose $\mu \ll \mathscr{L}^1$, or is purely atomic. Then the set of SSNB potentials between μ and ν is the set of solutions to*

$$\min_{f \in \mathcal{F}_{\ell,L,\mathcal{E}}} \|f' - \overline{\pi}\|^2_{L^2(\mu)}$$

where π is the unique optimal transport plan between μ and ν given by [Santambrogio, 2015, Theorem 2.9].

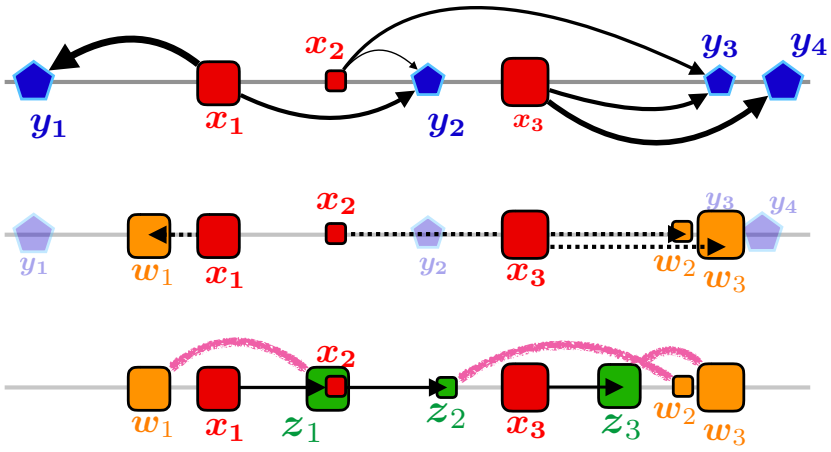


Figure 4.2: Top: optimal transport between two discrete measures μ, ν. Middle: the barycentric projection $\mathbf{w}$ of points $\mathbf{x}$ is displayed and corresponds to a Monge map (no mass splitting). Considering here for instance $\ell = 0.5$ and $L = 1$, the map that associates x_i to w_i is not 1-Lipschitz at pairs $(1, 2)$ or $(1, 3)$ and over-contracting in pair $(2, 3)$. Bottom: To compute points z_i that minimize their transport cost to the w_i (pink curves) while still ensuring $x_i \mapsto z_i$ is L-Lipschitz and strongly increasing amounts to solving the L-Lipschitz ℓ-strongly increasing isotonic regression problem (4.3).

Proof. Let $f : \mathbb{R} \to \mathbb{R}$. Then $f \in \mathcal{F}_{\ell,L,\mathcal{E}}$ if and only if it is ℓ-strongly convex and L-smooth on each set E_k, $k \in [\![K]\!]$, *i.e.* if and only if for any $k \in [\![K]\!]$, $\ell \leq f''|_{E_k} \leq L$.

For a measure ρ, let us write F_ρ and Q_ρ the cumulative distribution function and the quantile function (*i.e.* the generalized inverse of the cumulative distribution function). Then $Q_{\nabla f_\sharp \mu} = \nabla f \circ Q_\mu$.

Using the closed-form formula for the Wasserstein distance in dimension 1, the objective we wish to minimize (over $f \in \mathcal{F}_{\ell,L,\mathcal{E}}$) is:

$$W_2^2(f'_\sharp \mu, \nu) = \int_0^1 \left[f' \circ Q_\mu(t) - Q_\nu(t)\right]^2 dt.$$

Suppose μ has a density w.r.t the Lebesgue measure. Then by a change of variable, the objective becomes

$$\int_{-\infty}^{+\infty} \left[f'(x) - Q_\nu \circ F_\mu(x)\right]^2 d\mu(x) = \|f' - \overline{\pi}\|^2_{L^2(\mu)}.$$

Indeed, $Q_\nu \circ F_\mu$ is the optimal transport map from μ to ν, hence its own barycentric projection. The result follows.

Suppose now that μ is purely atomic, and write $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ with $x_1 \leq \ldots \leq x_n$. For $0 \leq i \leq n$, put $\alpha_i = \sum_{k=1}^i a_k$ with $\alpha_0 = 0$. Then

$$\begin{aligned} W_2^2(f'_\sharp \mu, \nu) &= \sum_{i=1}^n \int_{\alpha_{i-1}}^{\alpha_i} (f'(x_i) - Q_\nu(t))^2 dt \\ &= \sum_{i=1}^n a_i \left[f'(x_i) - \frac{1}{a_i}\left(\int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t) dt\right)\right]^2 \\ &\quad + \int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t)^2 dt - \frac{1}{a_i}\left(\int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t) dt\right)^2. \end{aligned}$$

Since $\sum_{i=1}^n \int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t)^2 dt - \frac{1}{a_i}\left(\int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t) dt\right)^2$ does not depend on f, minimizing $W_2^2(f'_\sharp \mu, \nu)$ over $f \in \mathcal{F}_{\ell,L,\mathcal{E}}$ is equivalent to solve

$$\min_{f \in \mathcal{F}_{\ell,L,\mathcal{E}}} \sum_{i=1}^n a_i \left[f'(x_i) - \frac{1}{a_i}\left(\int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t) dt\right)\right]^2.$$

There only remains to show that $\overline{\pi}(x_i) = \frac{1}{a_i}\int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t)\, dt$. Using the definition of the conditional expectation and the definition of π:

$$\begin{aligned} \overline{\pi}(x_i) &= \frac{1}{a_i}\int_{-\infty}^{+\infty} y\, \mathbf{1}\{x = x_i\}\, d\pi(x,y) \\ &= \frac{1}{a_i}\int_{-\infty}^{+\infty} y\, \mathbf{1}\{x = x_i\}\, d(Q_\mu, Q_\nu)_\sharp \mathscr{L}^1|_{[0,1]} \\ &= \frac{1}{a_i}\int_0^1 Q_\nu(t)\, \mathbf{1}\{Q_\mu(t) = x_i\}\, dt \\ &= \frac{1}{a_i}\int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t)\, dt. \end{aligned}$$

□

Discrete Computations Suppose $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ is discrete with $x_1 \leq \ldots \leq x_n$, and ν is arbitrary. Let us denote by Q_ν the (generalized) quantile function of ν. Writing π for the optimal transport plan between μ and ν, the barycentric projection $\overline{\pi}$ is explicit. Writing $\alpha_0 := 0$, $\alpha_i := \sum_{k=1}^i a_k$, one has $\overline{\pi}(x_i) = \frac{1}{a_i} \int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t)\, dt$.

If ν is also discrete, with weights $\mathbf{b} = (b_1, \ldots, b_m)$ and sorted support $\mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$, where $y_1 \leq \cdots \leq y_m$, one can recover the coordinates of the vector $(\overline{\pi}(x_i))_i$ of barycentric projections as

$$\mathbf{w} := \mathrm{diag}(\mathbf{a}^{-1})\mathbf{NW}(\mathbf{a}, \mathbf{b})\, \mathbf{y},$$

where $\mathbf{NW}(\mathbf{a}, \mathbf{b})$ is the so-called *North-west corner* solution [Peyré and Cuturi, 2019, §3.4.2] obtained in linear time w.r.t n, m by simply filling up greedily the transportation matrix from top-left to down-right. We deduce from Proposition 4.1 that a SSNB potential can be recovered by solving a weighted (and local, according to the partition $\mathcal{E}$) constrained isotonic regression problem (see Fig. 4.2):

$$\begin{aligned} &\min_{z \in \mathbb{R}^n} \sum_{i=1}^n a_i (z_i - w_i)^2 \\ &\text{s.t. } \forall k \leq K,\, \forall i, i+1 \in I_k, \\ &\qquad \ell(x_{i+1} - x_i) \leq z_{i+1} - z_i \leq L(x_{i+1} - x_i). \end{aligned} \tag{4.3}$$

The gradient of a SSNB potential $f_\star$ can then be retrieved by taking an interpolation of $x_i \mapsto z_i$ that is piecewise affine.

Algorithms solving the Lipschitz isotonic regression were first designed by Yeganova and Wilbur [2009] with a $\mathcal{O}(n^2)$ complexity. [Agarwal et al., 2010, Kakade et al., 2011] developed $\mathcal{O}(n \log n)$ algorithms. A Smooth NB potential can therefore be exactly computed in $\mathcal{O}(n \log n)$ time, which is the same complexity as of optimal transport in one dimension. Adding up the strongly increasing property, Problem (4.3) can also be seen as least-squares regression problem with box constraints. Indeed, introducing m variables $v_i \geq 0$, and defining z_i as the partial sum $\mathbf{v}$, namely $z_i = \sum_{j=1}^i v_j$ (or equivalently $v_i = z_i - z_{i-1}$ with $z_0 := 0$), and writing $u_i^- = \ell(x_{i+1} - x_i)$, $u_i^+ = L(x_{i+1} - x_i)$ one aims to find $\mathbf{v}$ that minimizes $\|A\mathbf{v} - \mathbf{w}\|_a^2$ s.t. $\mathbf{u}^- \leq \mathbf{v} \leq \mathbf{u}^+$, where A is the lower-triangular matrix of ones and $\|\cdot\|_a$ is the Euclidean norm weighted by a. In our experiments, we have found that a projected gradient descent approach to solve this problem performed in practice as quickly as more specialized algorithms and was easier to parallelize when comparing a measure μ to several measures ν.

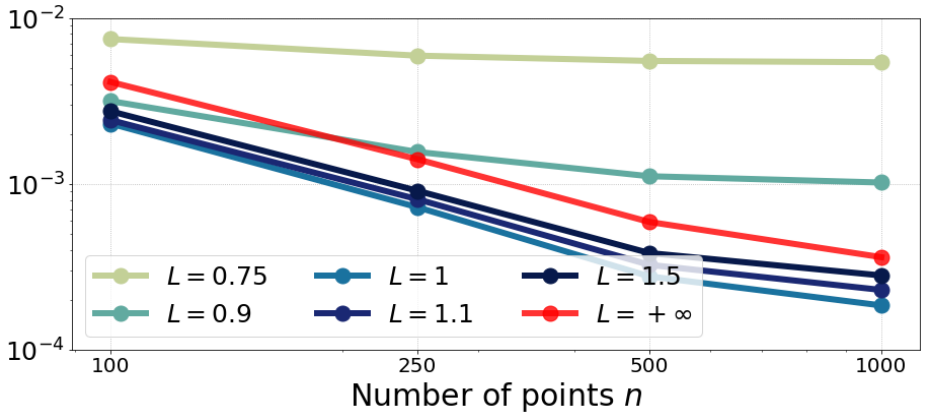


Figure 4.3: Take measures $\mu = \nu$ be the uniform measure over $[0, 1]$. For several n, we consider $\hat{\mu}_n, \hat{\nu}_n$ empirical measures over n iid samples from μ, ν, from which we compute a SSNB potential $\hat{f}_n$ with different values L, and $\ell = \min\{1, L\}$ (and $\ell = 0$ if $L = \infty$). We plot the estimation error $|W_2^2(\hat{\mu}_n, \hat{\zeta}_n) - W_2^2(\mu, \nu)|$ depending on n and L, averaged over 100 runs, where $\hat{\zeta}_n = \hat{f}_{n\sharp}\hat{\mu}_n$. If $L \leq \mathrm{Lip}(\mathrm{Id}) = 1$, the error does not converge to 0. Otherwise, the convergence is faster when L is closer to 1. The case $L = \infty$ corresponds to the classical OT estimator $\hat{\zeta}_n = \hat{\nu}_n$.

4.5 Estimation of the Wasserstein Distance and Monge Map

Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ be two compactly supported measures with densities w.r.t the Lebesgue measure in $\mathbb{R}^d$. Brenier theorem gives the existence of an optimal Brenier potential $f_\star$, *i.e.* $f_\star$ is convex and $\nabla f_{\star\sharp}\mu = \nu$. Our goal is twofold: estimate the map $\nabla f_\star$ and the value of $W_2(\mu, \nu)$ from samples.

Let $x_1, \dots, x_n \sim \mu$ and $y_1, \dots, y_n \sim \nu$ be i.i.d samples from the measures, and define $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ the empirical measures over the samples.

Let $\hat{f}_n$ be a SSNB potential between $\hat{\mu}_n$ and $\hat{\nu}_n$ with $\mathcal{E} = \{\mathbb{R}^d\}$. Then for $x \in \operatorname{supp}\mu$, a natural estimator of $\nabla f_\star(x)$ is given by a solution $\nabla \hat{f}_n(x)$ of (4.2). This defines an estimator $\nabla \hat{f}_n : \mathbb{R}^d \to \mathbb{R}^d$ of $\nabla f_\star$, that we use to define a plug-in estimator for $W_2(\mu, \nu)$:

Definition 4.3. *We define the SSNB estimator of* $W_2^2(\mu, \nu)$ *as*

$$\widehat{W}_2^2(\mu, \nu) := \int \|x - \nabla \hat{f}_n(x)\|^2 \, d\mu(x).$$

Since $\nabla \hat{f}_n$ is the gradient of a convex Brenier potential when $\mathcal{E} = \{\mathbb{R}^d\}$, it is optimal between μ and $\nabla \hat{f}_{n\sharp}\mu$ and $\widehat{W}_2(\mu, \nu) = W_2(\mu, \nabla \hat{f}_{n\sharp}\mu)$. If $\mathcal{E} \neq \{\mathbb{R}^d\}$, $\nabla \hat{f}_n$ is the gradient of a locally convex Brenier potential, and is not necessarily

globally optimal between μ and $\nabla \hat{f}_{n\sharp}\mu$. In that case $\widehat{W}_2(\mu,\nu)$ is an approximate upper bound of $W_2(\mu, \nabla \hat{f}_{n\sharp}\mu)$. In any case, the SSNB estimator can be computed using Monte-Carlo integration, whose estimation error does not depend on the dimension d.

Algorithm 5 Monte-Carlo approximation of the SSNB estimator

Input: $\hat{\mu}_n$, $\hat{\nu}_n$, partition $\mathcal{E}$, number N of Monte-Carlo samples
. $(u_i, z_i)_{i\leq n} \leftarrow$ solve SSNB (4.1) between $\hat{\mu}_n$, $\hat{\nu}_n$
for $j \in [\![N]\!]$ **do**
. Draw $\hat{x}_j \sim \mu$
. Find k s.t. $\hat{x}_j \in E_k$ (k-means)
. $\nabla \hat{f}_n(\hat{x}_j) \leftarrow$ solve QCQP (4.2)
end for
Output: $\widehat{W} = \left[(1/N)\sum_{j=1}^N \|\hat{x}_j - \nabla \hat{f}_n(\hat{x}_j)\|^2\right]^{1/2}$

We show that when the Brenier potential $f_\star$ is globally regular, the SSNB estimator is strongly consistent:

Proposition 4.2. *Choose $\mathcal{E} = \{\mathbb{R}^d\}$, $0 \leq \ell \leq L$. If $f_\star \in \mathcal{F}_{\ell,L,\mathcal{E}}$, it almost surely holds:*

$$\left|W_2(\mu,\nu) - \widehat{W}_2(\mu,\nu)\right| \underset{n\to\infty}{\longrightarrow} 0.$$

Proof. Since $\mathcal{E} = \{\mathbb{R}^d\}$, and using the triangular inequality for the Wasserstein distance,

$$\begin{aligned}
\left|W_2(\mu,\nu) - \widehat{W}_2(\mu,\nu)\right| &= \left|W_2(\mu,\nu) - W_2(\mu, \nabla \hat{f}_{n\sharp}\mu)\right| \\
&\leq W_2\left(\nabla \hat{f}_{n\sharp}\mu, \nu\right) \\
&\leq W_2\left(\nabla \hat{f}_{n\sharp}\mu, \nabla \hat{f}_{n\sharp}\hat{\mu}_n\right) \qquad (4.4) \\
&\quad + W_2\left(\nabla \hat{f}_{n\sharp}\hat{\mu}_n, \hat{\nu}_n\right) \qquad (4.5) \\
&\quad + W_2\left(\hat{\nu}_n, \nu\right). \qquad (4.6)
\end{aligned}$$

We now successively upper bound terms (4.4), (4.5), (4.6).

Since $\nabla \hat{f}_n$ is L-Lispchitz, almost surely:

$$(4.4) = W_2\left(\nabla \hat{f}_{n\sharp}\mu, \nabla \hat{f}_{n\sharp}\hat{\mu}_n\right) \leq L\, W_2\left(\mu, \hat{\mu}_n\right) \underset{n\to\infty}{\longrightarrow} 0$$

since almost surely, $\hat{\mu}_n \rightharpoonup \mu$ and μ has compact support, cf. [Santambrogio, 2015, Theorem 5.10]. For the same reason, almost surely:

$$(4.6) = W_2\left(\hat{\nu}_n, \nu\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Finally, since $f_\star \in \mathcal{F}_{\ell,L,\mathcal{E}}$ and $\nabla \hat{f}_n$ is an optimal SSNB potential, it almost surely holds:

$$(4.5) = W_2\left(\nabla \hat{f}_{n\sharp}\hat{\mu}_n, \hat{\nu}_n\right) \leq W_2\left(\nabla f_{\star\sharp}\hat{\mu}_n, \hat{\nu}_n\right) \underset{n\to\infty}{\longrightarrow} W_2\left(\nabla f_{\star\sharp}\mu, \nu\right) = 0$$

because $(\hat{\mu}_n, \hat{\nu}_n) \rightharpoonup (\mu, \nu)$, and by definition of $f_\star$, $\nabla f_{\star\sharp}\mu = \nu$. □

The study of the theoretical rate of convergence of this estimator is beyond the scope of this work. Recent results [Hütter and Rigollet, 2019, Flamary et al., 2019] show that assuming some regularity of the Monge map $\nabla f_\star$ leads to improved sample complexity rates. Numerical simulations (see section 4.6.1) seem to indicate a reduced estimation error for SSNB over the classical $W_2(\hat{\mu}_n, \hat{\nu}_n)$, even in the case where $\nabla f_\star$ is only locally Lipschitz and $\mathcal{E} \neq \{\mathbb{R}^d\}$. If $L < \mathrm{Lip}(\nabla f_\star)$, the SSNB estimator $\widehat{W}_2(\mu, \nu)$ is not consistent, as can be seen in Figure 4.3.

4.6 Experiments

All the computations were performed on a i9 2,9 GHz CPU, using MOSEK as a convex QCQP solver.

4.6.1 Numerical Estimation of Wasserstein Distances and Monge Maps

In this experiment, we consider two different settings:

1. **Global regularity:** μ is the uniform measure over the unit cube in $\mathbb{R}^d$, and $\nu = T_\sharp\mu$ where $T(x) = \Omega_d x$, where Ω_d is the diagonal matrix with diagonal coefficients: $0.8 + \frac{0.4}{d-1}(i-1)$, for $i \in [\![d]\!]$. T is the gradient of the convex function $f : x \mapsto \frac{1}{2}\|\Omega^{1/2}x\|^2$, so it is the optimal transport map. f is globally $\ell = 0.8$-strongly convex and $L = 1.2$-smooth.

2. **Local Regularity:** μ is the uniform measure over the unit ball in $\mathbb{R}^d$, and $\nu = T_\sharp\mu$ where $T(x_1, \ldots, x_d) = (x_1 + 2\,\mathrm{sign}(x_1), x_2, \ldots, x_d)$. As can be seen in Figure 4.4 (right), T splits the unit ball into two semi-balls. T is a subgradient of the convex function $f : x \mapsto \frac{1}{2}\|x\|^2 + 2|x_1|$, so it is the optimal transport map. f is $\ell = 1$-strongly convex, but is not smooth: ∇f is not even continuous. However, f is $L = 1$-smooth by parts.

For each of those two settings, we consider i.i.d samples $x_1, \ldots, x_n \sim \mu$ and $y_1, \ldots, y_n \sim \nu$ for different values of $n \in \mathbb{N}$, and denote by $\hat{\mu}_n$ and $\hat{\nu}_n$ the respective empirical measures on these points. Given a number of clusters $1 \leq K \leq n$, we compute the partition $\mathcal{E} = \{E_1, \ldots, E_K\}$ by running k-means with K clusters on data $x_1, \ldots, x_n$. In both settings, we run the algorithms with $\hat{\ell} = 0.6$ and $\hat{L} = 1.4$. We give experimental results on the statistical

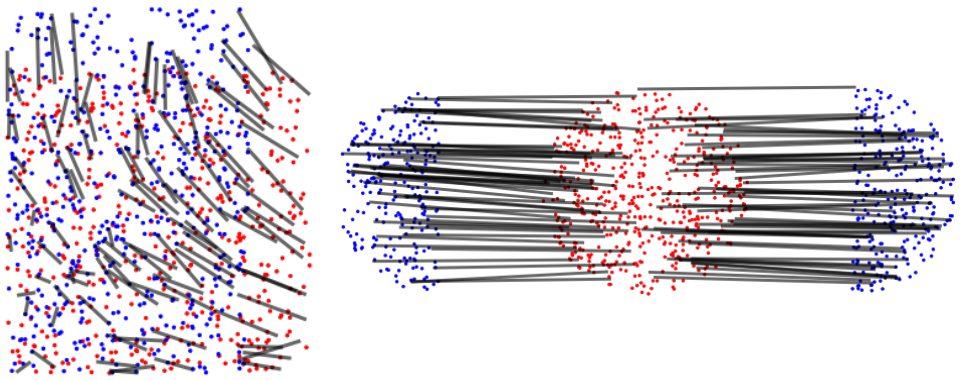


Figure 4.4: In the global (left) and local (right) regularity settings with $d = 2$, we plot $n = 500$ i.i.d samples from μ (red) and ν (blue). Black lines show the displacements $x \mapsto \nabla \hat{f}_n(x)$ for some new points $x \in \text{supp}(\mu) \setminus \text{supp}(\hat{\mu}_n)$, computed by solving QCQP (4.2).

performance of our SSNB estimator, computed using Monte-Carlo integration, compared to the classical optimal transport estimator $W_2(\hat{\mu}_n, \hat{\nu}_n)$. This performance depends on three parameters: the number n of points, the dimension d and the number of clusters K.

In Figure 4.5, we plot the estimation error depending on the cluster ratio K/n for fixed $n = 60$ and $d = 30$. In the global regularity setting (Figure 4.5 top), the error seems to be exponentially increasing in K, whereas the computation time decreases with K: there is a trade-off between accuracy and computation time. In the local regularity setting (Figure 4.5 bottom), the estimation error is large when the number of clusters is too small (because we ask for too much regularity) or too large. Interestingly, the number of clusters can be taken large and still leads to good results. In both settings, even a large number of clusters is enough to outperform the classical OT estimator, even when SSNB is computed with a much smaller number of points. Note that when $K/n = 1$, the SSNB estimator is basically equivalent (up to the Monte-Carlo integration error) to the classical OT estimator.

In Figure 4.6, we plot the estimation error depending on the number of points n, for fixed cluster ratio K/n and different dimension $d \in \{2, 30\}$. In both settings, and for both low ($d = 2$) and high ($d = 30$) dimension, the SSNB estimator seems to have the same rate as the classical OT estimator, but a much better constant in high dimension. This means that in high dimension, the SSNB estimator computed with a small number of points can be much more accurate than the classical OT estimator computed with a large number of points.

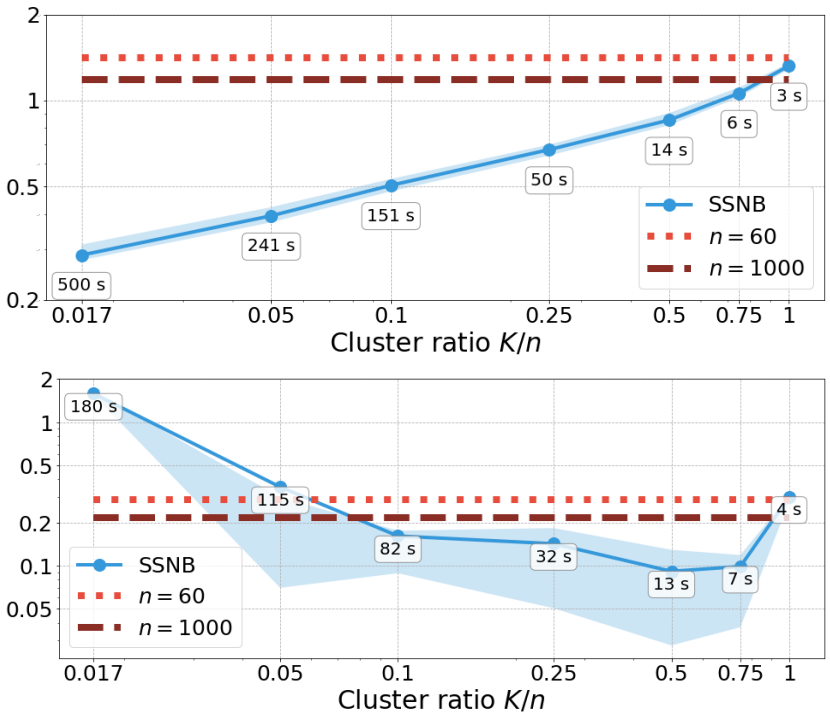


Figure 4.5: In the global (top) and local (bottom) regularity settings, we plot (on a log-log scale) the estimation error $|W_2(\mu,\nu) - \widehat{W}_2(\mu,\nu)|$ (blue line) depending on the cluster ratio K/n, with fixed number of points $n = 60$, dimension $d = 30$ and Monte-Carlo samples $N = 50$. The curves/shaded areas show the mean error/25%-75% percentiles over 20 data samples. The bubbles show the mean running time for the SSNB estimator computation. We plot the classical estimation error for $n = 60$ (light red dotted) and $n = 1000$ (dark red dashed) for comparison.

4.6.2 Domain Adaptation

Domain adaptation is a way to transfer knowledge from a source to a target dataset. The source dataset consists of labelled data, and the goal is to classify the target data. Courty et al. [2016], Perrot et al. [2016] proposed to use optimal transport (and different regularized version of OT) to perform such a task: the OT map from the source dataset to the target dataset (seen as empirical measures over the source/target data) is computed. Then each target data is labelled according to the label of its nearest neighbor among the transported source data.

The Caltech-Office datasets A, C, D, W contain images of ten different objects coming from four different sources: Amazon, Caltech-256, DSLR and Webcam. We consider DeCAF6 features [Donahue et al., 2014], which are sparse 4096-dimensional vectors. For each pair of source/target datasets, both

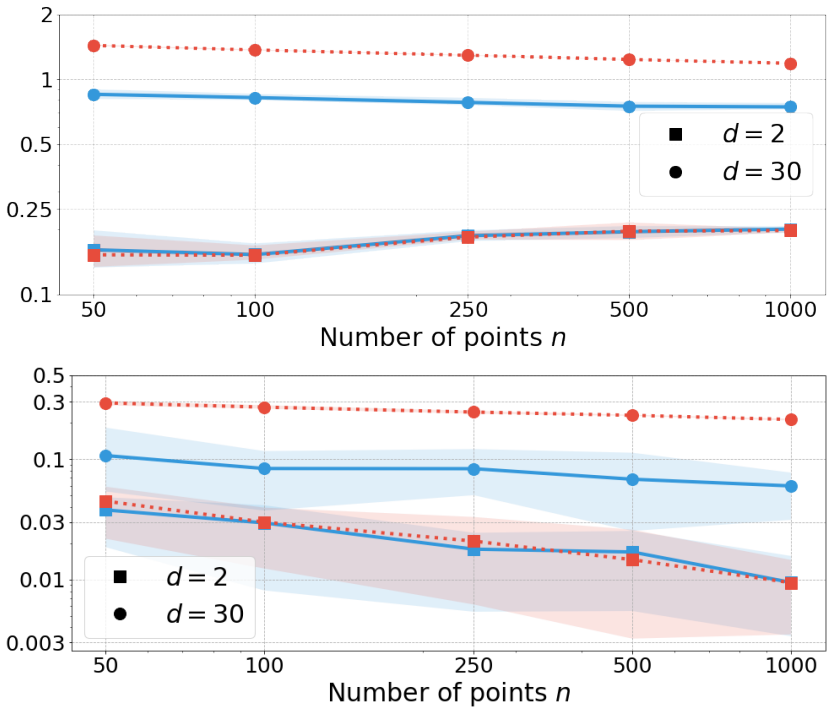


Figure 4.6: In the global (top) and local (bottom) regularity settings, we plot the estimation error of SSNB $|W_2(\mu,\nu) - \widehat{W}_2(\mu,\nu)|$ (blue line) and classical OT estimator $|W_2(\mu,\nu) - W_2(\hat{\mu}_n,\hat{\nu}_n)|$ (red dotted) depending on the number of points $n \in \{50, 100, 250, 500, 1000\}$, for dimension $d \in \{2, 30\}$. Here, the cluster ratios are taken constant equal to 0.5 (resp. 0.75) for the global (resp. local) regularity experiment. The number of Monte-Carlo samples is $N = 50$. The curves/shaded areas show the mean error/25%-75% percentiles over 20 data samples.

datasets are cut in half (train and test): hyperparameters are learnt on the train half and we report the mean ($\pm$ std) test accuracy over 10 random cuts, computed using a 1-Nearest Neighbour classifier on the transported source data. Results for OT, entropic OT, Group-Lasso and entropy regularized OT and SSNB are given in Table 4.1.

In order to compute the SSNB mapping, we *a)* quantize the source using k-means in each class with $k = 4$ centroids, *b)* learn the SSNB map between the 40 centroids and the target by solving (4.1), *c)* transport the source data by solving (4.2).

4.6.3 Color Transfer

Given a source and a target image, the goal of color transfer is to transform the colors of the source image so that it looks similar to the target image

Domains	OT	OT-IT	OT-GL	SSNB
A → C	74.0	81.0 ± 2.2	$\mathbf{86.6 \pm 1.0}$	83.9 ± 1.6
A → D	65.0	$\mathbf{84.3 \pm 5.2}$	$\mathbf{82.9 \pm 5.2}$	$\mathbf{79.2 \pm 3.7}$
A → W	65.0	$\mathbf{79.3 \pm 5.3}$	$\mathbf{81.6 \pm 3.1}$	$\mathbf{82.3 \pm 3.1}$
C → A	75.3	90.3 ± 1.3	$\mathbf{91.1 \pm 1.8}$	$\mathbf{91.9 \pm 1.1}$
C → D	65.6	$\mathbf{83.9 \pm 4.0}$	$\mathbf{85.8 \pm 2.3}$	81.1 ± 4.9
C → W	64.8	$\mathbf{77.1 \pm 3.9}$	$\mathbf{81.9 \pm 5.1}$	$\mathbf{79.3 \pm 2.8}$
D → A	69.6	88.3 ± 3.1	$\mathbf{90.5 \pm 1.9}$	$\mathbf{91.2 \pm 0.8}$
D → C	66.4	78.0 ± 2.9	$\mathbf{85.6 \pm 2.2}$	81.4 ± 1.9
D → W	84.7	$\mathbf{96.1 \pm 2.1}$	93.8 ± 2.1	$\mathbf{95.6 \pm 1.5}$
W → A	66.4	82.5 ± 4.3	$\mathbf{89.3 \pm 2.7}$	$\mathbf{89.8 \pm 2.9}$
W → C	62.5	74.8 ± 2.3	80.2 ± 2.4	$\mathbf{82.7 \pm 1.4}$
W → D	87.3	$\mathbf{97.5 \pm 2.1}$	$\mathbf{96.4 \pm 3.7}$	$\mathbf{95.9 \pm 3.2}$
Mean	70.6 ± 7.8	$\mathbf{84.4 \pm 7.0}$	$\mathbf{87.1 \pm 4.9}$	$\mathbf{86.2 \pm 6.0}$

Table 4.1: OT-IT: Entropy regularized OT. OT-GL: Entropy + Group Lasso regularized OT. Search intervals for OT-IT and OT-GL are $\varepsilon, \eta \in \{10^{-3}, \dots, 10^{3}\}$ with normalized cost, and for SSNB: $\ell \in \{0.2, 0.5, 0.7, 0.9\}$, $L \in \{0.3, 0.5, 0.7, 0.9, 1.3\}$. The best results are in bold.

color palette. Optimal transport has been used to carry out such a task, see *e.g.* [Bonneel et al., 2015, Ferradans et al., 2014, Rabin et al., 2014]. Each image is represented by a point cloud in the RGB color space identified with $[0, 1]^3$. The optimal transport plan π between the two point clouds give, up to a barycentric projection, a transfer color mapping.

It is natural to ask that similar colors are transferred to similar colors, and that different colors are transferred to different colors. These two demands translate into the smoothness and strong convexity of the Brenier potential from which derives the color transfer mapping. We therefore propose to compute a SSNB potential and map between the source and target distributions in the color space.

In order to make the computations tractable, we compute a k-means clustering with 30 clusters for each point cloud, and compute the SSNB potential using the two empirical measures on the centroids.

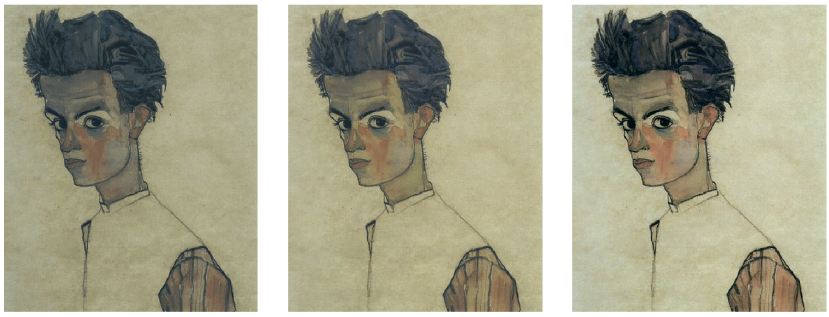
We then recompute a k-means clustering of the source point cloud with 1000 clusters. For each of the 1000 centroids, we compute its new color by solving QCQP (4.2). A pixel in the original image then sees its color changed according to the transformation of its nearest neighbor among the 1000 centroids.

In Figure 4.7, we show the color-transferred results using OT, or SSNB potentials for different values of parameters ℓ and L. Smaller values of L give more uniform colors, while larger values of ℓ give more contrast.

Conclusion

We have proposed in this work the first computational procedure to estimate optimal transport that incorporates smoothness and strongly convex (local) constraints on the Brenier potential, or, equivalently, that ensures that the optimal transport map has (local) distortion that is both upper and lower bounded. These assumptions are natural for several problems, both high and low dimensional, can be implemented practically and advance the current knowledge on handling the curse of dimensionality in optimal transport.

(a) Original Image (b) Target Image (c) Classical OT, $W \approx 0$



(d) $\ell = 0, L = 1$, $W \approx 1 \times 10^{-2}$ (e) $\ell = .5, L = 1$, $W \approx 1 \times 10^{-2}$ (f) $\ell = 1, L = 1$, $W \approx 2 \times 10^{-2}$

(g) $\ell = 0, L = 2$, $W \approx 4 \times 10^{-3}$ (h) $\ell = .5, L = 2$, $W \approx 5 \times 10^{-3}$ (i) $\ell = 1, L = 2$, $W \approx 2 \times 10^{-2}$

(j) $\ell = 0, L = 5$, $W \approx 2 \times 10^{-4}$ (k) $\ell = .5, L = 5$, $W \approx 1 \times 10^{-3}$ (l) $\ell = 1, L = 5$, $W \approx 2 \times 10^{-2}$

Figure 4.7: (a) Schiele's portrait. (b) Van Gogh's portrait. (c) Color transfer using classical OT. (d)-(l) Color transfer using SSNB map, for $\ell \in \{0, 0.5, 1\}$ and $L \in \{1, 2, 5\}$. The value W corresponds to the Wasserstein distance between the color distribution of the image and the color distribution of Van Gogh's portrait. The smaller W, the greater the fidelity to Van Gogh's portrait colors.

Conclusion

In this thesis, we have proposed variants of optimal transport distances that provide several advantages over classical Wasserstein distances. In part I, we showed how maximizing optimal transport with respect to the ground-cost function can help leverage the low-dimensional structures hidden in the measures (chapter 2) and how this very fact can be reinterpreted in terms regularization (chapter 3). In part II, we proposed a novel estimator for the Monge map which leverages the Brenier theorem and the regularity theory to define an optimal transport map on the whole space (chapter 4).

Follow-ups and perspectives to chapter 2

While the work in chapter 2 was partly motivated by the curse of dimensionality of optimal transport, we have not proved any improvement of the $\mathcal{O}(n^{-1/d})$ rate when using the projection robust Wasserstein (PRW) distance or the subspace robust Wasserstein (SRW) distance. Niles-Weed and Rigollet [2019] proved, for the PRW distance, a rate of $\mathcal{O}(n^{-1/k})$ when the measures only differ on a k-dimensional subspace. Since the SRW distance is strongly equivalent to the 2-Wasserstein distance, improved statistical bounds seem impossible to obtain in that case.

On the algorithmic side, we did not propose any algorithm for the PRW distance since it is the solution to a non-convex optimization problem. Lin et al. [2020], Huang et al. [2020b] recently proposed to use Riemannian optimization algorithms to compute the PRW distance.

On the application side, Huang et al. [2021] have used the PRW distance to define and compute robust barycenters of measures, while Wang et al. [2020], Masud et al. [2021] designed statistical tests based on the PRW distance. Alaya et al. [2020] propose to extend the PRW/SRW setting to the case where the two measures do not live on the same space.

Follow-ups and perspectives to chapter 3

In chapter 3, we showed how the link between maximizing and convexifying the optimal transport problem that we started to unveil in chapter 2 was actually a general fact. Chapter 3 ended on a deceiving note: the optimal adversarial

ground-cost function often happens to be separable, hence does not provide any interesting geometry (or dissimilarity measure) on the ground space.

Looking at the big picture, the adversarial choice of a ground-cost function is a robust way to choose a ground-cost function that is close, in some sense, to a fixed function that acts a bit like a prior. What if, instead of *optimizing* over the ground-cost function, we chose it as a small *random* perturbation of this very prior? Then the optimal transport value and plan become random as well, and understanding their randomness could be of interest. For example, in the discrete case:

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

where $x_1, \dots, x_n, y_1, \dots, y_m \in \mathbb{R}^d$ and a, b are probability weights. A simple choice of randomizing the ground-cost function in that case is to randomly perturbate the points with i.i.d Gaussian random vectors, to obtain random measures

$$\widetilde{\mu}_\sigma = \sum_{i=1}^{n} a_i \delta_{x_i + \sigma \xi_i}, \qquad \widetilde{\nu}_\sigma = \sum_{j=1}^{m} b_j \delta_{y_j + \sigma \xi'_j},$$

where $\xi_1, \dots, \xi_n, \xi'_1, \dots, \xi'_m \sim \mathcal{N}(0, \mathrm{Id})$ are independent, and $\sigma \geq 0$ controls the variance of the perturbations. Then almost surely, the optimal transport plan $P_\star(\widetilde{\mu}_\sigma, \widetilde{\nu}_\sigma)$ between $\widetilde{\mu}_\sigma$ and $\widetilde{\nu}_\sigma$ is unique and we can define its expected value:

$$\overline{P_\sigma} \stackrel{\text{def}}{=} \mathbb{E}\left[P_\star(\widetilde{\mu}_\sigma, \widetilde{\nu}_\sigma)\right] \in \mathbb{R}_+^{n \times m}.$$

Although I give no proof, the following intuitive facts should hold:

- $\lim_{\sigma \to 0} \overline{P_\sigma}$ exists and is an optimal transport plan between μ and ν;
- $\lim_{\sigma \to +\infty} \overline{P_\sigma} = a \otimes b$ is the independent coupling.

Then we could define the associated transport cost associated with $\overline{P_\sigma}$:

$$W^\sigma(\mu, \nu) \stackrel{\text{def}}{=} \sum_{i,j} \left(\overline{P_\sigma}\right)_{ij} c(x_i, y_j).$$

Using the intuitive facts above, we can then expect that:

$$\mathscr{T}_c(\mu, \nu) \xleftarrow[\sigma \to 0]{} W^\sigma(\mu, \nu) \xrightarrow[\sigma \to \infty]{} \sum_{i,j} a_i b_j c(x_i, y_j).$$

This is exactly what is verified by the Sinkhorn divergences [Genevay et al., 2018, Theorem 1]. Although I conjecture W^σ is not a Sinkhorn divergence, the parallel between them make W^σ an object of potential interest: is it related to some form of regularization? How can we define W^σ for non-discrete measures? What kind of properties does it exhibit? Does it suffer from a curse of dimensionality?

As we have seen, in the discrete case, randomizing the ground-cost function is equivalent to randomizing the position of the points, hence the supports of the measures. Could we randomize not only their supports, but the measures as a whole? For example, adding Gaussian noise to the points is just performing a heat kernel on the support of the measures. To randomize the weights, we could thus perform the heat kernel over the probability simplex. How can such random perturbations of the measures *in the space of measures* be informative in terms of optimal transport?

The third option consists of randomizing only the weights and fixing a ground-cost function. In that setting, could the associated random optimal transport cost (or plan) provide insightful information on the ground-cost function? For example, if the measures are distributions over the vertices of a graph, and the ground-cost function is the shortest-path distance over the graph, could random perturbations of the weights provide some kind of information on the structure of the graph? In other words, can optimally moving some mass over a graph inform about its structure?

Follow-ups and perspectives to chapter 4

The smooth and strongly convex nearest Brenier (SSNB) potentials defined in chapter 4 seem to exhibit nice statistical features, but we were unable to prove anything quite like them. Future works should try to understand the statistical properties of the SSNB estimators. On the algorithmic side, approximating the SSNB potentials is time-consuming, mainly because of the convex QCQP that has to be solved. One key question to make SSNB potentials usable at scale is therefore the one of accelerating their computation, even at the cost of a relaxation of the problem. Recently, Amos et al. [2017] defined input convex neural networks (ICNN), *i.e.* neural networks that are convex functions of their inputs. Such neural networks have been used by Korotin et al. [2019], Makkuva et al. [2020] to estimate Monge map by leveraging the convexity in some dual formulation of the 2-Wasserstein distance. Recently, Huang et al. [2020a] proposed to use ICNNs to estimate the convex Brenier potentials in the setting of normalizing flows. It would thus be natural to use ICNNs to approximate SSNB potentials: while the strong convexity of SSNB potentials would be easy to constrain as noticed in [Huang et al., 2020a, section 3.1], constraining their smoothness appears more intricate. One possibility would be to consider the Moreau-Yosida regularization [Moreau, 1965, Yosida, 1965] of the Brenier potential.

More generally, I believe that encoding the optimality of the Monge map by the convexity of the associated Brenier potential is still underemployed in machine learning, since regularizing the Brenier potential can be directly interpreted in terms of geometry, while the classical regularization on the transport plan is more difficult to interpret, even when using the results of chapter 3.

The Brenier theorem also gives a condition for a transport map to be optimal which does not involve the computation of a transport cost, which comes in handy when such computations are intractable. In some fields, such as statistical physics or in the study of point processes, measures of infinite mass appear (but they give finite mass to compact sets). How to define optimal transport maps in this setting? Let us consider the following example: $\mu = \sum_{x\in\mathbb{Z}^d} \delta_x$ and $\nu = \sum_{x\in\mathbb{Z}^d} \delta_{x+u}$ where $\|u\| < \frac{1}{2}$. Then any map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\sharp\mu = \nu$ induces an infinite displacement cost $C(T) = \sum_{x\in\mathbb{Z}^d} \|x - T(x)\|^2 = +\infty$, and in that sense, no map (or all the maps) is optimal. But, in some sense, it seems intuitive to say that the map $T_\star : x \mapsto x + u$ should be an optimal transport map, because any permutation of this map "locally increases the displacement cost". To formalize this suggestion, a first idea would be to restrict the measures to increasing compact domains $\Lambda_N = [-N, N]^d$, so that the displacement cost $\sum_{x\in\Lambda_N} \|x - T(x)\|^2$ is finite and can thus be minimized, providing a sequence (T_N) of compactly supported optimal maps. It would then remains to construct a map T defined on the whole space out of this sequence (T_N), which is no trivial matter. A much simpler construction is the following one based on the Brenier theorem. We will say that T is *compatible* with the transport of μ to ν if it verifies the following conditions:

- $T_\sharp\mu = \nu$,
- there exists a convex function $f : \mathbb{R}^d \to \mathbb{R}$ such that $T = \nabla f$.

If μ and ν were probability measures, being a *compatible* transport map would correspond to being an *optimal* transport map. But since μ and ν have infinite masses, this is not sufficient: think of the map $T : x + a\mathbf{1} + u$ where $\mathbf{1}$ is the vector of all ones, which is compatible with the transport of μ to ν for any $a \in \mathbb{Z}$. For any compatible transport map T sending μ to ν, we define its local cost on Λ_N:

$$C_{\Lambda_N}(T) \stackrel{\text{def}}{=} \frac{1}{\mu(\Lambda_N)} \int_{\Lambda_N} \|x - T(x)\|^2 \, d\mu(x) = W_2^2 \left(\frac{\mu_{|\Lambda_N}}{\mu(\Lambda_N)}, T_\sharp \left[\frac{\mu_{|\Lambda_N}}{\mu(\Lambda_N)} \right] \right),$$

from which we can define its global transport cost:

$$C(T) \stackrel{\text{def}}{=} \liminf_{N\to\infty} C_{\Lambda_N}(T).$$

Finally, an optimal transport map sending μ to ν should be a minimizer of:

$$\inf \{C(T), T \text{ is compatible with the transport}\}$$

and the value of this infimum is the associated transportation cost (it is in fact the cost needed to transport μ to ν "by unit of mass").

Bibliography

Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

Pankaj K Agarwal, Jeff M Phillips, and Bardia Sadri. Lipschitz unimodal and isotonic regression on paths and trees. In *Latin American Symposium on Theoretical Informatics*, pages 384–396. Springer, 2010.

Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. Network flows. 1988.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. Unsupervised hyper-alignment for multilingual word embeddings. In *International Conference on Learning Representations*, 2019.

Mokhtar Z. Alaya, Maxime Bérar, Gilles Gasso, and Alain Rakotomamonjy. Theoretical guarantees for bridging metric measure embedding and optimal transport, 2020.

Helmut Alt and Leonidas J Guibas. Discrete geometric shapes: Matching, interpolation, and approximation. In *Handbook of computational geometry*, pages 121–153. Elsevier, 2000.

Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Weed. Approximating the quadratic transportation metric in near-linear time. *arXiv preprint arXiv:1810.10046*, 2018a.

Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Weed. Massively scalable sinkhorn distances via the nyström method. *stat*, 1050:12, 2018b.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, 2006.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:214–223, 2017.

Aharon Ben-Tal and Arkadi Nemirovski. Robust convex optimization. *Mathematics of operations research*, 23(4):769–805, 1998.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

Dimitri P Bertsekas. *Control of uncertain systems with a set-membership description of the uncertainty.* PhD thesis, Massachusetts Institute of Technology, 1971.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures-wasserstein distance between positive definite matrices. *Expositiones Mathematicae, to appear*, 2018.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 880–889, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL http://proceedings.mlr.press/v84/blondel18a.html.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71:1–71:10, 2016.

James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47. Springer, 1986.

Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(19):805–808, 1987.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Donald Bures. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

Luis A Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

Luis A Caffarelli. Monotonicity properties of optimal transportation and the fkg and related inequalities. *Communications in Mathematical Physics*, 214 (3):547–563, 2000.

Luis A Caffarelli, Sergey A Kochengin, and Vladimir I Oliker. Problem of reflector design with given far-field scattering data. In *Monge Ampère equation: applications to geometry and optimization*, volume 226, page 13, 1999.

Guillermo D Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. *arXiv preprint arXiv:1209.1077*, 2012.

Scott Cohen and Leonidas Guibas. The earth mover's distance under transformation sets. In *Proceedings of the Seventh IEEE International Conference on Computer vision*, volume 2, pages 1076–1083. IEEE, 1999.

Andrew Cotter, Joseph Keshet, and Nathan Srebro. Explicit approximations of the gaussian kernel. *arXiv preprint arXiv:1109.4603*, 2011.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

Marco Cuturi and David Avis. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.

Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative modeling using the sliced wasserstein distance. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander Schwing. Max-sliced wasserstein distance and its use for gans. *arXiv preprint arXiv:1904.05877*, 2019.

Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145 (1-2):451–482, 2014.

Richard M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Arnaud Dupuy and Alfred Galichon. Personality traits and the marriage market. *Journal of Political Economy*, 122(6):1271–1319, 2014.

Arnaud Dupuy, Alfred Galichon, and Yifei Sun. Estimating matching affinity matrix under low-rank constraints. *Arxiv:1612.09585*, 2016.

Montacer Essid and Justin Solomon. Quadratically-regularized optimal transport on graphs. *arXiv preprint arXiv:1704.08200*, 2017.

Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/feydy19a.html.

Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010.

Alessio Figalli. *The Monge–Ampère equation and its applications*. 2017.

Rémi Flamary, Nicolas Courty, Alain Rakotomamonjy, and Devis Tuia. Optimal transport with laplacian regularization. In *NIPS 2014, Workshop on Optimal Transport and Machine Learning*, 2014.

Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

Rémi Flamary, Karim Lounici, and André Ferrari. Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*, 2019.

Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. 2019.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

Alfred Galichon and Bernard Salanié. Cupid's invisible hand: Social surplus and identification in matching models. *Available at SSRN 1804623*, 2015.

Matthias Gelbrich. On a formula for the l^2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

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

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–

1583. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/genevay19a.html.

Gauthier Gidel, Tony Jebara, and Simon Lacoste-Julien. Frank-Wolfe algorithms for saddle point problems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Clark R Givens, Rae Michael Shortt, et al. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2): 231–240, 1984.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. 2019.

Janice H Hammond. *Solving asymmetric variational inequality problems and systems of equations with generalized nonlinear programming algorithms*. PhD thesis, Massachusetts Institute of Technology, 1984.

Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative RNNs. In *International Conference on Machine Learning*, pages 2417–2426, 2016.

Chin-Wei Huang, Ricky TQ Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. *arXiv preprint arXiv:2012.05942*, 2020a.

Minhui Huang, Shiqian Ma, and Lifeng Lai. A riemannian block coordinate descent method for computing the projection robust wasserstein distance. *arXiv preprint arXiv:2012.05199*, 2020b.

Minhui Huang, Shiqian Ma, and Lifeng Lai. Projection robust wasserstein barycenter. *arXiv preprint arXiv:2102.03390*, 2021.

Geert-Jan Huizing, Gabriel Peyré, and Laura Cantini. Optimal transport improves cell-cell similarity inference in single-cell omics data. *bioRxiv*, 2021. doi: 10.1101/2021.03.19.436159. URL https://www.biorxiv.org/content/early/2021/03/20/2021.03.19.436159.

Jan-Christian Hütter and Philippe Rigollet. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

Soheil Kolouri, Charles E Martin, and Gustavo K Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo K Rohde. Generalized sliced wasserstein distances. *arXiv preprint arXiv:1902.00434*, 2019.

Jonathan Korman and Robert McCann. Optimal transportation with capacity constraints. *Transactions of the American Mathematical Society*, 367(3): 1501–1521, 2015.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

Theo Lacombe. *Statistics for Topological Descriptors using optimal transport*. Theses, Institut Polytechnique de Paris, September 2020. URL https://hal.archives-ouvertes.fr/tel-02979251.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael I Jordan. Projection robust wasserstein distance and riemannian optimization. *arXiv preprint arXiv:2006.07458*, 2020.

Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):840–853, 2007.

Dirk A Lorenz, Paul Manns, and Christian Meyer. Quadratically regularized optimal transport. *arXiv preprint arXiv:1903.01112*, 2019.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

Shoaib Bin Masud, Boyang Lyu, and Shuchin Aeron. Soft and subspace robust multivariate rank tests based on entropy regularized optimal transport. *arXiv preprint arXiv:2103.08811*, 2021.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, pages 4543–4553, 2019.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the wasserstein space of elliptical distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10258–10269. Curran Associates, Inc., 2018.

Boris Muzellec, Richard Nock, Giorgio Patrini, and Frank Nielsen. Tsallis regularized optimal transport and ecological inference. In *AAAI*, pages 2387–2393, 2017.

Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

Michael L Overton and Robert S Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62(1-3):321–357, 1993.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019.

François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages

1222–1232. PMLR, 26–28 Aug 2020. URL http://proceedings.mlr.press/v108/paty20a.html.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5072–5081, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/paty19a.html.

François-Pierre Paty and Marco Cuturi. Regularized optimal transport is ground cost adversarial. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7532–7542. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/paty20a.html.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *IEEE 12th International Conference on Computer Vision*, pages 460–467, 2009.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4197–4205. Curran Associates, Inc., 2016.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073.

Guido Philippis. *Regularity of optimal transport maps and applications*, volume 17. Springer Science & Business Media, 2013.

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

Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 256–269. Springer, 2015.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference*

on Scale Space and Variational Methods in Computer Vision, pages 435–446. Springer, 2011.

Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4852–4856. IEEE, 2014.

Alfréd Rényi. On projections of probability distributions. *Acta Mathematica Academiae Scientiarum Hungarica*, 3(3):131–142, 1952.

Philippe Rigollet and Jonathan Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11-12):1228–1235, 2018.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkQkBnJAb.

Filippo Santambrogio. *Optimal transport for applied mathematicians*. Birkhauser, 2015.

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.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *Proceedings of ICLR 2018*, 2018.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statististics*, 35:876–879, 1964.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66:1–66:11, 2015.

Adrien B Taylor. *Convex interpolation and performance estimation of first-order methods for convex optimization.* PhD thesis, 2017.

Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017.

Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003. ISBN 9780821833124.

Cedric Villani. *Optimal Transport: Old and New*, volume 338. Springer Verlag, 2009.

Jie Wang, Rui Gao, and Yao Xie. Two-sample test using projected wasserstein distance: Breaking the curse of dimensionality. *arXiv preprint arXiv:2010.11970*, 2020.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25 (4A):2620–2648, 2019.

L Yeganova and WJ Wilbur. Isotonic regression under lipschitz constraint. *Journal of optimization theory and applications*, 141(2):429–443, 2009.

Kôsaku Yosida. Functional analysis, volume 123 of. *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, 1965.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Transport optimal en grande dimension : obtention de régularité et de robustesse au moyen de la convexité et des projections

Mots clés : Transport optimal, Apprentissage statistique, Machine learning

Résumé : Le transport optimal (TO) a récemment gagné en popularité en apprentissage automatique pour comparer des mesures de probabilité. Contrairement aux dissimilarités plus classiques pour les distributions de probabilité, les distances de TO (ou distances de Wasserstein) permettent de comparer des distributions dont les supports sont disjoints en prenant en compte la géométrie de l'espace sous-jacent. Cet avantage est cependant entravé par le fait que ces distances sont généralement calculées en résolvant un programme linéaire, ce qui pose, lorsque l'espace sous-jacent est de grande dimension, des défis statistiques bien documentés et auxquels on se réfère communément sous le nom de "fléau" de la dimension. Trouver de nouvelles méthodologies qui puissent atténuer ce problème est donc un enjeu crucial si l'on veut que les algorithmes fondés sur le TO puissent fonctionner en pratique. Au-delà de cet aspect purement métrique, un autre intérêt de la théorie du TO réside en ce qu'elle fournit des outils mathématiques pour étudier des cartes qui peuvent transformer une mesure en une autre. Estimer de telles transformations qui soient à la fois optimales et qui puissent être généralisées en dehors des simples données, est un problème ouvert. Dans cette thèse, nous proposons un nouveau cadre d'estimation pour calculer des variantes des distances de Wasserstein. Le but est d'amoindrir les effets de la haute dimension en tirant partie des structures de faible dimension cachées dans les distributions. Cela peut se faire en projetant les mesures sur un sous-espace choisi de telle sorte à maximiser la distance de Wasserstein entre leurs projections. Outre cette nouvelle méthodologie, nous montrons que ce cadre d'étude s'inscrit plus largement dans un lien entre la régularisation des distances de Wasserstein et la robustesse. Dans la contribution suivante, nous partons du même problème d'estimation du TO en grande dimension, mais adoptons une perspective différente : plutôt que de modifier la fonction de coût, nous revenons au point de vue plus fondamental de Monge et proposons d'utiliser le théorème de Brenier et la théorie de la régularité de Caffarelli pour définir une nouvelle procédure d'estimation des cartes de transport lipschitziennes qui soient le gradient d'une fonction fortement convexe.

Title : Optimal Transport in High Dimension: Obtaining Regularity and Robustness using Convexity and Projections

Keywords : Optimal transport, Statistical Learning, Machine learning

Abstract : Optimal transport (OT) has recently gained popularity in machine learning as a way to compare probability distributions. Unlike more classical dissimilarities for probability measures, OT distances (or Wasserstein distances) can deal with distributions of disjoint supports by taking into account the geometry of the underlying ground space. This strength is, however, hampered by the fact that these distances are usually computed by solving a linear program, resulting, when this ground space is high-dimensional, in well documented statistical challenges, usually referred to as the "curse" of dimensionality. Finding new methodologies that can mitigate this issue is therefore crucial if one wants OT-based algorithms to perform well on real data. Beyond this purely metric aspect, another appealing feature of OT theory is that it provides mathematical tools to study maps that are able to morph a measure into another. Estimating such maps, that are both optimal and able to generalize outside the data, is an open problem. In this thesis, we propose a new estimation framework to compute proxies to the Wasserstein distance. That framework aims at handling high-dimensionality by taking advantage of the low-dimensional structures hidden in the distributions. This can be achieved by projecting the measures onto a subspace chosen so as to maximize the Wasserstein distance between their projections. In addition to this novel methodology, we show that this framework falls into a broader connection between regularization when computing Wasserstein distances and adversarial robustness. In the next contribution, we start from the same problem, estimation of OT in high dimensions, but adopt a different perspective: rather than changing the ground cost, we go back to the more fundamental Monge perspective on OT and use the Brenier theorem and Caffarelli's regularity theory to propose a new estimation procedure to characterize maps that are Lipschitz and gradients of strongly convex functions.

Institut Polytechnique de Paris
91120 Palaiseau, France