Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport

François-Pierre Paty
CREST, ENSAE,
Institut Polytechnique de Paris

Alexandre d'Aspremont
CNRS, ENS,
PSL Research University

Marco Cuturi
Google Brain,
CREST, ENSAE

Abstract

Estimating Wasserstein distances between two high-dimensional densities suffers from the curse of dimensionality: one needs an exponential (wrt dimension) number of samples to ensure that the distance between two empirical measures is comparable to the distance between the original densities. Therefore, optimal transport (OT) can only be used in machine learning if it is substantially regularized. On the other hand, 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. We propose in this work to draw inspiration from this theory and use regularity as a regularization tool. We give algorithms operating on two discrete measures that can recover nearly optimal transport maps with small distortion, or equivalently, nearly optimal Brenier potentials that are strongly convex and smooth. The problem boils down to solving alternatively a convex QCQP and a discrete OT problem, granting access to the values and gradients of the Brenier potential not only on sampled points, but also out of sample at the cost of solving a simpler QCQP for each evaluation. We propose algorithms to estimate and evaluate transport maps with desired regularity properties, benchmark their statistical performance, apply them to domain adaptation and visualize their action on a color transfer task.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

1 INTRODUCTION

Optimal transport (OT) has found practical applications in areas as diverse as supervised machine learning (Frogner et al., 2015; Abadeh et al., 2015; Courty et al., 2016), graphics (Solomon et al., 2015; Bonneel et al., 2016), generative models (Arjovsky et al., 2017; Salimans et al., 2018), NLP (Grave et al., 2019; Alaux et al., 2019), biology (Hashimoto et al., 2016; Schiebinger et al., 2019) or imaging (Rabin & Papadakis, 2015; Cuturi & Peyré, 2016). OT theory is useful in these 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. There is, however, an important difference between the OT definitions introduced in celebrated textbooks by Villani (2003; 2009) and Santambrogio (2015), and those used in the works cited above. In all of these 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 & 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 & 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.

2 REGULARITY IN OPTIMAL TRANSPORT

For $d \in \mathbb{N}$, we write $[\![d]\!] = \{1, ..., d\}$ and $\mathscr{L}^d$ for the Lebesgue measure in $\mathbb{R}^d$. We write $\mathscr{P}_2(\mathbb{R}^d)$ for the set of Borel probability measures with finite second-order moment.

Wasserstein distances, Kantorovich and Monge Formulations. For two probability measures $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, we write $\Pi(\mu, \nu)$ for the set of couplings

$$\Pi(\mu,\nu) = \{\pi \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d) \text{ s.t. } \forall A, B \subset \mathbb{R}^d \text{ Borel}, \\ \pi(A \times \mathbb{R}^d) = \mu(A), \pi(\mathbb{R}^d \times B) = \nu(B)\},$$

and define their 2-Wasserstein distance as the solution of the Kantorovich problem (Villani, 2009, §6):

$$W_2(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|_2^2 \, d\pi(x,y)\right)^{1/2}.$$

For Borel sets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, Borel map $T : \mathcal{X} \to \mathcal{Y}$ and $\mu \in \mathscr{P}(\mathcal{X})$, we denote by $T_\sharp \mu \in \mathscr{P}(\mathcal{Y})$ the push-forward of μ under T, *i.e.* the measure such that for any $A \subset \mathcal{Y}$, $T_\sharp \mu(A) = \mu\left(T^{-1}(A)\right)$. The Monge (1781) formulation of OT consists in considering maps such that $T_\sharp \mu = \nu$, instead of couplings. Both formulations are equal when feasible maps exist, namely

$$W_2(\mu,\nu) = \left(\inf_{T : T_\sharp \mu = \nu} \int \|x - T(x)\|^2 \, d\mu(x)\right)^{1/2}.$$

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.

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 & Guibas, 1999; Alt & 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 push-foward of μ and ν to define:

Definition 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 1. *The existence of an SSNB potential is proved in the supplementary material. 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 1.*

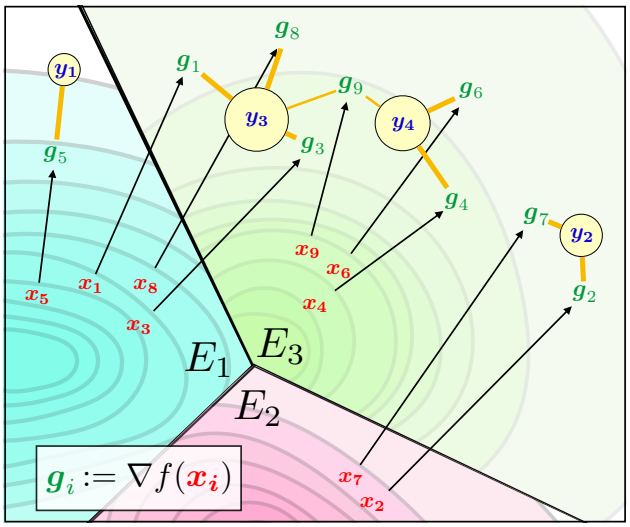


Figure 1: Points x_i mapped onto points $g_i := \nabla f(x_i)$ for a function f that is locally smooth strongly convex. SSNB potentials are such that the measure of endpoints g_i is as close as possible (in W_2 sense) to the measure supported on the y_j. Here this would be the sum of the squares of the length of these orange sticks.

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 non-negative 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 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 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:*

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

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

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

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

We refer to the supplementary material for the proof.

We provide algorithms to compute a SSNB potential in dimension $d\ge 2$ when μ,ν are discrete measures. In order to solve Problem (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 2 below) of μ as an initialization for the points z.

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 2 (Barycentric Projection)**.** *Let $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$, 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 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 1. *Let $\mu,\nu\in\mathcal{P}_2(\mathbb{R})$ and $0\le\ell\le L$. Suppose $\mu\ll\mathcal{L}^1$, 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).

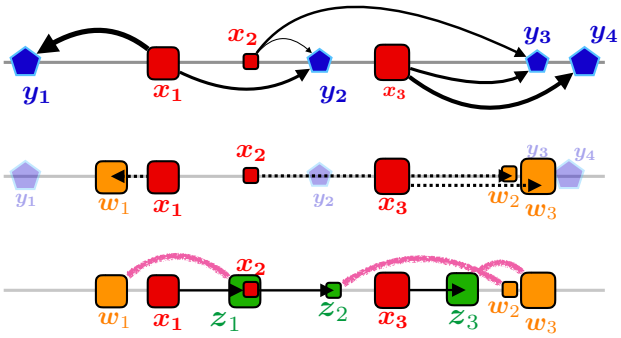


Figure 2: Top: optimal transport between two discrete measures μ,ν. Middle: the barycentric projection $\mathbf{w}$ of points $\mathbf{x}$ is displayed and corresponds to a Monge map (no mass splitting). Considering here for instance $\ell=0.5$ and $L=1$, the map that associates x_i to w_i is not 1-Lipschitz 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 (3).

Discrete Computations. Suppose $\mu=\sum_{i=1}^n a_i\delta_{x_i}$ is discrete with $x_1\le\dots\le x_n$, 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 $\overline{\pi}(x_i) = \frac{1}{a_i}\int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t)\,dt$ (proof in the supplementary material).

If ν is also discrete, with weights $\mathbf{b}=(b_1,\dots,b_m)$ and sorted support $\mathbf{y}=(y_1,\dots,y_m)\in\mathbb{R}^m$, 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 & 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 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. 2):

$$\min_{z \in \mathbb{R}^n} \sum_{i=1}^{n} a_i (z_i - w_i)^2 \tag{3}$$
$$\text{s.t. } \forall k \leq K, \forall i, i+1 \in I_k,$$
$$\ell(x_{i+1} - x_i) \leq z_{i+1} - z_i \leq L(x_{i+1} - x_i).$$

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 & 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 (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.

5 ESTIMATION OF 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 $x_1, \ldots, x_n \sim \mu$ and $y_1, \ldots, y_n \sim \nu$ 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.

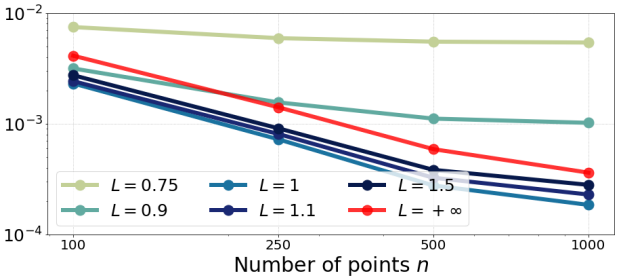


Figure 3: Take measures $\mu = \nu$ be the uniform measure over $[0, 1]$. For several n, we consider $\hat{\mu}_n, \hat{\nu}_n$ empirical measures over n iid samples from μ, ν, from which we compute a SSNB potential $\hat{f}_n$ with different values L, and $\ell = \min\{1, L\}$ (and $\ell = 0$ if $L = \infty$). 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 $\hat{\zeta}_n = \hat{f}_{n\sharp}\mu_n$. 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 $\hat{\zeta}_n = \hat{\nu}_n$.

Let $\hat{f}_n$ be a SSNB potential between $\hat{\mu}_n$ and $\hat{\nu}_n$ with $\mathcal{E} = \{\mathbb{R}^d\}$. Then for $x \in \mathrm{supp}\,\mu$, a natural estimator of $\nabla f_\star(x)$ is given by a solution $\nabla \hat{f}_n(x)$ of (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 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 1 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 (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 (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 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:*

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

The study of the theoretical rate of convergence of this estimator is beyond the scope of this work. Recent results (Hütter & 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 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 3.

6 EXPERIMENTS

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

6.1 Numerical Estimation of Wasserstein Distances and Monge Maps

In this experiment, we consider two different settings:
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.
Local Regularity: μ is the uniform measure over the unit ball in $\mathbb{R}^d$, and $\nu = T_\sharp\mu$ where $T(x_1,\dots,x_d) = (x_1 + 2\,\mathrm{sign}(x_1), x_2, \dots, x_d)$. As can be seen in Figure 4 (bottom), 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 $x_1,\dots,x_n \sim \mu$ and $y_1,\dots,y_n \sim \nu$ 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 $\mathcal{E} = \{E_1,\dots,E_K\}$ by running k-means with K clusters on data $x_1,\dots,x_n$. In both settings, we run the algorithms with $\hat{\ell} = 0.6$ and $\hat{L} = 1.4$. We give experimental results on the statistical performance of our SSNB estimator, computed using Monte-Carlo integration, compared to the classical optimal transport estimator $W_2(\hat{\mu}_n, \hat{\nu}_n)$.

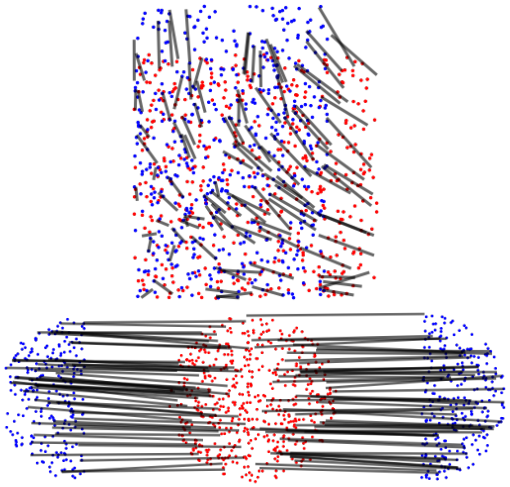


Figure 4: In the global (top) and local (bottom) regularity settings with $d = 2$, we plot $n = 500$ i.i.d samples from μ (red) and ν (blue). Black lines show the displacements $x \mapsto \nabla\hat{f}_n(x)$ for some new points $x \in \mathrm{supp}(\mu) \setminus \mathrm{supp}(\hat{\mu}_n)$, computed by solving QCQP (2).

This performance depends on three parameters: the number n of points, the dimension d and the number of clusters K.

In Figure 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 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 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 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 seem 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 that the classical OT estimator computed with a large number of points.

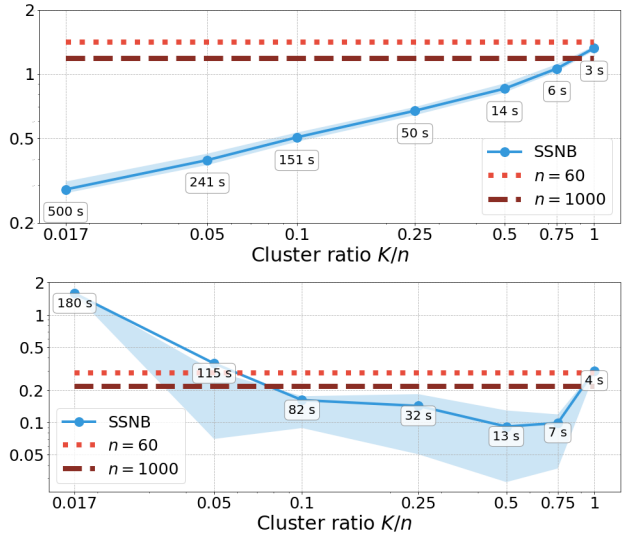


Figure 5: In the global (top) and local (bottom) regularity settings, we plot (on a log-log scale) the estimation error $|W_2(\mu,\nu) - \widehat{W}_2(\mu,\nu)|$ (blue line) depending on the cluster ratio K/n, with fixed number of points $n = 60$, dimension $d = 30$ and Monte-Carlo samples $N = 50$. 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.

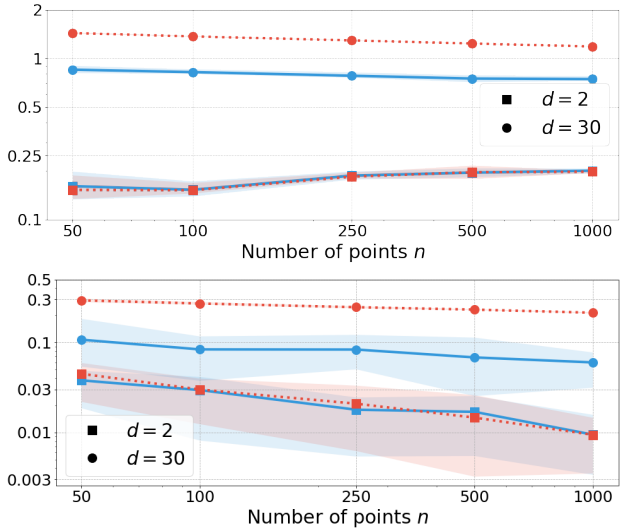


Figure 6: In the global (top) and local (bottom) regularity settings, we plot the estimation error of SSNB $|W_2(\mu,\nu) - \widehat{W}_2(\mu,\nu)|$ (blue line) and classical OT estimator $|W_2(\mu,\nu) - W_2(\hat{\mu}_n,\hat{\nu}_n)|$ (red dotted) depending on the number of points $n \in \{50, 100, 250, 500, 1000\}$, 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.

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 datasets are cut in half (train and test): hyperparameters are learnt on the train half and we report the mean (± 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 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 (1), *c)* transport the source data by solving (2).

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

Table 1: OT-IT: Entropy regularized OT. OT-GL: Entropy + Group Lasso regularized OT. Search intervals for OT-IT and OT-GL are $\epsilon, \eta \in \{10^{-3}, \dots, 10^{3}\}$ 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.

6.3 Color Transfer

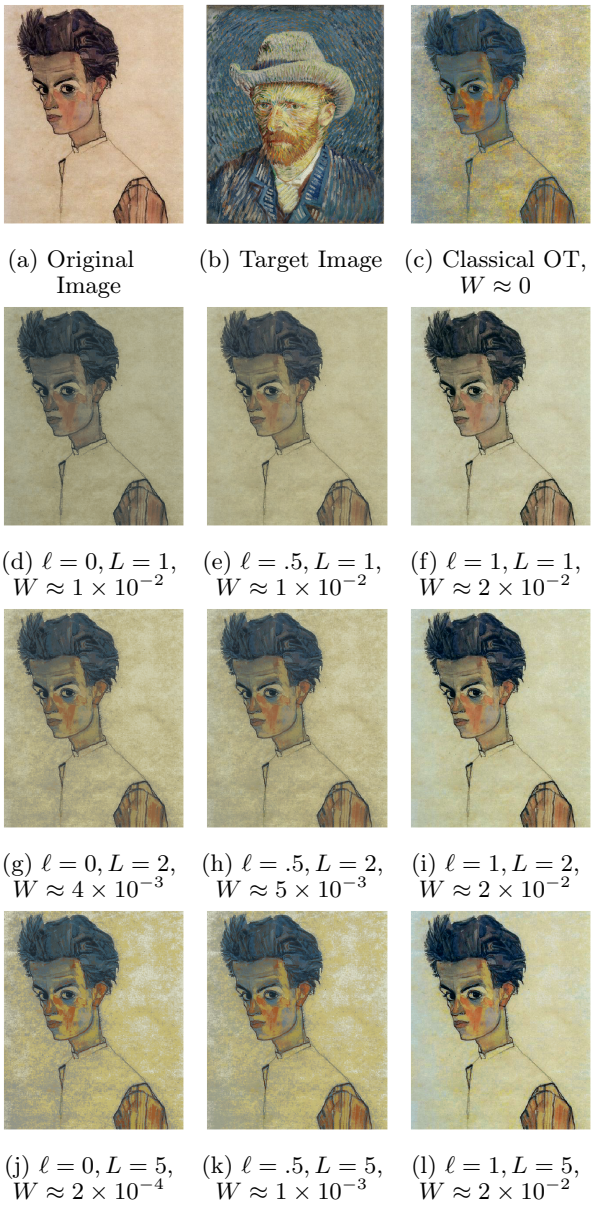
Given a source and a target image, the goal of color transfer is to transform the colors of the source image so that it looks similar to the target image color palette. Optimal transport has been used to carry out such a task, see *e.g.* (Bonneel et al., 2015; Ferradans et al., 2014; Rabin et al., 2014). 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 (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 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.



(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 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.

7 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.

Acknowledgments. We thank J. Altschuler for his remark that the QCQP (1) considered in Theorem 1 is in fact convex (unlike in the optimization literature where it is not) and can therefore be exactly solved and not approximated through an SDP.

F.P.P and M.C. acknowledge the support of a "Chaire d'excellence de l'IDEX Paris Saclay". A.A. is at the département d'informatique de l'ENS, École normale supérieure, UMR CNRS 8548, PSL Research University, 75005 Paris, France, and INRIA Sierra project-team. AA would like to acknowledge support from the *ML and Optimisation* joint research initiative with the *fonds*

AXA pour la recherche and Kamet Ventures, a Google focused award, as well as funding by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pp. 1576–1584, 2015.

Agarwal, P. K., Phillips, J. M., and Sadri, B. Lipschitz unimodal and isotonic regression on paths and trees. In *Latin American Symposium on Theoretical Informatics*, pp. 384–396. Springer, 2010.

Alaux, J., Grave, E., Cuturi, M., and Joulin, A. Unsupervised hyper-alignment for multilingual word embeddings. In *International Conference on Learning Representations*, 2019.

Alt, H. and Guibas, L. J. Discrete geometric shapes: Matching, interpolation, and approximation. In *Handbook of computational geometry*, pp. 121–153. Elsevier, 2000.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, 2006.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:214–223, 2017.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Bonneel, N., Peyré, G., and Cuturi, M. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71:1–71:10, 2016.

Brenier, Y. Décomposition polaire et réarrangement monotone des champs de vecteurs. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(19):805–808, 1987.

Caffarelli, L. A. Monotonicity properties of optimal transportation and the fkg and related inequalities. *Communications in Mathematical Physics*, 214(3): 547–563, 2000.

Caffarelli, L. A., Kochengin, S. A., and Oliker, V. I. Problem of reflector design with given far-field scattering data. In *Monge Ampère equation: applications to geometry and optimization*, volume 226, pp. 13, 1999.

Cohen, S. and Guibas, L. The earth mover's distance under transformation sets. In *Proceedings of the Seventh IEEE International Conference on Computer vision*, volume 2, pp. 1076–1083. IEEE, 1999.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Cuturi, M. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pp. 2292–2300, 2013.

Cuturi, M. and Peyré, G. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pp. 647–655, 2014.

Drori, Y. and Teboulle, M. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

Dudley, R. M. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40 (1):40–50, 1969.

Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Figalli, A. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195 (2):533–560, 2010.

Figalli, A. *The Monge–Ampère equation and its applications*. 2017.

Flamary, R., Courty, N., Rakotomamonjy, A., and Tuia, D. Optimal transport with laplacian regularization. In *NIPS 2014, Workshop on Optimal Transport and Machine Learning*, 2014.

Flamary, R., Lounici, K., and Ferrari, A. Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*, 2019.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2053–2061, 2015.

Grave, E., Joulin, A., and Berthet, Q. Unsupervised alignment of embeddings with wasserstein procrustes. 2019.

Hashimoto, T., Gifford, D., and Jaakkola, T. Learning population-level diffusions with generative RNNs. In *International Conference on Machine Learning*, pp. 2417–2426, 2016.

Hütter, J.-C. and Rigollet, P. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

Kakade, S. M., Kanade, V., Shamir, O., and Kalai, A. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pp. 927–935, 2011.

Mémoli, F. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pp. 666–704, 1781.

Panaretos, V. M. and Zemel, Y. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019.

Perrot, M., Courty, N., Flamary, R., and Habrard, A. Mapping estimation for discrete optimal transport. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4197–4205. Curran Associates, Inc., 2016.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073.

Rabin, J. and Papadakis, N. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 256–269. Springer, 2015.

Rabin, J., Ferradans, S., and Papadakis, N. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 4852–4856. IEEE, 2014.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkQkBnJAb.

Santambrogio, F. *Optimal transport for applied mathematicians*. Birkhauser, 2015.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66:1–66:11, 2015.

Taylor, A. B. *Convex interpolation and performance estimation of first-order methods for convex optimization.* PhD thesis, 2017.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017.

Villani, C. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003. ISBN 9780821833124.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer Verlag, 2009.

Yeganova, L. and Wilbur, W. Isotonic regression under lipschitz constraint. *Journal of optimization theory and applications*, 141(2):429–443, 2009.