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

François-Pierre Paty
CREST, ENSAE,
Institut Polytechnique de Paris
francois.pierre.paty@ensae.fr

Alexandre d'Aspremont
CNRS, ENS,
PSL Research University
aspremon@ens.fr

Marco Cuturi
Google Brain,
CREST, ENSAE
cuturi@google.com

Abstract

The problem of estimating Wasserstein distances in high-dimensional spaces suffers from the curse of dimensionality: One needs an exponential (w.r.t. dimension) number of samples for the distance between two measures to be comparable to that evaluated using i.i.d samples. Therefore, using the optimal transport (OT) geometry in machine learning involves regularizing it, one way or another. One of the greatest achievements of the OT literature in recent years lies in regularity theory: one can prove under suitable hypothesis that the OT map between two measures is Lipschitz, or, equivalently when studying 2-Wasserstein distances, that the Brenier convex potential (whose gradient yields an optimal map) is a smooth function. We propose in this work to go backwards, and adopt instead regularity as a regularization tool. We propose algorithms working on discrete measures that can recover nearly optimal transport maps that have small distortion, or, equivalently, nearly optimal Brenier potential that are strongly convex and smooth. For univariate measures, we show that computing these potentials is equivalent to solving an isotonic regression problem under Lipschitz and strong monotonicity constraints. For multivariate measures the problem boils down to a non-convex QCQP problem, which can be relaxed to a semidefinite program. Most importantly, we recover as the result of this optimization the values and gradients of the Brenier potential on sampled points, but show how that they can be more generally evaluated on any new point, at the cost of solving a QP for each new evaluation. Building on these two formulations we propose practical algorithms to estimate and evaluate transport maps with desired smoothness/strong convexity properties, illustrate their statistical performance and visualize maps on a color transfer task.

1 Introduction

Optimal transport (OT) has found practical applications in areas as diverse as supervised machine learning [21, 1, 13], graphics [33, 8], generative models [6, 30], NLP [22, 3], biology [23, 32] or imaging [28, 15]. OT theory is useful for 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 textbooks such as the celebrated references by Villani [36, 37] or more recently in the exhaustive presentation by Santambrogio [31], and those used in the works cited above. In all of these applications, some form of regularization is needed to ensure 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 [17][26, §3]. Regularization, defined explicitly or implicitly as an approximation algorithm, is therefore crucial to ensure that OT is meaningful and can work at scale.

Preprint. Under review.

Brenier Potentials and Regularity theory. In the OT literature, regularity has a distinct meaning, one that is usually associated to the properties of the optimal Monge map [37, §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 the Brenier [9] 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 [10] 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" $\nu \propto e^{-W}\gamma_d$ (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 shows that the Monge map is the gradient of a Brenier [9] potential that is 1-smooth.

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. More specifically,

1. Given two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, 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. We show that nearest-Brenier potentials can be recovered as the solution of a bilevel QCQP/Wasserstein optimization problem.
2. 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 (the derivative of a convex function) must be strongly increasing and Lipschitz. A projected gradient descent approach can be used to solve this problem efficiently.
3. In the multivariate case, we show that the QCQP problem can be relaxed as a SDP, using recent advances in mathematical programming to quantify the worst-case performance of first order methods when used on smooth strongly convex functions [35, 16].
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 QP for each new point.
5. We implement and test these algorithms on various tasks, in which smooth strongly convex potentials add statistical stability, and illustrate them on a color transfer task.

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 has the solution of the Kantorovich problem [37, §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 [25] formulation of OT is, when this minimization is feasible, equivalent to that of Kantorovich, 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 Wasserstein: 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 theorem [9] 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 [37, 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 [20].

Regularity of OT maps 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 $\mathcal{E} = (E_1, \dots, E_K)$ of $\mathbb{R}^d$, we will more generally say that f is 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.

Results on the regularity of the Brenier potential were first obtained by Caffarelli [10]. For measures $\mu = e^V \gamma_d$ and $\nu = e^{-W}\gamma_d$, where $V, W : \mathbb{R}^d \to \mathbb{R}$ are convex and γ_d is the standard Gaussian measure on $\mathbb{R}^d$, the Caffarelli contraction theorem states that the optimal Brenier potential f^* (defined up to a constant) between μ and ν is 1-smooth. Although global smoothness is not always verified, the following theorem by Figalli [19] shows that local regularity holds in a general setting:

Theorem 1 (Theorem 3.5 in [19])**.** *Suppose $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ have compact support and densities f, g w.r.t $\mathscr{L}^d$ bounded away from zero and infinity, and denote by T the optimal Monge map sending μ to ν. Then there exist two negligible sets $X \subset \operatorname{supp}(\mu)$, $Y \subset \operatorname{supp}(\nu)$ such that $T : \operatorname{supp}\mu \setminus X \to \operatorname{supp}\nu \setminus Y$ is locally α-Hölder for some $\alpha > 0$.*

3 Regularity as Regularization

Contrary to the viewpoint adopted in the OT literature [11, 20], 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-strong and L-smooth, the map ∇f has distortion $\ell\|x - y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. When $\ell = L = 1$, ∇f must be a translation. Lifting the assumption that f is convex, one would recover the case where ∇f is an isometry [12, 4, 3].

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^* a L-smooth ℓ-strongly convex nearest Brenier (SSNB) potential between μ and ν if*

$$f^* \in \underset{f \in \mathcal{F}_{\ell,L,\mathcal{E}}}{\arg\min}\, W_2\left[\nabla f_\sharp \mu, \nu\right].$$

Remark 1. *For a SSNB potential we consider the associated transport value between μ and its nearest approximation of ν:*

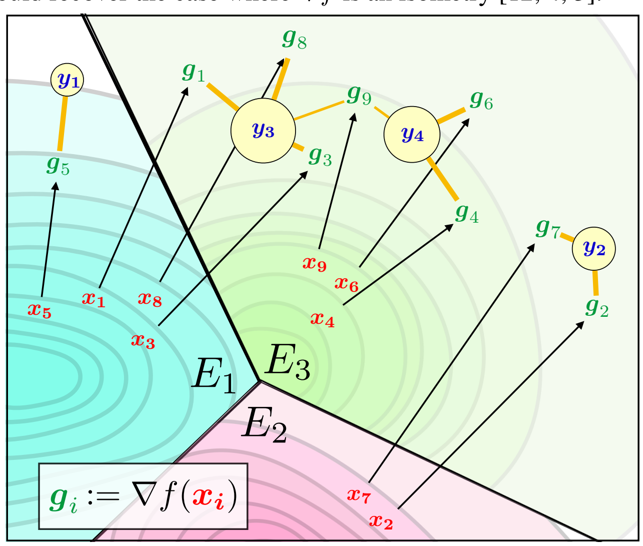


Figure 1: Points x_i mapped onto points $g_i := \nabla f(x_i)$ for a function f that is locally smooth strongly convex. SSNB potentials are such that the measure of endpoints g_i are as close as possible (in Wasserstein sense) to the measure supported on th y_j. Here this would be the sum of the squares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

$$W_2\left(\mu, \nabla f^*{}_\sharp \mu\right) = \left[\int \|x - \nabla f^*(x)\|^2 \, d\mu(x)\right]^{1/2}.$$

This quantity cannot define a metric between μ and ν because it is not symmetric in the formulation above and $W_2(\mu, \nabla f^*{}_\sharp\mu) = 0 \not\Rightarrow \mu = \nu$ (take any ν that is not a Dirac and $\mu = \delta_{\mathbb{E}[\nu]}$).

Remark 2. *The existence of an SSNB potential is proved in the supplementary material. When $E = \{\mathbb{R}^d\}$, a SSNB potential defines an optimal transport between μ and $\nabla f^*{}_\sharp\mu$. For more general partitions E one only has that property locally, and f^* can therefore be interpreted as a piecewise convex potential, giving rise to piecewise optimal transport maps, as illustrated in Figure 1.*

Algorithmic formulation as a bilevel 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 $x_1, \ldots x_n \in \mathbb{R}^d$, $y_1, \ldots y_m \in \mathbb{R}^d$, and $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_m)$ 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 μ). 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 consider optimizing over the set of convex functions with prescribed smoothness and strong-convexity constants (see for instance [34, 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 2. *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:*

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

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

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

$$\min_{v \in \mathbb{R},\, g \in \mathbb{R}^d} v \tag{2}$$

$$\text{s.t. } \forall i \in I_k,\ v \geq u_i + \langle z_i, x - x_i \rangle + \frac{1}{2(1-\ell/L)}\left(\frac{1}{L}\|g - z_i\|^2 + \ell\|x - x_i\|^2 - 2\frac{\ell}{L}\langle z_i - g, x_i - x \rangle\right).$$

We refer to the supplementary material for the proof.

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 [5, Def.5.4.2], which is the most geometrically meaningful way to recover a map from a coupling.

Definition 2 (Barycentric Projection)**.** *Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, 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 [5] shows that $\overline{\pi}$ is the gradient a convex function. It is then admissible for the SSNB optimization problem defined in Theorem 2 as soon as it verifies regularity (Lipschitzness) and curvature (strongly increasing). Although the barycentric projection map is not optimal in genera, the following proposition shows that it is however optimal for univariate measures:

Proposition 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 [31, Theorem 2.9].

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 $\overline{\pi}(x_i) = \frac{1}{a_i}\int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t)\,dt$ (proof in the supplementary material).

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

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

where $\mathbf{NW}(\mathbf{a}, \mathbf{b})$ is the so-called *North-west corner* solution [27, §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 can 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):

$$\min_{z\in\mathbb{R}^n} \sum_{i=1}^n a_i (z_i - w_i)^2 \tag{3}$$

s.t. $\forall k \leq K,\ \forall i, i+1 \in I_k,\ \ell(x_{i+1} - x_i) \leq z_{i+1} - z_i \leq L(x_{i+1} - x_i)$.

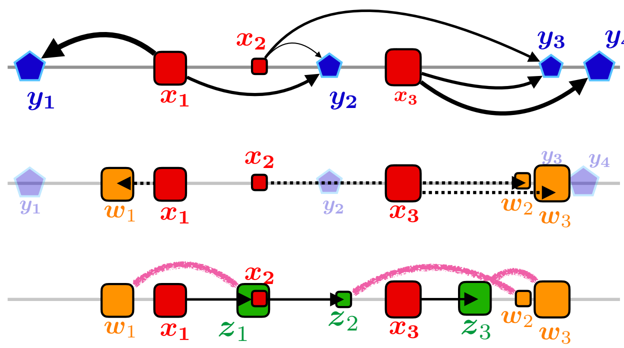


Figure 2: Top: optimal transport between two discrete measure μ, ν. Middle: the barycentric projection $\mathbf{w}$ of points $\mathbf{x}$ is displayed and corresponds to a Monge map (no mass splitting). Considering here for instance $\ell = 0.5$ and $L = 1$, the map that associates x_i to w_i is not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

The gradient of a SSNB potential f^* 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 [38] with a $\mathcal{O}(n^2)$ complexity. [2, 24] 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.

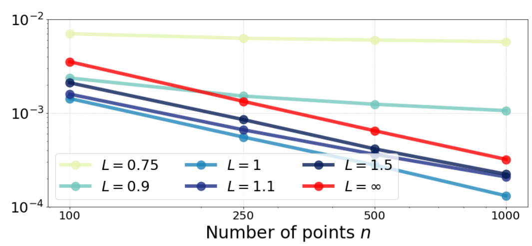


Figure 3: Take measures $\mu = \nu = \mathcal{U}([0, 1])$. For several n, we consider $\hat{\mu}_n, \hat{\nu}_n$ empirical measures over n iid samples from μ, ν, from which we compute a SSNB potential $\hat{f}_n$ with different values L, and $\ell = \min\{1, L\}$ (and $\ell = 0$ if $L = \infty$). We plot the 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 $\hat{\zeta}_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$. If L is chosen smaller than $\mathrm{Lip}(\mathrm{Id}) = 1$, 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 $\hat{\zeta}_n = \hat{\nu}_n$.

5 Semidefinite Relaxations in the Higher-dimensional Case

In this section, we provide algorithms to compute a SSNB potential in dimension $d \geq 2$ when μ, ν are discrete measures. In order to solve Problem (1), we will alternate between minimizing over

$(z_1, \ldots, z_n, u)$ and computing a coupling P solving the OT problem. The OT computation can be efficiently carried out using Sinkhorn's algorithm [14]. The other minimization is a non-convex QCQP, separable in K, that we lift as semidefinite program (SDP).

SDP relaxation Let $k \in [\![K]\!]$ and write $n_k = |E_k|$. For $z_1, \ldots, z_{n_k} \in \mathbb{R}^d$, we define $G \in \mathbb{R}^{(2n_k+m)\times(2n_k+m)}$ to be the Gram matrix associated with points $x_i \in E_k$, $z_1, \ldots, z_{n_k}$ and $y_1, \ldots, y_m$. To simplify notations, we will write $G(\alpha_i, \beta_j)$ for the coefficient in G corresponding to $\langle \alpha_i, \beta_j \rangle$ where $\alpha, \beta \in \{x, y, z\}$.

To relax the original problem, we can do a change of variables, introducing the matrix variables $Z_i = z_i z_i^T$. This is of course a nonconvex constraint in (z_i, Z_i) but we can relax it to $z_i z_i^T \preceq Z_i$ which is a Schur complement. By construction, we also have $\mathrm{trace}(Z_i) = G(z_i, z_i)$, hence $\|z_i\|_2^2 \leq G(z_i, z_i)$.

Initialize $z_1, \ldots, z_n \in \mathbb{R}^d$
repeat
  $C \leftarrow \left(\|z_i - y_j\|^2\right)_{ij}$
  $P \leftarrow \mathrm{OT}(a, b, C)$
  for $k = 1, \ldots, K$ **do**
    $\{z_i \,|\, i \in I_k\} \leftarrow \mathrm{SDP}_k(P)$
  end
until *convergence*;

Algorithm 1: Alternate Minimization for SSNB computation. $\mathrm{SDP}_k(P)$ implements a SDP solver for problem (4).

The locations $x_i \in E_k, y_1, \ldots, y_m$ are known and fixed, the terms $G(y_j, z_i)$ and $G(x_j, z_i)$ are linear, written $G(y_j, z_i) = z_i^T y_j$ and $G(x_j, z_i) = z_i^T x_j$ for $z_i \in \mathbb{R}^d$. Overall, for a fixed transport plan $P \in \mathcal{U}(a, b)$, we have to solve the following SDPs (one for each $k \in [\![K]\!]$):

$$\begin{aligned}
&\min_{G \succeq 0, u \in \mathbb{R}^{I_k}} \left\{ \sum_{i \in I_k} \sum_{j=1}^{m} P_{i,j} \left[G(y_j, y_j) + G(z_i, z_i) - 2G(y_j, z_i) \right] \right\} \qquad (4)\\
&\text{s.t. } \forall i, j \in I_k,\; u_i \geq u_j + G(z_j, x_i) - G(z_j, x_j) \\
&\qquad + \frac{1}{2(1 - \ell/L)} \left(\frac{1}{L} \left[G(z_i, z_i) + G(z_j, z_j) - 2G(z_i, z_j) \right] \right. \\
&\qquad \left. + \ell \|x_i - x_j\|^2 - 2\frac{\ell}{L} \left[G(z_j, x_j) - G(z_j, x_i) - G(z_i, x_j) + G(z_i, x_i) \right] \right) \\
&\quad \forall i \in I_k,\; \|z_i\|_2^2 \leq G(z_i, z_i),\; \forall j \in [\![m]\!],\; G(y_j, z_i) = z_i^T y_j,\; G(x_j, z_i) = z_i^T x_j.
\end{aligned}$$

This semidefinite relaxation will be tight when the ambient dimension d is of the same order as the number of points [35]. When d is smaller, there is a gap between the optimum of the SDP and that of the QCQP, and we need to find (or approximate) a low rank solution.

6 Estimation of Wasserstein Distance and Monge Map

Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ be compactly supported measures with densities w.r.t the Lebesgue measure in $\mathbb{R}^d$. Let f^* be an optimal Brenier potential such that $\nabla f^*{}_\sharp \mu = \nu$. Our goal is twofold: estimate the map ∇f^* and the value of $W_2(\mu, \nu)$.

Draw n i.i.d samples $x_1, \ldots, x_n \sim \mu$ and $y_1, \ldots, y_n \sim \nu$, and let $\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n := \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$.

Let $\hat{f}_n$ be a SSNB potential with $\mathcal{E} = \{\mathbb{R}^d\}$. Then for $x \in \operatorname{supp} \mu$ a natural estimator of $\nabla f^*(x)$ is given by a solution $\nabla \hat{f}_n(x)$ of (2).

Data: $\hat{\mu}_n$, $\hat{\nu}_n$, partition $\mathcal{E}$, N MC iterations
$(u_i, z_i)_{i \leq n} \leftarrow$ solve SSNB (1) using Algo 1
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 QP (2)
end
return
  $\hat{W} = \left[(1/N) \sum_{j=1}^N \|\hat{x}_j - \nabla \hat{f}_n(\hat{x}_j)\|^2 \right]^{1/2}$

Algorithm 2: Monte-Carlo (MC) approximation of the SSNB estimator.

This defines an estimator $\nabla \hat{f}_n : \mathbb{R}^d \to \mathbb{R}^d$ of ∇f^*, that we use to estimate $W_2(\mu, \nu)$:

Definition 3. *We define the SSNB estimator $\hat{W}_2(\mu, \nu)$ of $W_2(\mu, \nu)$ as $W_2(\mu, \nabla \hat{f}_{n\sharp} \mu)$.*

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 $\nabla \hat{f}_{n\sharp} \mu$. Then $W_2^2(\mu, \nabla \hat{f}_{n\sharp} \mu) = \int \|x - \nabla \hat{f}_n(x)\|^2 \, d\mu(x)$ can be computed using Monte-Carlo integration, whose estimation error does not depend upon the dimension d.

If $\mathcal{E} \neq \{\mathbb{R}^d\}$, $\nabla \hat{f}_n$ is the gradient of a locally convex Brenier potential, and not necessarily globally optimal. In that case $\int \|x - \nabla \hat{f}_n(x)\|^2 \, d\mu(x)$ is an approximate upper bound of $W_2^2(\mu, \nabla \hat{f}_{n\sharp}\mu)$.

Proposition 2. *Choose $\mathcal{E} = \{\mathbb{R}^d\}$, $0 \leq \ell \leq L$. If $f^* \in \mathcal{F}_{\ell,L,\mathcal{E}}$:*

$$\left| W_2(\mu, \nu) - W_2(\mu, (\nabla \hat{f}_n)_\sharp \mu) \right| \leq W_2 \left((\nabla \hat{f}_n)_\sharp \mu, \nu \right) \leq \|\nabla \hat{f}_n - \nabla f^*\|_{L^2(\mu)} \underset{n \to \infty}{\longrightarrow} 0 \text{ a.s.}$$

The study of the theoretical rate of convergence of this estimator is left for future work. Numerical simulations (see Figure 4,right) seem to indicate a faster rate of convergence compared to the classical $W_2(\hat{\mu}_n, \hat{\nu}_n)$, even in the case where ∇f^* is only locally Lipschitz and $\mathcal{E} \neq \{\mathbb{R}^d\}$. If $L < \mathrm{Lip}(\nabla f^*)$, the SSNB estimator $\hat{W}_2(\mu, \nu)$ is not consistent, as can be seen in Figure 3.

Note that in practice, the real values of ℓ and L are unknown. They can estimated by computing the optimal assignment $\sigma^* \in \arg\min_{\sigma : [\![n]\!] \to [\![n]\!]} \sum_{i=1}^n \|x_i - y_{\sigma(i)}\|^2$ and looking at the minimum and maximum values of $\|y_{\sigma(i)} - y_{\sigma(j)}\| / \|x_i - x_j\|$.

7 Experiments

All the computations were performed on a Mac Book Pro, using MOSEK as a SDP and QP solver.

7.1 Estimation of a Locally Lipschitz Monge Map

In this experiment, we consider μ 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 (upper left), 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. Clearly, f is $\ell = 1$-strongly convex, but is not smooth: ∇f is not even continuous. However, f is $L = 1$-smooth by part.

In Figure 4 (bottom left), we consider empirical measures $\hat{\mu}_n, \hat{\nu}_n$ over $n = 1000$ points. We run a k-means over $\mathrm{supp}\,\mu$ to compute $K = 400$ clusters $E_1, \ldots, E_K$. We run algorithm 1 to compute a SSNB potential $\hat{f}_n$. For several random points $x \in \mathrm{supp}\,\mu$ that are not in the support of $\hat{\mu}_n$, we compute the estimated SSNB map $\nabla \hat{f}_n(x)$ by solving the QP (2).

In Figure 4 (right), we consider empirical measures $\hat{\mu}_n, \hat{\nu}_n$ for different values of dimension $d \in \{2, 20, 100\}$ and of number of points $n \in \{10, 50, 100, 500\}$. We plot the estimation error of the SSNB estimator with $\ell = 0$ and $L = 1$ (with $K = 0.4n$ k-means clusters and $N = 50$ Monte-Carlo samples) and of the classical discrete OT estimator. The SSNB estimator seems to converge faster than the classical discrete OT estimator.

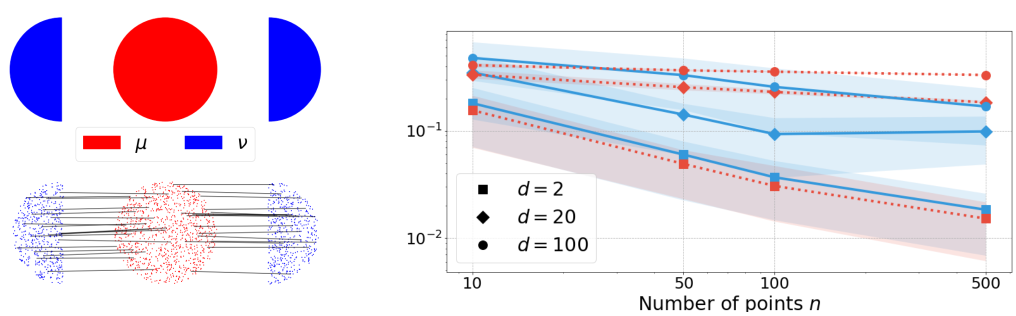


Figure 4: (Top left) Measures μ, ν. (Bottom left) Empirical measures $\hat{\mu}_n, \hat{\nu}_n$ on $n = 1000$ points. The black segments correspond to the displacement vectors $\nabla \hat{f}_n(x) - x$ for several unseen points $x \in \mathrm{supp}\,\mu$. (Right) Estimation error $|W_2(\mu, \nu) - \hat{W}_2(\mu, \nu)|$ (blue line) and $|W_2(\mu, \nu) - W_2(\hat{\mu}_n, \hat{\nu}_n)|$ (red dotted) depending on the number of points n and dimension $d \in \{2, 20, 100\}$, averaged over 100 samples. The shaded areas show the 25%-75% percentiles over the runs.

7.2 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.* [7, 18, 29]. 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.

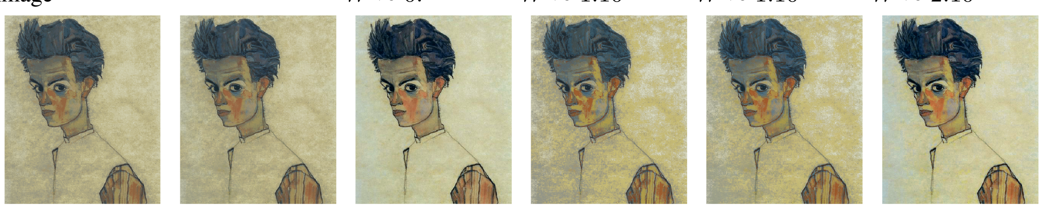
It is natural to ask that similar colors are transferred to similar colors, and that different colors are transferred to different colors. These two demands translate into the smoothness and strong convexity of the Brenier potential from which derives the color transfer mapping. We therefore propose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

In order to make to 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 solving QP (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 5, we show the color-transferred results using OT, or SSNB potentials for different values of parameters ℓ and L. Larger images are available in the supplementary material.

(a) Original Image (b) Target Image (c) Classical OT, $W \approx 0.$ (d) $\ell = 0, L = 1,$ $W \approx 1.10^{-2}$ (e) $\ell = .5, L = 1,$ $W \approx 1.10^{-2}$ (f) $\ell = 1, L = 1,$ $W \approx 2.10^{-2}$



(g) $\ell = 0, L = 2,$ $W \approx 4.10^{-3}$ (h) $\ell = .5, L = 2,$ $W \approx 5.10^{-3}$ (i) $\ell = 1, L = 2,$ $W \approx 2.10^{-2}$ (j) $\ell = 0, L = 5,$ $W \approx 2.10^{-4}$ (k) $\ell = .5, L = 5,$ $W \approx 1.10^{-3}$ (l) $\ell = 1, L = 5,$ $W \approx 2.10^{-2}$

Figure 5: (a) Schiele's portrait. (b) Van Gogh's portrait. (c) Color transfer using classical OT. (c)-(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. Smaller L values give more uniform colors, while larger ℓ values 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.

References

[1] 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.

[2] 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.

[3] Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. Unsupervised hyperalignment for multilingual word embeddings. *Proceedings of ICLR, arXiv preprint arXiv:1811.01124*, 2018.

[4] Helmut Alt and Leonidas J Guibas. Discrete geometric shapes: Matching, interpolation, and approximation. In *Handbook of computational geometry*, pages 121–153. Elsevier, 2000.

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

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:214–223, 2017.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] Nicolas Courty, Rémi Flamary, Devis Tuia, and Thomas Corpetti. Optimal transport for data fusion in remote sensing. In *2016 IEEE International Geoscience and Remote Sensing Symposium*, pages 3571–3574. IEEE, 2016.

[14] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

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

[16] 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.

[17] RM Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[18] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

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

[20] Alessio Figalli. *The Monge–Ampère equation and its applications*. 2017.

[21] 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.

[22] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. 2019.

[23] Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative RNNs. In *International Conference on Machine Learning*, pages 2417–2426, 2016.

[24] 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.

[25] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

[26] Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019.

[27] 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.

[28] 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.

[29] 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.

[30] 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`.

[31] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Birkhauser, 2015.

[32] 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.

[33] 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.

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

[35] 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.

[36] Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003. ISBN 9780821833124.

[37] Cedric Villani. *Optimal Transport: Old and New*, volume 338. Springer Verlag, 2009.

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

A Proofs

Proof for Definition 1 We write a proof 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:

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

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 $\epsilon > 0$ and let $N \in \mathbb{N}$ such that $n \geq N \Rightarrow \|\nabla f_n - g\|_\infty \leq \epsilon$. Then for $n \geq N$ and $x \in \mathbb{R}^d$,

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

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

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

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

$$f^*(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \int_0^1 \langle \nabla f_n(tx), x\, \rangle\, dt = \int_0^1 \langle \lim_{n\to\infty} \nabla f_n(tx), x\, \rangle\, dt = \int_0^1 \langle g(tx), x\, \rangle\, dt$$

so f^* is differentiable and $\nabla f^* = g$. Using Lebesgue's dominated convergence theorem, uniform (hence pointwise) convergence of $(\nabla f_n)_{n\in\mathbb{N}}$ to ∇f^* shows that $(\nabla f_n)_\sharp\mu \rightharpoonup (\nabla f^*)_\sharp\mu$. Then classical optimal transport stability theorems (*e.g.* theorem 1.51 in [31]) show that

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

i.e. f^* is a minimizer.

Proof of Theorem 2 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 $z_1, \ldots, z_n \in \mathbb{R}^d$ 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 [34, 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 $i, j \in I_k$,

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

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

The second part of the theorem is a direct application of [34, Theorem 3.14].

Proof of Proposition 1 Let $f : \mathbb{R} \to \mathbb{R}$. Then $f \in \mathcal{F}_{\ell,L,\mathcal{E}}$ if and only if it is convex and L-smooth on each set E_k, $k \in [\![K]\!]$, *i.e.* if and only if for any $k \in [\![K]\!]$, $0 \leq f''|_{E_k} \leq L$.

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

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

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:

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

Proof of Proposition 2 Using the triangular inequality for the Wasserstein distance,

$$\left|W_2(\mu,\nu) - W_2(\mu, (\nabla \hat{f}_n)_\sharp \mu)\right| \leq W_2\left((\nabla \hat{f}_n)_\sharp \mu, \nu\right).$$

Then using the fact that $(\nabla \hat{f}_n, \nabla f^*)_\sharp \mu$ is an admissible transport plan between $(\nabla \hat{f}_n)_\sharp \mu$ and ν:

$$W_2\left((\nabla \hat{f}_n)_\sharp \mu, \nu\right) = W_2\left((\nabla \hat{f}_n)_\sharp \mu, (\nabla f^*)_\sharp \mu\right) \leq \left(\int \|x-y\|^2\, d(\nabla \hat{f}_n, \nabla f^*)_\sharp \mu\right)^{1/2} = \|\nabla \hat{f}_n - \nabla f^*\|_{L^2(\mu)}.$$

Using stability of optimal transport, for example [37, Theorem 5.19],

$$(\mathrm{Id}, \nabla \hat{f}_n)_\sharp \mu \rightharpoonup (\mathrm{Id}, \nabla f^*)_\sharp \mu \text{ a.s.}$$

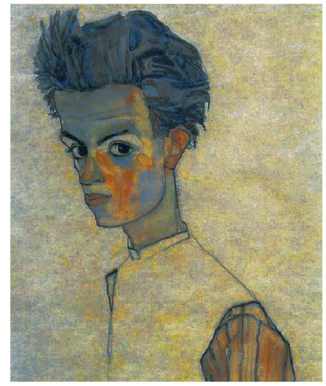
Since μ is compactly supported and $\nabla \hat{f}_n$ is Lipschitz, [31, Lemma 2.25] shows that $\|\nabla \hat{f}_n - \nabla f^*\|_{L^2(\mu)} \underset{n\to\infty}{\longrightarrow} 0$.

B Color Transfer

Higher-quality images for the color transfer application, with the same parameters.

(a) Original Image

(b) Target Image



(c) Classical OT, $W \approx 0$.

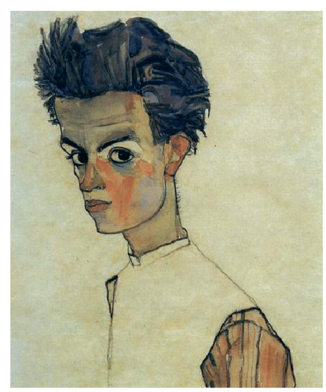
(d) $\ell = 0, L = 1, W \approx 1.10^{-2}$

(e) $\ell = 0.5, L = 1, W \approx 1.10^{-2}$

(f) $\ell = 1, L = 1, W \approx 2.10^{-2}$

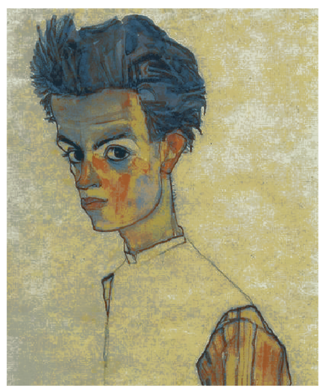
(g) $\ell = 0, L = 2, W \approx 4.10^{-3}$

(h) $\ell = 0.5, L = 2, W \approx 5.10^{-3}$

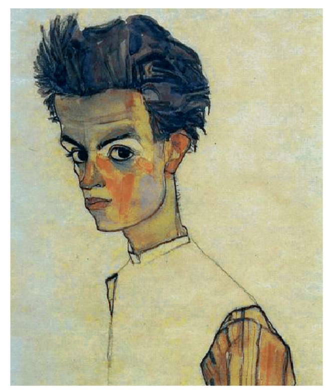


(i) $\ell = 1, L = 2, W \approx 2.10^{-2}$

(a) $\ell = 0, L = 5, W \approx 2.10^{-4}$



(b) $\ell = 0.5, L = 5, W \approx 1.10^{-3}$



(c) $\ell = 1, L = 5, W \approx 2.10^{-2}$