Abstract

The problem of estimating Wasserstein distances between two densities living in high-dimensional spaces suffers from the curse of dimensionality: One needs an exponential (w.r.t. dimension) number of i.i.d. samples from both measures to ensure that the distance between these two samples is comparable to the distance between the original densities. Therefore, the optimal transport (OT) geometry can only be used in machine learning if the OT problem is substantially regularized. On the other hand, one of the greatest achievements of the OT literature in recent years lies in regularity theory, where regular carries here a different meaning: Caffarelli [10] showed that the OT map between two well behaved measures is Lipschitz, or, equivalently when considering 2-Wasserstein distances, that the Brenier convex potential (whose gradient yields an optimal map) is a smooth function. We propose in this work to draw inspiration from this theory, and use regularity as a regularization tool. We propose algorithms operating on two discrete measures that can recover nearly optimal transport maps between these measures that have small distortion, or, equivalently, nearly optimal Brenier potentials that are both strongly convex and smooth. For univariate measures, we show that computing these potentials is equivalent to solving an isotonic regression problem with Lipschitz and strong monotonicity constraints. For multivariate measures, the problem boils down to solving alternatively a convex QCQP and a discrete OT problem. We recover as a result of these optimization the values and gradients of the Brenier potential on sampled points, but also show, more generally, that “new” values and gradients (transport maps) of that potential can be computed out of sample, at the cost of solving a simpler QCQP for each evaluation. Building on these two formulations we propose algorithms to estimate and evaluate transport maps with desired smoothness/strong convexity properties, benchmark their statistical performance, apply them to domain adaptation and visualize their action on a color transfer task.

1 Introduction

Optimal transport (OT) has found practical applications in areas as diverse as supervised machine learning [24][13], graphics [38][8], generative models [6][35], NLP [25][2], biology [26][37] or imaging [33][15]. 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 [41][42] and Santambrogio [36], 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 [18][31][§3]. 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 [42][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 [9] theorem, states
that the OT problem between $\mu$ and $\nu$ 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 $\nabla f$ 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 $\mu \propto e^W \gamma_d$ towards a “thinner than Gaussian” measure $\nu \propto e^{-W} \gamma_d$ (here $\gamma_d$ is the Gaussian measure on $\mathbb{R}^d$, $\gamma_d \propto e^{-||x||^2}$, 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 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 $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, a $L$-smooth and $\ell$-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 $\nu$. We call such potential nearest-Brenier because they provide the “nearest” way to transport $\mu$ to a measure like $\nu$ using a smooth and strongly convex potential.

2. When $\mu, \nu$ are discrete probability measures, we show that nearest-Brenier potentials can be recovered as the solution of a bilevel 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 [10] [17], yet result 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 (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.

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

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

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

and define their 2-Wasserstein distance as the solution of the Kantorovich problem [42] §6:

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

For Borel sets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, Borel map $T : \mathcal{X} \to \mathcal{Y}$ and $\mu \in \mathcal{P}(\mathcal{X})$, we denote by $T_\sharp \mu \in \mathcal{P}(\mathcal{Y})$ the push-forward of $\mu$ under $T$, i.e. the measure such that for any $A \subset \mathcal{Y}, T_\sharp \mu(A) = \mu(T^{-1}(A))$. The Monge 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

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

**Convexity and Transport: The Brenier Theorem.** Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $f : \mathbb{R}^d \to \mathbb{R}$ convex and differentiable $\mu$-a.e. Then $\nabla f$, as a map from $\mathbb{R}^d$ to $\mathbb{R}^d$ is optimal for the Monge formulation of OT between the measures $\mu$ and $\nabla f \sharp \mu$. The Brenier theorem [19] shows that if $\mu = \rho \mathcal{L}^d$ ($\mu$ is absolutely continuous w.r.t. $\mathcal{L}^d$ with density $\rho$) and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, there always exists a convex $f$ such that $\nabla f \sharp \mu = \nu$, i.e. there exists an optimal Monge map sending $\mu$ to $\nu$ that is the gradient of a convex function $f$. Such a convex function $f$ is called a Brenier potential between $\mu$ and $\nu$. If moreover $\nu = g \mathcal{L}^d$, that is $\nu$ has density $g$, a change of variable formula shows that $f$ should be solution to the Monge-Ampère [42] Eq.12.4] equation

$$\det(\nabla^2 f) = \frac{g}{\rho^{d-1}}.$$

The study of the Monge-Ampère equation is the key to obtain regularity results on $f$ and $\nabla f$, see the recent survey by Figalli [21].

**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 $\ell$-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 $\ell$-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 \(\mathcal{F}_{\ell, L, E}\) for the set of such functions.

**Regularity of OT maps.** 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 \(V, W\) are convex functions and \(\gamma_d\) is the standard Gaussian measure on \(\mathbb{R}^d\), Caffarelli’s contraction theorem states that any optimal Brenier potential \(f^*\) between \(\mu\) and \(\nu\) is 1-smooth. More general results have been proposed by Figalli [20] who showed that local regularity holds in a general setting: loosely speaking, one can obtain “local Hölder regularity by part” as soon as the measures have bounded densities and compact support.

### 3 Regularity as Regularization

Contrary to the viewpoint adopted in the OT literature [11, 21], 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 \(\mu, \nu\). Note that if a convex potential is \(\ell\)-strongly convex and \(L\)-smooth, the map \(\nabla f\) has distortion \(\ell \|x-y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L \|x-y\|\). Therefore, setting \(\ell = L = 1\) enforces that \(\nabla f\) 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 \(\nabla f\) is an isometry, considered in [12, 4, 3]. Note that this distortion also plays a role when estimating the Gromov-Wasserstein distance between general metric spaces [29] and was notably rise to piecewise optimal transport maps, as illustrated in Figure 1.

**Near-Brenier smooth strongly convex potentials.** We will seek functions \(f\) that are \(\ell\)-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 \(\nu\). If \(\nabla f_{\sharp \mu}\) were to be exactly equal to \(\nu\), such a function would be called a Brenier potential. We quantify that nearness in terms of the Wasserstein distance between the push-forward of \(\mu\) and \(\nu\) to define:

**Definition 1.** Let \(E\) be a partition of \(\mathbb{R}^d\) and \(0 \leq \ell \leq L\). For \(\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)\), we call \(f^*\) a (\(E\)-locally) \(L\)-smooth \(\ell\)-strongly convex nearest Brenier (SSNB) potential between \(\mu\) and \(\nu\) if

\[
\begin{align*}
    f^* & \in \arg\min_{f \in \mathcal{F}_{\ell, L, E}} W_2(\nabla f_{\sharp \mu}, \nu),
\end{align*}
\]

**Remark 1.** The existence of an SSNB potential is proved in the supplementary material. When \(E = \{\mathbb{R}^d\}\),

the gradient of any SSNB potential \(f^*\) defines an optimal transport map between \(\mu\) and \(\nabla f^*\_{\sharp \mu}\). The associated transport value \(W_2^2(\mu, \nabla f^*\_{\sharp \mu})\) does not define a metric between \(\mu\) and \(\nu\) because it is not symmetric, and \(W_2(\mu, \nabla f^*\_{\sharp \mu}) = 0 \iff \mu = \nu\) (take any \(\nu\) that is not Dirac and \(\mu = \delta_{\{\nu\}\})\). 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 2.

**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 \(a = (a_1, \ldots, a_n)\) and \(b = (b_1, \ldots, b_m)\) are probability weight vectors. We write \(U(a, b)\) for the transportation polytope with marginals \(a\) and \(b\), namely the set of \(n \times m\) matrices with nonnegative entries such that their row-sum and column-sum are respectively equal to \(a\) and \(b\). Set a desired smoothness \(L > 0\) and strong-convexity parameter \(\ell \leq L\), and choose a partition \(E\) of \(\mathbb{R}^d\) (in our experiments \(E\) is either \(\{\mathbb{R}^d\}\), or computed using a \(K\)-means partition of \(\{\mu, \nu\}\) using a multiplication parameter \(L \in \mathbb{N}\). For \(k \in [K]\), we write \(I_k = \{i \in [n] \mid \text{s.t. } x_i \in E_k\}\). 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

![Figure 1: Points \(x_i\) mapped onto points \(g_i := \nabla f(x_i)\) for a function \(f\) that is locally smooth strongly convex.](image)
set of convex functions with prescribed smoothness and strong-convexity constants (see for instance [39, 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 $f \in \mathcal{F}_{\ell,L,E}$ can be recovered as:

$$\begin{aligned}
\min_{z_i \in \mathbb{R}^d} W_2^2 \left( \sum_{i=1}^{n} a_i \delta_{z_i}, \nu \right) := \min_{P \in \mathcal{U}(a,b)} \sum_{i,j} P_{ij} \left( z_i - y_j \right)^2 \\
\text{s.t. } \forall k \leq K, \forall i, j \in I_k,
\quad u_i \geq u_j + (z_j, x_i - x_j) \\
\quad + \frac{1}{2(1 - \ell/L)} \left( \frac{1}{L} \left( z_i - z_j \right)^2 + \ell \| x_i - x_j \|^2 \\
\quad - 2\frac{\ell}{L} (z_j - z_i, x_j - x_i) \right).
\end{aligned} \tag{1}$$

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

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

We refer to the supplementary material for the proof.

We provide algorithms to compute a SSNB potential in dimension $d \geq 2$ when $\mu, \nu$ 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 convex QCQP, separable in $K$ smaller convex QCQP that can be solved efficiently.

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 $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$, and take $\pi$ an optimal transport plan between $\mu$ and $\nu$. The barycentric projection of $\pi$ is defined as the map $\pi : x \mapsto \mathbb{E}_{(X,Y) \sim \pi} [Y | X = x]$.

Theorem 12.4.4 in [5] shows that $\pi$ is the gradient of 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 $0 \leq \ell \leq L$. Suppose $\mu \ll \mathcal{L}^1$, or is purely atomic. Then the set of SSNB potentials between $\mu$ and $\nu$ is the set of solutions to

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

where $\pi$ is the unique optimal transport plan between $\mu$ and $\nu$ given by [39, Theorem 2.9].

Figure 2: Top: optimal transport between two discrete measure $\mu, \nu$. Middle: the barycentric projection $w$ of points $x$ 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 $\ell$-strongly increasing isotonic regression problem [3].

Discrete computations. Suppose $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ is discrete with $x_1 \leq \ldots \leq x_n$, and $\nu$ is arbitrary. Let us denote by $Q_{x_{\nu}}$ the (generalized) quantile function of $\nu$. Writing $\pi$ for the optimal transport plan between $\mu$ and $\nu$, the barycentric projection $\pi$ is explicit. Writing $a_0 := 0$, $\alpha_i := \sum_{k=1}^{i} a_k$, one has $\pi(x_i) = \frac{1}{a_i} \int_{\alpha_{i-1}}^{\alpha_i} Q_{x_{\nu}}(t) \, dt$ (proof in the supplementary material).

If $\nu$ is also discrete, with weights $b = (b_1, \ldots, b_m)$ and sorted support $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$, where...
\[ y_1 \leq \cdots \leq y_n, \] one can recover the coordinates of the vector \( (\boldsymbol{\pi}(x_i))_i \) of barycentric projections as
\[ \mathbf{w} := \text{diag}(a^{-1}) \mathbf{NW}(\mathbf{a}, \mathbf{b}) y, \]
where \( \mathbf{NW}(\mathbf{a}, \mathbf{b}) \) is the so-called North-west corner solution \([22] \S 3.4.2\) 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):
\[
\min_{z \in \mathbb{R}^n} \sum_{i=1}^n a_i (z_i - w_i)^2 \\
\text{s.t. } \forall k \leq K, \forall i, i + 1 \in I_k, \\
\ell(x_{i+1} - x_i) \leq z_{i+1} - z_i \leq L(x_{i+1} - x_i).
\]
The gradient of a SSNB potential \( f^* \) 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 [13] with an \( O(n^2) \) complexity. [2-28] developed \( O(n \log n) \) algorithms. A Smooth NB potential can therefore be exactly computed in \( O(n \log n) \) time, which is the same complexity as that 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 \( v \), namely \( z_i = \sum_{j=1}^i v_j \) (or equivalently \( v_i = z_i - z_{i-1} \) with \( z_0 := 0 \)), and writing \( w_i = \ell(x_{i+1} - x_i) \), one aims to find \( \mathbf{v} \) that minimizes \( \| A\mathbf{v} - w \|_a^2 \) subject to \( \mathbf{u} \leq \mathbf{v} \leq \mathbf{u}^* \), 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 \( \mu \) to several measures \( \nu \).

## 5 Estimation of Wasserstein Distance and Monge Map

Let \( \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d) \) 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 \( f^* \), i.e. \( f^* \) is convex and \( \nabla f^*_\mu = \nu \). Our goal is twofold: estimate the map \( \nabla f^* \) 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.

Let \( \hat{f}_n \) be a SSNB potential between \( \hat{\mu}_n \) and \( \hat{\nu}_n \) with \( \mathcal{E} = \{ \mathbb{R}^d \} \). Then for \( x \in \text{supp} \mu \), a natural estimator of \( \nabla f^*(x) \) 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 \( \nabla f^* \), 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 \( \hat{W}_2(\mu, \nu) := \int \| x - \nabla \hat{f}_n(x) \|_2^2 \, dm(x) \).

Since \( \nabla \hat{f}_n \) is the gradient of a convex Brenier potential when \( \mathcal{E} = \{ \mathbb{R}^d \} \), it is optimal between \( \mu \) and \( \nabla \hat{f}_n \mu \), and \( \hat{W}_2(\mu, \nu) = W_2(\mu, \nabla \hat{f}_n \mu) \). 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 \( \mu \) and \( \nabla \hat{f}_n \mu \). In that case \( \hat{W}_2(\mu, \nu) \) is an approximate upper bound of \( W_2(\mu, \nabla \hat{f}_n \mu) \). 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

\[
\begin{array}{l}
(\hat{u}_i, \hat{z}_i)_{i \leq n} \leftarrow \text{solve SSNB [1] between } \hat{\mu}_n, \hat{\nu}_n \\
\text{for } j \in [N] \text{ do} \\
\quad \text{Draw } \hat{x}_j \sim \mu \\
\quad \text{Find } k \text{ s.t. } \hat{x}_j \in E_k \text{ (k-means)} \\
\quad \nabla \hat{f}_n(\hat{x}_j) \leftarrow \text{solve QCQP [2]}
\end{array}
\]

**end for**

**Output:** \( \hat{W} = \left( \frac{1}{N} \sum_{j=1}^N \| \hat{x}_j - \nabla \hat{f}_n(\hat{x}_j) \|_2^2 \right)^{1/2} \)
We show that when the Brenier potential $f^*$ is globally regular, the SSNB estimator is strongly consistent:

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

$$
|W_2(\mu, \nu) - \hat{W}_2(\mu, \nu)| \leq W_2 \left( \nabla \hat{f}_n, \mu, \nu \right)
\leq \|\nabla \hat{f}_n - \nabla f^*\|_{L^2(\mu)} \xrightarrow{n \to \infty} 0.
$$

The study of the theoretical rate of convergence of this estimator is beyond the scope of this work. Recent results [27, 23] show that assuming some regularity of the Monge map $\nabla f^*$ leads to improved sample complexity rates. Numerical simulations (see section 6.1) seem to indicate an reduced estimation error for SSNB over the classical $W_2(\hat{\mu}_n, \hat{\nu}_n)$, even in the case where $\nabla f^*$ is only locally Lipschitz and $\mathcal{E} \neq \{\mathbb{R}^d\}$. If $L < \text{Lip}(\nabla f^*)$, the SSNB estimator $\hat{W}_2(\mu, \nu)$ 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:** $\mu$ is the uniform measure over the unit cube in $\mathbb{R}^d$, and $\nu = T_2 \mu$ where $T(x) = \Omega_2 x$, with $\Omega_2 = \text{Diag}(0.8 + \frac{0.5}{d}, \ldots, 0.8 + \frac{0.5}{d})$. $T$ 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:** $\mu$ is the uniform measure over the unit ball in $\mathbb{R}^d$, and $\nu = T_2 \mu$ where $T(x_1, \ldots, x_d) = (x_1 + 2 \text{sign}(x_1), x_2, \ldots, x_d)$. As can be seen in Figure 4, $T$ 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: $\nabla f$ is not even continuous. However, $f$ is $L = 1$-smooth by part.

For each of those two settings, we consider i.i.d samples $x_1, \ldots, x_n \sim \mu$ and $y_1, \ldots, y_n \sim \nu$ for different values of $n \in \mathbb{N}$, and denote by $\hat{\mu}_n$ and $\hat{\nu}_n$ the respective empirical measures on these points. Given a number of clusters $1 \leq K \leq n$, we compute the partition $\mathcal{E} = \{E_1, \ldots, E_K\}$ by running k-means with $K$ clusters on data $x_1, \ldots, x_n$. In both settings, we run the algorithms with $\ell = 0.6$ and $L = 1.4$. We give experimental results on the statistical performance of our SSNB estimator, computed using Monte-Carlo integration, compared to the classical optimal transport estimator $W_2(\hat{\mu}_n, \hat{\nu}_n)$.

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 $n = 500$ and $d = 30$. In the global regularity setting (Figure 5 top), the error seem 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.
We consider DeCAF6 features [16], which are sparse. 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 [16], which are sparse 4906-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, entropy regularized OT, Group-Lasso and entropy regularized OT and SSNB are given in Table 1.

### 6.2 Domain Adaptation

Domain adaptation is a way to transfer knowledge from a source dataset to a target dataset. The source dataset consists of labelled data, and the goal is to classify the target data. In [13], optimal transport (and different regularized version of OT) is used 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 [16], which are sparse 4906-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, entropy regularized 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, b) learn the SSNB map between the $10 \times k$ 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 | 84.8 ± 1.1 |
| A → D   | 65.0 | 84.3 ± 5.2 | 82.9 ± 5.2 | 82.3 ± 3.0 |
| A → W   | 65.0 | 79.3 ± 5.3 | 81.6 ± 3.1 | 82.7 ± 2.6 |
| C → A   | 75.3 | 90.0 ± 1.3 | 91.1 ± 1.8 | 91.7 ± 1.0 |
| C → D   | 65.0 | 83.9 ± 4.0 | 85.8 ± 2.3 | 81.9 ± 4.8 |
| C → W   | 64.8 | 77.1 ± 7.9 | 81.9 ± 5.1 | 83.9 ± 4.3 |
| D → A   | 69.0 | 88.3 ± 3.1 | 90.5 ± 1.9 | 92.6 ± 0.9 |
| D → C   | 66.4 | 78.0 ± 2.9 | 85.6 ± 2.2 | 83.2 ± 2.1 |
| D → W   | 84.7 | 96.1 ± 2.1 | 93.8 ± 2.1 | 94.7 ± 2.7 |
| W → A   | 66.4 | 82.5 ± 4.3 | 89.3 ± 2.7 | 91.8 ± 1.3 |
| W → C   | 62.5 | 74.8 ± 2.3 | 80.2 ± 2.4 | 81.5 ± 1.4 |
| W → D   | 87.3 | 97.5 ± 2.1 | 96.4 ± 3.7 | 94.9 ± 3.2 |
| Mean     | 70.6 | 81.4 ± 1.7 | 87.1 ± 1.9 | 87.1 ± 1.9 |

Table 1: Results for OT are based on the whole datasets. OT-IT: Entropy regularized OT. OT-GL: Entropy + Group Lasso regularized OT. Search intervals for OT-IT and OT-GL are $\epsilon, \eta \in \{10^{-3}, \ldots, 10^{3}\}$, and for SSNB: $\ell \in \{0.2, 0.5, 0.7\}$, $L \in \{0.5, 0.7, 1.2\}$, $k \in \{1, 5\}$. [10]
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. [7, 19, 34]. Each image is represented by a point cloud in the RGB color space identified with \([0, 1]^3\). The optimal transport plan \(\pi\) 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 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 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 \(\ell\) and \(L\). Larger images are available in the supplementary material.

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.
References

[1] Soroosh Shafieezadeh Abadeh, Peyman 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 hyper-alignment for multilingual word embeddings. In International Conference on Learning Representations, 2019.

[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 Courtuy, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence, 39(9):1853–1865, 2016.

[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] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In International conference on machine learning, pages 647–655, 2014.

[17] 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.

[18] Richard M. Dudley. The speed of mean Glivenko-Cantelli convergence. Annals of Mathematical Statistics, 40(1):40–50, 1969.

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

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

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

[22] Rémi Flamary, Nicolas Courty, Alain Rakotomamonjy, and Devis Tuia. Optimal transport with laplacian regularization. In NIPS 2014, Workshop on Optimal Transport and Machine Learning, 2014.

[23] Rémi Flamary, Karim Lounici, and André Ferrari. Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation. arXiv preprint arXiv:1905.10155, 2019.

[24] 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.

[25] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. 2019.
Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative RNNs. In *International Conference on Machine Learning*, pages 2417–2426, 2016.

Jan-Christian Hütter and Philippe Rigollet. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

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

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

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

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.

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

Julien Rabin, 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.

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=rkQkBnJA

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

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

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. *Convolutional Wasserstein distances: efficient optimal transportation on geometric domains*. *ACM Transactions on Graphics*, 34(4):66:1–66:11, 2015.

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

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

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

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

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
\[
W_2[(\nabla f_n)_{2\mu}, \nu] \leq \frac{1}{n + 1} + \inf_{f \in \mathcal{F}_{L,L}} W_2[(\nabla f)_{2\mu}, \nu].
\]

Let \( x_0 \in \text{supp}(\mu) \). 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 \( \nu[B(0, R)] \geq 1 - \frac{1}{2} \). Then for \( C > 0 \) large enough, there exists an \( n \in \mathbb{N} \) such that:
\[
W_2^2[(\nabla f_n)_{2\mu}, \nu] = \min_{\pi \in \Pi(\mu, \nu)} \int \| \nabla f_n(x) - y \|^2 \, d\pi(x, y)
\]
\[
\geq \int \| \nabla f_n(x) - \text{proj}_{B(0, R)}[\nabla f_n(x)] \|^2 \, d\mu(x)
\]
\[
\geq \int_{B(x_0, r) \cap \text{supp}(\mu)} \| \nabla f_n - \text{proj}_{B(0, R)}[\nabla f_n] \|^2 \, d\mu
\]
\[
\geq \frac{1}{2} V \min_{y \in B(0, R)} \| \nabla f_n(x) - y \|^2
\]
\[
\geq \frac{1}{2} V(C - LR - R)
\]
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 \( \phi, \psi \) be two extractions and \( \alpha, \beta \) such that \( f_{\phi(n)}(x) \rightarrow \alpha \) and \( f_{\psi(n)}(x) \rightarrow \beta \). Then
\[
|\alpha - \beta| = \lim_{n \rightarrow \infty} \left| \int_0^1 \langle \nabla f_{\phi(n)}(tx) - \nabla f_{\psi(n)}(tx), x \rangle \, dt \right|
\]
\[
\leq \lim_{n \rightarrow \infty} \| x \| \| \nabla f_{\phi(n)} - \nabla f_{\psi(n)} \|_\infty = 0.
\]

This shows that \( (f_n)_{n \in \mathbb{N}} \) converges pointwise to some function \( f^* \). In particular, \( f^* \) is convex. For \( x \in \mathbb{R}^d \), using Lebesgue’s dominated convergence theorem,
\[
f^*(x) = \lim_{n \rightarrow \infty} f_n(x) = \lim_{n \rightarrow \infty} \int_0^1 \langle \nabla f_n(tx), x \rangle \, dt
\]
\[
= \int_0^1 \langle \lim_{n \rightarrow \infty} \nabla f_n(tx), x \rangle \, dt = \int_0^1 \langle g(tx), x \rangle \, dt
\]
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 \( \nabla f^* \) shows that \( (\nabla f_n)_{2\mu} \rightharpoonup (\nabla f^*)_{2\mu} \). Then classical optimal transport stability theorems (e.g. theorem 1.51 in [36]) show that
\[
W_2[(\nabla f^*)_{2\mu}, \nu] = \lim_{n \rightarrow \infty} W_2[(\nabla f_n)_{2\mu}, \nu]
\]
\[
= \inf_{f \in \mathcal{F}_{L,L}} W_2[(\nabla f)_{2\mu}, \nu],
\]
i.e. \( f^* \) is a minimizer.

Proof of Theorem 1 For \( f \in \mathcal{F}_{L,L,E} \), \( \nabla f_\mu = \sum_{i=1}^n a_i \delta_{z_i} \). Writing \( z_i = \nabla f(x_i) \), we wish to minimize \( W_2^2 \left( \sum_{i=1}^n a_i \delta_{z_i} \right) \) over all the points \( z_1, \ldots, z_n \in \mathbb{R}^d \) such that there exists \( f \in \mathcal{F}_{L,L,E} \) with \( \nabla f(x_i) = z_i \) for all \( i \in [n] \). Following [39] 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 \( i, j \in I_k \),
\[
u_i \geq \nu_j + \langle z_j, x_i - x_j \rangle + \frac{1}{2(1 - \ell/L)} \left( \frac{L}{\ell} \| z_i - z_j \|^2 + \ell \| x_i - x_j \|^2 - 2 \ell \langle z_j - z_i, x_j - x_i \rangle \right)
\]
Then minimizing over \( f \in \mathcal{F}_{L,L,E} \) is equivalent to minimizing over \( (z_1, \ldots, z_n, u) \) under these interpolation constraints.

The second part of the theorem is a direct application of [39] Theorem 3.14).

Proof of Proposition 1 Let \( f : \mathbb{R} \rightarrow \mathbb{R} \). Then \( f \in \mathcal{F}_{L,L,E} \) 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 \( \rho \), let us write \( F_\rho \) and \( Q_\rho \) the cumulative distribution function and the quantile function (i.e. the generalized inverse of the cumulative distribution function). Then \( Q_\rho \circ f_\mu = \nabla f \circ Q_\mu \).

Using the closed-form formula for the Wasserstein distance in dimension 1, the objective we wish to minimize (over \( f \in \mathcal{F}_{L,L,E} \)) is:
\[
W_2^2(f_\mu, \nu) = \int_0^1 [f' \circ Q_\mu(t) - Q_\nu(t)]^2 \, dt.
\]

Suppose \( \mu \) has a density w.r.t. the Lebesgue measure. Then by a change of variable, the objective becomes
\[
\int_{-\infty}^{+\infty} [f'(x) - Q_\nu \circ F_\mu(x)]^2 \, d\mu(x) = \| f' - \pi \|_{L^2(\mu)}^2.
\]
Indeed, \( Q_\nu \circ F_\mu \) is the optimal transport map from \( \mu \) to \( \nu \), hence its own barycentric projection. The result
Suppose now that \( \mu \) 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 \( \alpha_i = \sum_{k=i}^{n} a_k \) with \( a_0 = 0 \). Then

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

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 \( W_2^2(f^*_\mu, \nu) \) over \( f \in F_{\ell, L, \epsilon} \) is equivalent to solve

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

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

\[
\pi(x_i) = \frac{1}{a_i} \int_{-\infty}^{\infty} y 1\{x = x_i\} d\pi(x, y) \\
= \frac{1}{a_i} \int_{-\infty}^{\infty} y 1\{x = x_i\} d(Q_\mu, Q_\nu)_t |[0, 1] \\
= \frac{1}{a_i} \int_{0}^{1} Q_\nu(t) 1\{Q_\mu(t) = x_i\} dt \\
= \frac{1}{a_I} \int_{\alpha_{i-1}}^{\alpha_i} Q_\nu(t) dt.
\]

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

\[
|W_2(\mu, \nu) - W_2(\mu, (\nabla \hat{f}_n)_{\sharp} \mu)| \leq W_2(\nabla \hat{f}_n)_{\sharp} \mu, \nu \).
\]

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 \( \nu \):

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

Using stability of optimal transport, for example [42, Theorem 5.19],

\[
(Id, \nabla \hat{f}_n)_{\sharp} \mu \rightarrow (Id, \nabla f^*)_{\sharp} \mu \text{ a.s.}
\]

Since \( \mu \) is compactly supported and \( \nabla \hat{f}_n \) is Lipschitz, [38, Lemma 2.25] shows that \( \|\nabla \hat{f}_n - \nabla f^*\|_{L^2(\mu)} \rightarrow 0 \).
(a) $\ell = 1, L = 2, \quad W \approx 2 \times 10^{-2}$
(b) $\ell = 0, L = 5, \quad W \approx 2 \times 10^{-4}$
(c) $\ell = 0.5, L = 5, \quad W \approx 1 \times 10^{-3}$
(d) $\ell = 1, L = 5, \quad W \approx 2 \times 10^{-2}$