A MULTISCALE APPROACH TO OPTIMAL TRANSPORT

QUENTIN MÉRIGOT
LABORATOIRE JEAN KUNTZMANN, UNIVERSITÉ DE GRENOBLE AND CNRS

Abstract. In this paper, we propose an improvement of an algorithm of Aurenhammer, Hoffmann and Aronov to find a least square matching between a probability density and finite set of sites with mass constraints, in the Euclidean plane. Our algorithm exploits the multiscale nature of this optimal transport problem. We iteratively simplify the target using Lloyd's algorithm, and use the solution of the simplified problem as a rough initial solution to the more complex one. This approach allows for fast estimation of distances between measures related to optimal transport (known as Earth-mover or Wasserstein distances). We also discuss the implementation of these algorithms, and compare the original one to its multiscale counterpart.

1. Introduction

Engineer and mathematician Gaspard Monge proposed the following problem [Mon81]: what is the cheapest way to transport a pile of sand into a hole with minimum cost, knowing that moving an individual particle from a position x to another position y has a cost $c(x, y)$? This problem gave birth to the field of optimal transport, which has been very vivid in the past twenty years, with applications in geometry, probability and PDEs (see e.g. [Vil09]).

However, Monge's problem was an engineering problem, and it is not very surprising that various form of optimal transport appeared in many applied fields. In computer vision, distances defined using optimal transport have been used as a way to compare the color histograms of images in [RTG00] under the name of *Earth mover distance*. Optimal transport on the circle has been used for transferring the *hue* of an image to another [DSS10]. In combination with Lloyd's algorithm, optimal transport seems a interesting tool for optimal quantization [BSD09]. More recently, methods inspired by (or relying on) optimal transport have been proposed as a tool in several parts of geometry processing: surface comparison [LD11], surface reconstruction from data corrupted with outliers [CCSM10, MDGD$^+$10], construction of optimized primal-dual triangulations [MMD11], reconstruction with sharp corners and edges [dGCSAD11].

Yet, the lack of a practical method to compute optimal transports maps except in 1D has hindered the development of many of these applications. Even in the simplest planar situation, namely with $c(x, y) = \|x - y\|^2$, there is a lack of reasonably fast and widely usable method able to efficiently compute optimal transport maps.

1.1. L^2 optimal transport.

In the remaining of the paper, we will deal with the L^2 optimal transport, i.e. where the cost for moving a certain amount of mass from a point x to a point y is proportional to square of the Euclidean distance $\|x - y\|^2$. This model is well-understood theoretically, and has several nice properties (such as uniqueness of the solution) that follow from strict concavity of the cost. Numerical schemes have been proposed by Brenier-Benamou [BB00], Loeper [LR05] and Angenent-Haker-Tannenbaum [AHT03] to solve L^2 optimal transport. However, numerical instabilities makes them difficult to use for general problems:

for instance, [LR05] requires a lower bound on the density of the source measure, while the gradient descent algorithm of [AHT03] suffers from a drift effect which produces optimal maps that are not transport plans. Another possibility to find optimal transport plans is by discretizing the source and/or target measure. These discrete approach include linear programming, the Hungarian method, and a variant known as Bertsekas' auction algorithm [Ber88]. These methods work for general cost functions, and are often unable to take advantage of the geometric simplifications that occur when working specifically with the squared Euclidean metric.

A promising approach, both from the practical and the theoretical point of view has been proposed in [AHA98]. In this approach, the source measure has a density ρ while the target measure is given by a sum of Dirac masses supported on a finite set S. The fact that the source measure has a density ensures the existence and uniqueness of the optimal transport map. In this case, solving the optimal transport problem amounts to finding a weight vector $(w_p)_{p\in S}$ such that the power diagram of (S, w) has the following property: for every point s in S, the proportion of the mass of ρ contained in the corresponding power cell should be equal to the mass of the Dirac mass at s (see Section 2.2). In the same article, the authors also introduced a convex function whose minimum is attained at this optimal weight vector, thus transforming the optimal transport problem into an unconstrained convex optimization problem on $\mathbb{R}^N$ where N is the number of points in the set S.

1.2. **Contributions.** We revisit the approach of [AHA98] with implementation in mind. Our main contribution is a multiscale approach for solving the unconstrained convex minimization problem introduced in [AHA98], and thus to solve L^2 optimal transport.

Let us sketch briefly the main idea. In order to solve the optimal transport problem between a measure with density μ such as a grayscale image and a discrete measure ν, we build a sequence $\nu_0 = \nu, \dots, \nu_L$ of simplifications of ν using Lloyd's algorithm. We start by solving the much easier transport problem between μ and the roughest measure ν_L using standard convex optimization techniques. Then, we use the solution of this problem to build an initial guess for the optimal transport between μ and ν_{L-1}. We then proceed to convex optimization starting from this guess to solve the optimal transport between μ and ν_{L-1}. This is repeated until we have obtained a solution to the original problem. If the original target measure has a density, we use Lloyd's algorithm to obtain the first discretization ν. Note that this idea was independently proposed in [Bos10].

This procedure provides a significant speedup, up to an order of magnitude, for computing optimal transport plans. Moreover, at every step of the algorithm it is possible to obtain a lower and upper bound on the Wasserstein distance (also known as Earth-mover distances [RTG00]) between the source measure μ and the original target measure ν. Using this approach, one can obtain rough estimates of Wasserstein distances between two images with a speedup of up to two order of magnitude over the simple convex optimization approach.

2. Background

We briefly recap the few concepts of measure theory and optimal transport that we use, before explaining the relation between the L^2 optimal transport and power diagrams. We also recall how optimal transport can be turned into an unconstrained convex optimization problem.

2.1. **Measure theory and Optimal transport.** A *non-negative measure* μ on the space $\mathbb{R}^d$ is a map from (measurable) subsets of $\mathbb{R}^d$ to a non-negative numbers, which is *additive* in the sense that $\mu\left(\cup_{i\in\mathcal{N}} B_i\right) = \sum_i \mu(B_i)$ whenever (B_i) is a

FIGURE 1. Interpolation between the standard pictures *Photograph* and *Peppers* obtained with our algorithm (see Section 5.3). The target image was quantized with 625 Dirac masses for the first row, and 15625 Dirac masses for the second row. The cells in these pictures are interpolation between power cells (on the left) and Voronoi cells (on the right), obtained by linearly varying the weights from their value on the left picture to zero. Their color vary so that the product of the area of a cell multiplied by its gray level remains constant over time.

countable family of disjoint (measurable) subsets. The *total mass* of a measure μ is $\text{mass}(\mu) := \mu(\mathbb{R}^d)$. A measure μ with unit total mass is called a *probability measure*. The *support* of a measure μ, denoted by $\text{spt}(\mu)$ is the smallest closed set whose complement has zero measure.

The optimal transport problem involves two probability measures: a source measure μ, and a target measure ν. We will always suppose that the source measure μ has a density, i.e. there exists a non-negative function ρ on $\mathbb{R}^d$ such that for every (measurable) subset B of $\mathbb{R}^d$,

$$\mu(B) := \int_B \rho(x)\mathrm{d}x.$$

On the other hand, we will assume that the target measure ν is discrete, supported on a finite set S of $\mathbb{R}^d$. This means that there exists a family of positive coefficients $(\lambda_p)_{p\in S}$ such that for every subset B,

$$\nu(B) = \sum_{p\in S\cap B} \lambda_p.$$

The above formula is equivalent to writing ν as the sum $\sum_{p\in S} \lambda_p \delta_p$, where δ_p is the unit *Dirac mass* at p. The integral of a continuous function ϕ with respect to these two measures is $\int_{\mathbb{R}^d} \phi(x)\mathrm{d}\mu(x) := \int_{\mathbb{R}^d} \phi(x)\rho(x)\mathrm{d}x$, and $\int_{\mathbb{R}^d} \phi(x)\mathrm{d}\nu(x) := \sum_{p\in S} \lambda_p \phi(p)$.

In Section 3.1 we will see how to adapt the proposed method to the case where the source and target measures both have density.

Transport map. The *pushforward* of a measure μ by a map $T : \mathbb{R}^d \to \mathbb{R}^d$ is another measure $T_\#\mu$ defined by the equation $T_\#\mu(B) := \mu(T^{-1}(B))$ for every subset B of $\mathbb{R}^d$. A map T is called a *transport map* between μ and ν if the pushforward of μ by T is ν. We denote by $\Pi(\mu,\nu)$ the set of transport maps between μ and ν.

For instance, a map T is a transport map between the source measure μ and the target measure ν described at the beginning of this paragraph if and only if for

every point p in the support S of ν,

$$\lambda_p = \mu(T^{-1}(\{p\})) \quad \left[= \int_{T^{-1}(\{p\})} \rho(x)\mathrm{d}x \right].$$

Optimal transport maps. The *cost* of a transport map T between the source measure μ with density ρ and the target measure ν is defined by:

$$c(T) := \int_{\mathbb{R}^d} \|x - T(x)\|^2 \mathrm{d}\mu(x) \quad \left[= \int_{\mathbb{R}^d} \|x - T(x)\|^2 \rho(x)\mathrm{d}x \right]$$

The problem of optimal transport, also called *Monge's problem*, is to find a transport map T_{opt} whose cost is minimal among all transport maps between μ and ν, i.e.

$$T_{\mathrm{opt}} := \arg\min\{c(T);\, T \in \Pi(\mu, \nu)\}$$

The non-convexity of both the cost function c and of the set of transport plans makes it difficult, in general, to prove the existence of a minimum. However, in this specific case where μ has a density and the cost is the squared Euclidean norm the existence follows from [Bre91].

Wasserstein distance. The *Wasserstein distance* between two probability measures μ and ν, μ having a density, is the square root of the minimal transport cost. We denote it by $\mathrm{Wass}_2(\mu, \nu)$. Intuitively, the Wasserstein distance measures the minimum global cost of transporting every bit of μ onto ν, supposing that moving an infinitesimal amount $\mathrm{d}\mu(x)$ from x to y is equal to $\|x - y\|^2 \mathrm{d}\mu(x)$.

Note that our definition above is not symmetric, as it requires the source measure to have a density. However, this restriction can be leveraged using the notion of *transport plan* instead of transport map, leading to a definition of Wasserstein distance between any pair of probability measures [Vil09, Ch. 6]).

2.2. **Power diagrams.** Let S be a finite set of points in $\mathbb{R}^d$, and $w : S \to \mathbb{R}$ be a given weight vector. The *power diagram* or *weighted Voronoi diagram* of (S, w) is a decomposition of the ambient space in a finite number of cells, one for each point in S, defined by the property that a point x belongs to $\mathrm{Vor}_S^w(p)$ iff for every q in S, one has $\|x - p\|^2 - w(p) \leq \|x - q\|^2 - w(q)$. Note that if the weights are all zero, this coincides with the usual Voronoi diagram.

Given such a decomposition, we will consider the application T_S^w which maps every point x in the power cell $\mathrm{Vor}_S^w(p)$ to its center p. This map is well-defined everywhere except on the boundary of the power cells, but since this set has zero Lebesgue measure this has no consequence for us. The pushforward of the source measure μ by the map T_S^w is a sum of Dirac masses centered at every point p in the set S, whose mass is the μ-mass of the corresponding power cell:

$$T_S^w|_{\#}\,\mu = \sum_{p \in S} \mu(\mathrm{Vor}_S^w(p))\delta_p,$$

In [AHA98], the following theorem was proven. Note that this result, as well as Theorem 2, can be also obtained as a consequence of Brenier theorem [Bre91].

Theorem 1. For any probability measure μ with density, and a weighted set of points (S, w), the map T_S^w is an optimal transport map between the measure μ and the pushforward measure $T_S^w|_{\#}\,\mu$. Consequently,

$$\mathrm{Wass}_2\left(\mu, T_S^w|_{\#}\,\mu\right) = \left(\sum_{p \in S} \int_{\mathrm{Vor}_S^w(p)} \|x - p\|^2 \rho(x)\mathrm{d}x\right)^{1/2}.$$

This theorem gives examples of optimal transport plans and the discrete probability measures supported on S that can be written as the pushforward $T_S^w|_\# \mu$, where w is a set of weights on S. As we will see in the next section (Theorem 2), it turns out that every probability measure supported on S can be written in this way. Said otherwise, the optimal transport maps between μ and a measure supported on S can always be written as $T_{S,w}$ for some weight vector w.

2.3. **An unconstrained convex optimization problem. Adapted weight vectors.** Let μ be a probability measure with density ρ, and ν a discrete probability measure supported on a finite set S, with $\nu = \sum_{p\in S} \lambda_p \delta_p$. A weight vector $w : S \to \mathbb{R}$ on S is called *adapted* to the couple of measures (μ, ν) if for every site p in S, one has

$$\lambda_p = \mu(\mathrm{Vor}_S^w(p)) \quad \left[= \int_{\mathrm{Vor}_S^w(p)} \rho(x)\mathrm{d}x \right]. \tag{1}$$

By Theorem 1, if the weigh vector w is adapted to the couple (μ, ν) then the optimal transport between μ and the discrete measure ν is given by the map $T_{S,w}$.

Moreover, Theorem 2 below asserts that finding a weight vector adapted to the couple (μ, ν) amounts to finding a global minimum of the function Φ below, thus turning the very constrained original problem (minimization among convex maps) into an unconstrained convex optimization problem. Note that this theorem follows from the discussion in Section 5 of [AHA98].

$$\Phi(w) := \sum_{p\in S} \left(\lambda_p w(p) - \int_{\mathrm{Vor}_S^w(s)} (\|x - s\|^2 - w(p))\mathrm{d}\mu(x) \right) \tag{2}$$

Theorem 2. Given a measure μ with density ρ on $\mathbb{R}^d$, and $\nu = \sum_{s\in S} \lambda_p \delta_p$, the following three statements are equivalent:

(i) the power map T_S^w realizes an optimal transport between the measures μ and ν ;
(ii) w is adapted to the couple (μ, ν);
(iii) w is a global minimizer of the convex function Φ.

We will recall briefly how the value of the gradient of the function Φ can be computed at a weight vector w. Incidentally, the steps necessary to this computation almost sketch a complete proof of the theorem. Consider the map

$$\Psi(w) := \sum_{p\in S} \int_{\mathrm{Vor}_S^w(p)} (\|x - p\|^2 - w(p))\mathrm{d}\mu(x)$$

which is related to the our function by the equation $\Phi(w) = \sum_{p\in S} \lambda_p w(p) - \Psi(w)$. The map Ψ is concave, as we will show by writing it as an infimum of linear functions. Consider *any* map T from $\mathbb{R}^d$ to the finite set S. By definition of the power cell, for every point x in $\mathrm{Vor}_S^w(p)$ one has

$$\|x - p\|^2 - w(p) \leq \|x - T(x)\|^2 - w(T(x))$$

As a consequence, the function Ψ can be rewritten as

$$\Psi(w) = \inf_T \Psi_T(w), \text{ where}$$

$$\Psi_T(w) := \int_{\mathbb{R}^2} (\|x - T(x)\|^2 - w(T(x)))\mathrm{d}\mu(x).$$

Since the functions Ψ_T all depend linearly on w, the function Ψ is concave, and Φ is convex. Moreover, it is easy to check that the values of the functions Ψ and $\Psi_{T_S^w}$

coincide for the weight vector w. Consequently, their gradient coincide to at that point, and a simple computation shows :

$$\frac{\partial \Psi}{\partial w(p)}(w) = \int_{\mathrm{Vor}_S^w(p)} \rho(x)\mathrm{d}x \tag{3}$$

$$\frac{\partial \Phi}{\partial w(p)}(w) = \lambda_p - \int_{\mathrm{Vor}_S^w(p)} \rho(x)\mathrm{d}x \tag{4}$$

In particular, the second equation shows that the gradient $\nabla\Phi$ vanishes at a weight vector w if and only if w satisfies Eq. (1).

3. Multiscale Approach for Minimizing Φ

The efficiency of an optimization technique relies on two important choices. The most important one is the choice of descent algorithm, as it is well-known that the difference in efficiency between (for instance) the first order simple gradient descent algorithm and the second order Newton methods can be tremendous [Fle87].

The second one is the choice of the position from where the optimization is started. Its importance shouldn't be disregarded, even for convex optimization, as the second-order convergence in Newton's descent does only happen in a small basin around the global minimizer.

In this section, we introduce our multiscale algorithm for finding a global minimum of Φ. We start by building a decomposition of the target measure ν, i.e. a sequence of discrete measures $\nu_0 := \nu, \nu_1, \dots, \nu_L$ that are simpler and simpler as L increases. The level ℓ of the decomposition is then used to construct a good initial weight vector for the optimal transport problem $(\mu, \nu_{\ell-1})$, in a hierarchical way.

3.1. Decomposition of the target measure.

A *decomposition* of the target measure $\nu = \sum_{p\in S} \lambda_p \delta_p$ is a sequence of discrete probability measures $(\nu_\ell)_{\ell\geq 0}$ such that $\nu_0 = \nu$ and ν_ℓ is supported on a set S_ℓ:

$$\nu_\ell = \sum_{p\in S_\ell} \lambda_{p,\ell}\delta_p.$$

Moreover, for every level ℓ we are given a transport map between ν_ℓ and $\nu_{\ell+1}$, that is a map π_ℓ from the support S_ℓ of ν_ℓ to the (smaller) support $S_{\ell+1}$ of $\nu_{\ell+1}$ with the additional property that for every point p in $S_{\ell+1}$,

$$\lambda_{p,\ell+1} = \sum_{q\in\pi_\ell^{-1}(p)} \lambda_{q,\ell} \tag{5}$$

The decomposition that we will consider in practice are constructed using Lloyd's algorithm, as explained in Section 4.2. This means that the transport map π_ℓ maps every point p in S_ℓ to its nearest neighbor in $S_{\ell+1}$.

We remark that having access to such a transport map between ν_ℓ and $\nu_{\ell+1}$ allows to bound the Wasserstein distance between these two measures. By considering the composition of transport maps, it is also possible to bound the distance between e.g. ν and ν_L. Letting $\pi = \pi_{L-1} \circ \ldots \circ \pi_0$ one has:

$$\mathrm{Wass}_2(\nu, \nu_L) \leq \left(\sum_{p\in S} \lambda_p \|p - \pi(p)\|^2\right)^{1/2} \tag{6}$$

3.2. **Algorithm.** We are given the source probability measure μ, and a decomposition $(\nu_\ell)_{0\leq\ell\leq L}$ with L levels of the target measure. The goal is to use the solution of the optimal transport problem from μ to $\nu_{\ell+1}$ at level $\ell+1$ to construct the initial weight vector for the optimal transport problem between μ and ν_ℓ at level ℓ. As before, we will consider the weight vectors at level ℓ as functions from the support S_ℓ of μ_ℓ to $\mathbb{R}$. The function Φ_ℓ that we optimize at step ℓ is given by the same formula as in Eq. (2).

The multiscale descent algorithm is summarized in Algorithm 1. Note that the algorithm does not depend on the choice of the convex optimization scheme (L-BFGS), which we will discuss later in Section 4.

Algorithm 1 Multiscale minimization of $\Phi := \Phi_0$

```
w_L := 0
for ℓ = L − 1 to 0 do
  set w_{ℓ,0}(p) := w_{ℓ+1}(π_ℓ(p)) for every p ∈ S_ℓ
  k := 0
  repeat
    compute w_{ℓ,k+1} from w_{ℓ,k} using L-BFGS on Φ_ℓ
    set v_{k+1} := ∇Φ_ℓ(w_{ℓ,k+1}), k := k + 1
  until ‖v_k‖_q > ε
  set w_ℓ := w_{ℓ,k}
end for
```

In the stopping condition $\|.\|_q$ denotes the usual L^q-norm where $q > 1$ or $q = +\infty$. In particular,

$$\|\nabla\Phi(w)\|_\infty = \sup_{p\in S} |\lambda_p - \mu(\mathrm{Vor}_S^w(p))|$$
$$\|\nabla\Phi(w)\|_1 = \sum_{p\in S} |\lambda_p - \mu(\mathrm{Vor}_S^w(p))|$$

The first quantity measures the maximum error that has been made by considering the weight vector w instead of the optimal one. In particular, if $\|\nabla\Phi(w)\| \geq \min_{p\in S}\lambda_p$, then one is sure that all the cells in the power diagram of (S,w) intersect μ non-trivially. This is important especially for visualization purpose, as the existence of cells with zero μ-mass lead to black spots in the interpolated pictures (see Section 5.3). The choice of $\|.\|_1$ plays a different role which we describe in the next paragraph.

3.3. **Computation of Wasserstein distances.** Simple lower and upper bounds on Wasserstein distance can be obtained at every step of the multiscale algorithm, using the fact that $\|\nabla\Phi(w)\|_1$ corresponds to the twice amount mass that has been misaffected. This follows from the following proposition:

Proposition 1. Let w be the a weight vector on S, and consider the image measure $\tilde{\nu} := T_{S,w}|_\# \mu$. Then,

$$\mathrm{Wass}_2(\nu,\tilde{\nu}) \leq D \times \|\nabla\Phi(w)\|_1^{2/1},$$

where D is the diameter of the support S of ν.

Proof. By definition, both this measure $\tilde{\nu}$ and the target measure ν are supported on the same set S. Moreover,

$$m := \|\nabla\Phi(w)\|_1 = \sum_{p\in S} \left| \lambda_p - \int_{\mathrm{Vor}_S^w(p)} \rho(x)\mathrm{d}x \right|$$

corresponds to the amount of mass that has been mistransported. The cost of transporting this mass back at the right place in S is at most $\sqrt{mD^2}$, where $D = \mathrm{diam}(S)$. □

As a consequence of this proposition, stopping Algorithm 1 at level ℓ with weight vector w_ℓ yields the following estimation of $\mathrm{Wass}_2(\mu,\nu)$:

$$\left|\mathrm{Wass}_2(\mu,\nu) - \left(\int_{\mathbb{R}^d} \|x - T_{S_\ell,w_\ell}(x)\|^2 \rho(x)\mathrm{d}x\right)^{1/2}\right| \tag{7}$$
$$\leq D \times \|\nabla\Phi_\ell(w_\ell)\|_1^{1/2} + \mathrm{Wass}_2(\nu,\nu_\ell)$$

Said otherwise, if one wants to compute the Wasserstein distance between μ and ν up to a certain error ε, it is not necessary to consider the levels of the decomposition below the first level ℓ_0 such that $\mathrm{Wass}_2(\nu_{\ell_0},\nu) < \varepsilon$. Note that this quantity can be estimated thanks to Eq. (6). The effectiveness of this approach is discussed in Section 5.2.

Proof. of Eq. (7) By Theorem 1, the map T_{S_ℓ,w_ℓ} is an optimal transport between the measure μ and $\tilde{\nu} = T_{S_\ell,w_\ell}|_\# \mu$. This means that the Wasserstein distance $\mathrm{Wass}_2(\mu,\tilde{\nu})$ is equal to c_ℓ. Moreover, by the reverse triangle inequality,

$$|\,\mathrm{Wass}_2(\mu,\nu) - \mathrm{Wass}_2(\mu,\tilde{\nu})| \leq \mathrm{Wass}_2(\tilde{\nu},\nu_\ell) + \mathrm{Wass}_2(\nu_\ell,\nu).$$

One then concludes using the previous proposition. □

3.4. **Convergence of Optimal Transport Maps.** In this paragraph, we discuss the soundness of constructing an initial weight vector for the optimal transport problem (μ,ν_ℓ) from an adapted weight vector for the problem $(\mu,\nu_{\ell+1})$. The result of this section are summarized in the following theorem. Note that the definition of zero-mean convex potential is given below, and is necessary to define uniquely the adapted weight vector: without this, adapted weight vectors are defined up to an additive constant.

Theorem 3. Let ν and $(\nu_n)_{n\geq 1}$ be discrete probability measures supported on finite sets S and $(S_n)_{n\geq 1}$ respectively, such that $\lim_n \mathrm{Wass}_2(\nu,\nu_n) = 0$. Let:

$$w_n : S_n \to \mathbb{R} \text{ be adapted to } (\mu,\nu_n)$$
$$w : S \to \mathbb{R} \text{ be adapted to } (\mu,\nu)$$

Suppose that both weight vectors yield zero-mean convex potentials (see below), and that the assumptions of Proposition 2 are satisfied. Then, for every sequence of points $p_n \in S_n$ converging to a point p in S, one has $w(p) = \lim w_n(p_n)$.

Before proving this theorem, we need to introduce a few definitions and auxiliary results.

Convex potential. Let ν be a probability measure supported on a finite set S, and let w denote the weight vector adapted to the optimal transport problem between μ and ν. Set

$$\phi_S^w(x) := \frac{1}{2}\left(\|x\|^2 - \min_{p\in S} \|x-p\|^2 - w(p)\right) \tag{8}$$
$$= \max_{p\in S} \langle x|p\rangle + \frac{1}{2}(w(p) - \|p\|^2) \tag{9}$$

where $\langle v|w\rangle = \sum_i v_i w_i$ denotes the usual Euclidean scalar product. From these two formulations, it is easy to see that the function ϕ_S^w is convex, and that its gradient $\nabla\phi_S^w$ coincides with the transport map T_S^w. We call such a function a *convex potential* for the optimal transport plan. Since adding a constant to the

weight vector (or to ϕ_S^w) does not change the transport plan, we consider the *zero-mean convex potential* which is uniquely defined by the extra assumption that $\int_{\mathbb{R}^d} \phi_S^w(x)\rho(x)\mathrm{d}x = 0$.

Proposition 2. Let μ and ν be two probability measures, μ having density ρ and ν supported on a finite set S. Let (ν_n) be a sequence of probability measures supported on finite sets (S_n), s.t. $\lim_n \mathrm{Wass}_2(\nu_n, \nu) = 0$. Assume that:

(i) the support of ρ is the closure of a connected open set Ω with regular (piecewise $\mathcal{C}^1$) boundary ;
(ii) there exists a positive constant m such that $\rho \geq m$ on Ω;
(iii) the support of all the measures ν_n is contained in a fixed ball $\mathrm{B}(0, L)$;

Denote $\phi := \phi_S^w$ (resp. $\phi_n := \phi_{S_n}^{w_n}$) the zero-mean convex potential of the optimal transport between μ and ν (resp. ν_n). Then, ϕ_n converges to ϕ uniformly on Ω as n grows to infinity.

This proposition is similar to [Bos10, Theorem 1], but without the requirement that the source and target measure have to be supported on convex sets. It relies on the following result (cf [Vil09], Corollary 5.23):

Fact. *Let ν_n be a sequence of measures converging to μ, and T_n (resp. T) denotes the optimal optimal transport map between μ and ν_n (resp. ν). Then, for every positive ε,*

$$\lim_{n\to+\infty} \mu(\Delta_\varepsilon(T, T_n)) = 0 \tag{10}$$

where $\Delta_\varepsilon(T, T_n) := \{x \in \mathbb{R}^d;\ \|T(x) - T_n(x)\| \geq \varepsilon\}$.

Proof. of Proposition 2. For almost every x in Ω, the gradient $\nabla\phi_n(x) = T_n(x)$ is included in the support of ν_n, hence in the ball $\mathrm{B}(0, L)$ by (iii). The same holds for T, so that the inequality $\|T - T_n\| \leq 2L$ holds for almost every x in Ω. For every $p \geq 1$,

$$\begin{aligned}
\|T - T_n\|_{\mathrm{L}^p(\mu)} &:= \int_\Omega \|T(x) - T_n(x)\|^p \rho(x)\mathrm{d}x \\
&= \int_{\Omega\setminus\Delta_\varepsilon(T,T_n)} \|T(x) - T_n(x)\|^p \rho(x)\mathrm{d}x \\
&\quad + \int_{\Delta_\varepsilon(T,T_n)} \|T(x) - T_n(x)\|^p \rho(x)\mathrm{d}x \\
&\leq \varepsilon^p + (2L)^p \mu(\Delta_\varepsilon(T, T_n))
\end{aligned}$$

Using Eq. (10), we obtain the convergence of T_n to T in the $\mathrm{L}^p(\mu)$-sense. Thanks to the the assumptions (i) and (ii), we can apply the Poincaré inequality on the domain (Ω, μ) to the zero-mean potentials ϕ_n and ϕ to get:

$$\begin{aligned}
\|\phi_n - \phi\|_{\mathrm{L}^p(\mu)} &:= \int_{\mathbb{R}^d} \|\phi_n(x) - \phi(x)\| \rho(x)\mathrm{d}x \\
&\leq \mathrm{const}(\Omega, \rho) \|T_n - T\|_{\mathrm{L}^p(\mu)}.
\end{aligned}$$

In other words, ϕ_n converges to ϕ in the $\mathrm{L}^p(\mu)$-sense. Since the support of all target measures is contained in a ball of size L, $\|\nabla\phi_n\| \leq L$, and ϕ_S^w is L-Lipschitz. Hence, ϕ_n also converges uniformly to ϕ on the support of μ. □

Proof. of Theorem 3. We begin by proving that there for every point p in S, there exists a sequence of points $q_n \in S_n$ converging to p such that $w_n(q_n)$ also converges to $w(p)$. Applying Eq. (10) with ε equal to half the minimum distance between two points in S ensures that T_n converges to T on a set F with full Lebesgue measure in

Ω. Choose a point x in the intersection of the cell $\mathrm{Vor}_S^w(x)$ and of F, and consider the sequence $q_n = T_n(p)$. This sequence converges to p, and by definition one has :

$$\phi_n(p_n) - \phi(p) = \frac{1}{2}(\|x-p\|^2 - \|x-p_n\|^2 + w(p) - w(p_n))$$

Using the uniform convergence of ϕ_n to ϕ, one deduces that $w(q_n)$ converges to $w(p)$.

We now prove by contradiction that if (p_n) converges to p, then $\limsup w_n(p_n)$ is at most $w(p)$. Suppose not: taking subsequence if necessary, the limit of $w_n(p_n)$ is larger than $w(p)$ by a positive η. For every point x in Ω, we use the triangle inequality to get

$$\|x-p_n\|^2 - w_n(p_n) \leq \|x-q_n\|^2 - w_n(q_n) + r_n \tag{11}$$
$$\text{with } r_n := \|q_n - p_n\|^2 + 2D\|p_n - q_n\| + w_n(q_n) - w_n(p_n)$$

and D is the maximum distance between a point in Ω and a point in the ball $\mathrm{B}(0,L)$ defined by the assumption (iii). Using the convergence of (p_n) and (q_n) to the same point p, and the assumption on the limits of $w_n(p_n)$ and $w_n(q_n)$, we obtain $\lim_{n\to+\infty} r_n \leq -\eta$. Combining this with Eq. (11) shows that the cell of p_n in the power diagram $\mathrm{Vor}_{S_n}^{w_n}$ does not intersect Ω for n large enough. This contradicts the hypothesis that p_n is a Dirac with positive mass in the support of ν_n.

The proof that $\liminf w_n(p_n)$ is larger than $w(p)$ is very similar but a bit longer, as it requires the use of the zero-mean assumption for the convex potentials. These two bounds for $\liminf / \sup w_n(p_n)$ conclude the proof. □

4. Implementation

In the first paragraph, we give some details our implementation of the convex optimization method proposed in [AHA98] for a fixed target measure. Then, we explain how we compute the hierarchical decomposition of the target measure needed for the multiscale algorithm.

4.1. For a fixed target measure.

Solving optimal transport between a probability measure μ with density ρ and a discrete measure $\nu = \sum_{p\in S} \lambda_p \delta_p$ amounts to finding the minimum of the convex function Φ given in Theorem 2.(iii):

$$\Phi(w) = \sum_{p\in S} \left(\lambda_p w(p) - \int_{\mathrm{Vor}_S^w(p)} (\|x-p\|^2 - w(p))\rho(x)\mathrm{d}x \right)$$
$$\text{with } \frac{\partial \Phi}{\partial w(p)}(w) = \lambda_p - \int_{\mathrm{Vor}_S^w(p)} \rho(x)\mathrm{d}x$$

We need three ingredients to achieve this goal: an efficient and robust implementation of power diagram computation, robust numerical integration functions, and convex optimization software. In this paragraph, we discuss and motivate our choices regarding these three aspects.

Power diagrams. We use the Regular_triangulation_2 package from CGAL [cga]. It is tempting to try to avoid recomputing the whole power diagram for every evaluation of the function Φ by using the same approach that was used in [MMdCTAD09] to maintain the Delaunay triangulation. However, as shown in Figure 3(a), the topology of the power diagram keeps changing until the very last steps of the optimization, thus discarding this approach.

Numerical integration. In our C++ implementation, a measure μ with density ρ is represented by an object which can answer the following two queries. Given

a convex polygon $P = [a_0, \dots, a_N = a_0]$, and a function f from P to $\mathbb{R}$, the class should provide a way to compute:

(1) the **mass** of P, i.e. $\int_P \rho(x)\mathrm{d}x$;
(2) the **integral** of f over P, i.e. $\int_P f(x)\rho(x)\mathrm{d}x$.

In practice, we only use it the second query for the functions $f : x \mapsto \|x - x_0\|^2$. We developped two different models of measure with density.

The first one is the uniform measure on a convex polygon R. In this case, computing the mass of a polygon P amounts to computing the area of the intersection $P \cap R$ of two convex polygons. The integral of the squared distance function $x \mapsto \|x-x_0\|^2$ over the polygon $P \cap R$ is computed by triangulating P and summing the integral over each triangle T. The integral on T can be obtained in closed-form: if one denotes by $\mathrm{cov}(T, x_0)$ the covariance matrix of T with base point x_0, then

$$\int_T \|x - x_0\|^2 \mathrm{d}x = \mathrm{cov}(T, x_0)_{11} + \mathrm{cov}(T, x_0)_{22}$$

The second model corresponds to the density obtained from a grayscale image. We assume that the density ρ is constant on each square pixel $p_{i,j} = [i, i+1) \times [j, j+1)$, equal to the value $a_{i,j}$. We then consider:

$$\int_P \rho(x)\mathrm{d}x = \sum_{i,j} a_{i,j}\, \mathrm{area}(P \cap p_{i,j}) \tag{12}$$

$$\int_P f(x)\rho(x)\mathrm{d}x \simeq \sum_{i,j} a_{i,j}\, \mathrm{area}(P \cap p_{i,j}) f(i,j) \tag{13}$$

Note that it is not possible to simply replace the area of $P \cap p_{i,j}$ by zero or one depending on whether P intersects $p_{i,j}$ or not, thus disallowing a more efficient GPU implementation. However, since the area of $P \cap p_{i,j}$ needs to be computed only for pixels containing edges or vertices of P the algorithm we use remains rather efficient. Pixels on edges are dealt with while applying Bresenham's algorithm to raster the polygon. The coverage of pixels containing vertices of P is obtained simply by computing the intersection of the polygon P with the square $p_{i,j}$.

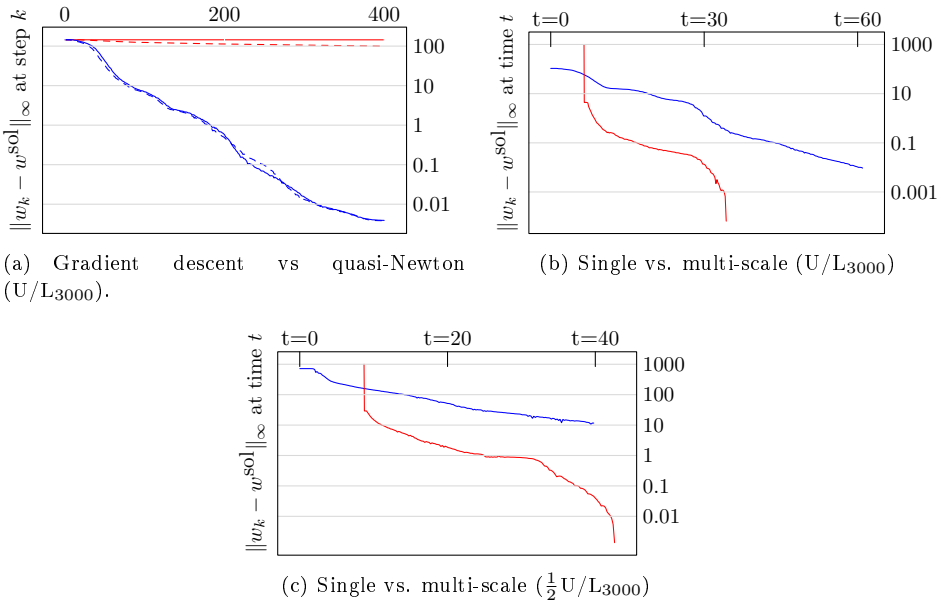
Convex optimization. We tried several approaches for the actual convex optimization. All of these methods use the following rough scheme to construct the sequence of weight vectors (w_k):

(i) Determine a descent direction d_k ;
(ii) Determine a timestep s_k, and set $w_{k+1} = w_k + s_k d_k$.

Methods to choose the descent direction d_k include gradient methods, where d_k is simply $-\nabla\Phi(w_k)$, Newton methods for which $d_k = -[\mathrm{D}^2\Phi(w_k)]^{-1}\nabla\Phi(w_k))$ and quasi-Newton methods. In quasi-Newton methods $\mathrm{D}^2\Phi(w_k)$ is not computed exactly, but estimated from previous evaluations of the gradients. We chose the widely used low-storage version of the BFGS scheme [Fle87], implemented in C in libLBFGS.

The timestep s_k is determined by a search along the line starting from w_k with direction d_k. Here again, the literature is very vast, as there is a trade-off between finding a good step size (the best choice would be to minimize the function $s \mapsto \Phi(w_k + sd_k)$) and requiring as few functions evaluations as possible — recall that in our case a function evaluation requires the construction of a complete Power diagram!

Figure 2.(a) shows that gradient descent methods are outperformed by quasi-Newton ones, regardless of the choice of line search. It also shows that the choice of line search method is not as important — barring the fixed-step scheme. For all



(a) Gradient descent vs quasi-Newton ($\mathrm{U/L}_{3000}$).

(b) Single vs. multi-scale ($\mathrm{U/L}_{3000}$)

(c) Single vs. multi-scale ($\frac{1}{2}\mathrm{U/L}_{3000}$)

FIGURE 2. Speed of convergence, measured by the L^∞ distance between the weight vector at a given time/step and the optimal one. (a) Comparison of simple convex optimization algorithms: gradient descent (red) with fixed step (solid) or strong Wolfe line search (dashed), and low-storage BFGS algorithm (blue) with strong Wolfe (solid) or Moré-Thuente line-search (dashed). (b) and (c) Comparison between the original algorithm of [AHA98] (red) and the multiscale one (blue).

remaining experiments, we use the low-storage BFGS method with Moré-Thuente line search [MT94].

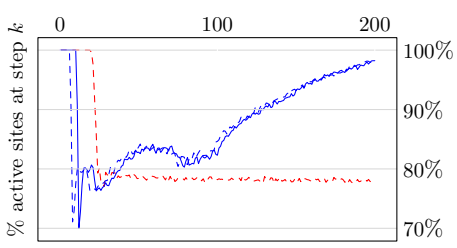
4.2. **Decomposition of the target measure.** Suppose for now that the measure ν is discrete; we will explain in the next paragraph how to convert an image to such a measure. From this measure, we construct a sequence of discrete probability measures (ν_ℓ), with

$$\nu_\ell = \sum_{p \in S_\ell} \lambda_{p,\ell} \delta_p$$

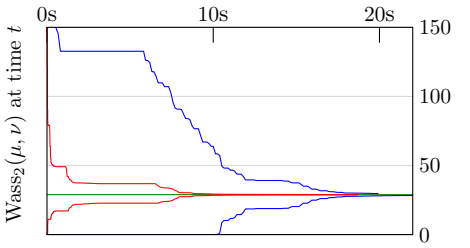
such that $\nu_0 = \nu$, and that the number of points of the support of ν_ℓ decreases as ℓ increases. The parameters of our algorithm are the number L of levels in the decomposition, and for each level ℓ, the number of points $n(\ell)$ in the support of the measure ν_ℓ. In practice, we found that choosing $n(\ell) = n(0)/k^\ell$ with $k = 5$ usually provides good results.

Lloyd's algorithm. Theorem 3 suggests that if we want to be able to construct a good initial weight vector for the problem (μ, ν_ℓ) from a weight vector adapted to $(\mu, \nu_{\ell+1})$ we need to have $\nu_{\ell+1}$ as close as possible to ν_ℓ in the Wasserstein sense. Given the constraints that $\nu_{\ell+1}$ is supported on $n(\ell+1)$ points, this means

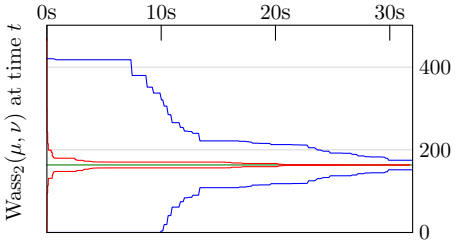
$$\nu_{\ell+1} \in \arg\min\{\mathrm{Wass}_2(\bar{\nu}, \nu_\ell);\ |\mathrm{spt}(\bar{\nu})| \leq n(\ell+1)\}.$$



(a) Number of active sites (U/L$_{3000}$)



(b) Bounds on Wasserstein distance (U/L$_{3000}$)



(c) Bounds on Wasserstein distance ($\frac{1}{2}$U/L$_{3000}$)

FIGURE 3. (a) Percentage of points in the support of the target measure whose Power cell intersects the support of the source measure during the execution of various convex optimization algorithms (colors are the same as in Fig. 2(a)). (b) and (c) Estimation of Wasserstein distance: in red (resp. blue), the lower and upper bounds obtained by the multiscale (resp. original) algorithm as a function of time, and in green the correct value.

This minimization problem is equivalent to a weighted k-means problem, with $k = n(\ell+1)$. Since it is hopeless to solve this problem exactly, we use the standard Lloyd's iterative algorithm to find a good local minimum.

We initialize the algorithm using a random sample $S^0_{\ell+1}$ of $n(\ell+1)$ points drawn independently from ν_ℓ. We then apply Lloyd's descent step to $S^n_{\ell+1}$ to obtain $S^{n+1}_{\ell+1}$, stopping when the points do not move more than a given threshold between two successive steps. This procedure provides us with the support $S_{\ell+1}$ of our measure. We define π_ℓ to be the application which maps a point p in S_ℓ to its nearest neighbor in $S_{\ell+1}$. The values of $(\lambda_{p,\ell+1})_{p\in S_{\ell+1}}$ are defined by Eq. (5).

Initial quantization of the target measure. Often, the target measure is not a discrete measure but a measure ν_I with density $\sigma : \Omega \to \mathbb{R}$ (such as a grayscale image). In this case we apply Lloyd's algorithm to the measure ρ' in order to obtain an initial quantization $\nu = \sum_{p\in S} \lambda_p \delta_p$ of the original measure ν_I with a prescribed number of points N.

5. RESULTS

We will use the following datasets in our experiments. We denote by λU the uniform probability measure on the square $\lambda S = [0, \lambda 512] \times [0, \lambda 512]$. For $\lambda = 1$ we will simply write U and S. By L, we denote is the standard grayscale picture of Lena on the square S. Given a measure with density D, we will denote by D_N a quantization of this measure with N points, obtained using Lloyd's algorithm. The decomposition of measures we work with are all obtained with the same parameters:

source/target	original	multiscale	speedup
$\frac{1}{2}U$ / U_{10000}	577s	143s	4.0
$\frac{1}{4}U$ / U_{10000}	1180s	189s	6.2
$\frac{1}{8}U$ / U_{10000}	1844s	241s	7.6
U / L_{10000}	216s	52s	4.1

TABLE 1. Running time of the original and multiscale algorithm to find a weight vector such that $\|\nabla\Phi(w)\|_\infty < 10^{-6}$.

5 levels in the decomposition (including the original one), and level ℓ being made of $N/5^\ell$ Dirac masses.

5.1. **Comparisons with the original approach.** In Figure 2(b) and 2(c) we show the evolution of the $\|.\|_\infty$ distance between the weight vector obtained at a given time, and the optimal one w^{sol}. This optimal weight vector had been previously obtained by running the algorithm with a target accuracy of $\|\nabla\Phi(w)\|_\infty < 10^{-9}$.

The advantage of our multiscale method over the original convex optimization is especially important when the source and target measure are far from each other. Table 1 compares the running time of the original and multiresolution algorithms to compute a weight vector adapted to the problem of optimally transporting λU to U_{1000} with a given accuracy $\|\nabla\Phi(w)\|_\infty < \varepsilon$. The speedup increases as λ goes to zero, i.e. as the measure λU becomes more concentrated around the lower-left corner of the original square S.

5.2. **Computation of Wasserstein distances.** We use the approach described in Section 3.3 to obtain lower and upper bounds on the Wasserstein distance between μ and ν at every step of the algorithm. Figure 3(b) and 3(c) compare the evolution of these two bounds as a function of the runtime of the original and the multiscale algorithm.

5.3. **Displacement interpolation of images.** The concept of displacement interpolation of two probability measures was introduced in [McC97]. It uses optimal transport maps as a replacement for the linear interpolation $\mu_t = (1-t)\mu + t\nu$. Displacement interpolation can be a useful tool for the interpolation of grayscale image, when the gray value of a pixel can be interpreted as a density of some quantity (e.g. satellite views of clouds, preprocessed so that the gray level ranges from black to white depending on the thickness of the cloud). We make use of the transport map computed using the multiscale algorithm. Recall that in order to apply this algorithm to a target measure with density $\sigma : \Omega \to \mathbb{R}$, we had to compute a first quantization of σ, $\nu = \sum_{p\in S} \lambda_p \delta_p$ using Lloyd's algorithm. By construction of ν, and by definition of the optimal weight vector ω, one has for every point p in S

$$\int_{\mathrm{Vor}_S(p)\cap\Omega} \sigma(x)\mathrm{d}x = \lambda_p = \int_{\mathrm{Vor}_{S,w}(p)\cap\Omega} \rho(x)\mathrm{d}x.$$

This suggests a way to construct an interpolation between σ and ρ. Given a time t, consider the weight vector $w_t = tw$, and the corresponding Power diagram (Vor_{S,w_t}). Now, we define the interpolant ρ_t at time t as the only piecewise-constant function ρ_t on Ω obtained by spreading the mass of λ_p on the intersection of the cell $\mathrm{Vor}_{S,w_t}(p)$ with Ω, i.e. for every point x in $\mathrm{Vor}_{S,w_t}(p)$, define $\rho_t(x) := \lambda_p / \operatorname{area}(\mathrm{Vor}_{S,w_t}(p))$. An example of this interpolation is presented in Figure 4.

FIGURE 4. First and second rows: An interpolation between a picture of G. Monge and photograph of B. Riemann (with N =625 and 15k respectively). The intermediary steps are obtained using McCann's displacement interpolation [McC97] of the two corresponding measures, which can be computed from the L^2 optimal transport.

6. Discussion

In this paper we have presented a simple way to increase the efficiency of the convex optimization algorithm introduced in [AHA98] to solve the optimal transport problem. We also discussed how our multiscale approach can be used to obtain fast estimation of Wasserstein distances between images.

This first step suggests that, in order to obtain faster computations of optimal transport, one has to better understand the geometry of the function Φ. For instance, it is currently not possible to obtain complexity estimates for this approach because: (i) nothing is known about the shape and size of the basin around the minimizer where Newton's method has quadratic convergence and (ii) the stability result (Theorem 3) is not quantitative. Understanding these two problems could open the way to even more efficient computations of optimal transport maps.

We also believe that this multiscale approach can be useful in the solution of more geometric problems with a similar structure. An example of such a problem is Minkowski's problem: given a set of normals $\vec{n}_1, \dots, \vec{n}_N$ and a set of areas $\lambda_1, \dots, \lambda_N$ such that $\sum_i \lambda_i \vec{n}_i$ vanishes, find a convex polytope whose facets normals are among the $(\vec{n}_i)$, and such that the facet corresponding to $\vec{n}_i$ has an area of exactly λ_i. This problem has a similar multiscale structure as optimal transport, and can be also solved by minimizing a convex functional [LRO06], and would probably benefit from a multiscale approach. A second example is the problem of designing a reflector antenna with prescribed image measure at infinity, which can also be formally cast as an optimal transport problem (Section 4.2.5 in [Oli03]).

Acknowledgements. ANR grant GIGA ANR-09-BLAN-0331-01 and Université Grenoble I MSTIC grant GEONOR.

References

[AHA98] F. Aurenhammer, F. Hoffmann, and B. Aronov, *Minkowski-type theorems and least-squares clustering*, Algorithmica **20** (1998), no. 1, 61–76.

[AHT03] S. Angenent, S. Haker, and A. Tannenbaum, *Minimizing flows for the monge-kantorovich problem*, SIAM journal on mathematical analysis **35** (2003), no. 1, 61–97.

[BB00] J.D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numerische Mathematik **84** (2000), no. 3, 375–393.

[Ber88] D.P. Bertsekas, *The auction algorithm: A distributed relaxation method for the assignment problem*, Annals of Operations Research **14** (1988), no. 1, 105–123.

[Bos10] D. Bosc, *Numerical approximation of optimal transport maps*, Preprint, 2010.

[Bre91] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Communications on pure and applied mathematics **44** (1991), no. 4, 375–417.

[BSD09] Michael Balzer, Thomas Schlömer, and Oliver Deussen, *Capacity-constrained point distributions: a variant of Lloyd's method*, ACM Trans. Graph. **28** (2009), 86:1–86:8.

[CCSM10] F. Chazal, D. Cohen-Steiner, and Q. Mérigot, *Geometric inference for probability measures*, Foundation of Computational Mathematics (2010), to appear.

[cga] CGAL, *Computational Geometry Algorithms Library*, http://www.cgal.org.

[dGCSAD11] F. de Goes, D. Cohen-Steiner, P. Alliez, and M. Desbrun., *An optimal transport approach to robust reconstruction and simplification of 2D shapes*, Preprint, 2011.

[DSS10] J. Delon, J. Salomon, and A. Sobolevski, *Fast Transport Optimization for Monge Costs on the Circle*, SIAM Journal on Applied Mathematics **70** (2010), no. 7, 2239–2258.

[Fle87] R. Fletcher, *Practical methods of optimization*, John Wiley & Sons, 1987.

[LD11] Y. Lipman and I. Daubechies, *Conformal Wasserstein distances: comparing surfaces in polynomial time*, Advances in Mathematics (2011).

[LR05] G. Loeper and F. Rapetti, *Numerical solution of the Monge-Ampère equation by a Newton's algorithm*, Comptes Rendus Mathematique **340** (2005), no. 4, 319–324.

[LRO06] T. Lachand-Robert and É. Oudet, *Minimizing within convex bodies using a convex hull method*, SIAM Journal on Optimization **16** (2006), no. 2, 368–379.

[McC97] R.J. McCann, *A Convexity Principle for Interacting Gases*, Advances in Mathematics **128** (1997), no. 1, 153–179.

[MDGD+10] P. Mullen, F. De Goes, M. Desbrun, D. Cohen-Steiner, and P. Alliez, *Signing the Unsigned: Robust Surface Reconstruction from Raw Pointsets*, Computer Graphics Forum **29** (2010), no. 5, 1733–1741.

[MMD11] P. Mullen, F. Memari, De Goes, and M. Desbrun, *Hodge-Optimized Triangulations*, Proceedings of ACM SIGGRAPH 2011, 2011.

[MMdCTAD09] P. Machado Manhães de Castro, J. Tournois, P. Alliez, and O. Devillers, *Filtering relocations on a Delaunay triangulation*, Computer Graphics Forum, vol. 28, Wiley Online Library, 2009, pp. 1465–1474.

[Mon81] G. Monge, *Mémoire sur la théorie des déblais et de remblais*, Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique etde Physique pour la même année, 1781, pp. 666–704.

[MT94] J.J. Moré and D.J. Thuente, *Line search algorithms with guaranteed sufficient decrease*, ACM Transactions on Mathematical Software (TOMS) **20** (1994), no. 3, 286–307.

[Oli03] V. Oliker, *Mathematical aspects of design of beam shaping surfaces in geometrical optics*, Trends in Nonlinear Analysis, Springer Verlag, 2003, p. 193.

[RTG00] Y. Rubner, C. Tomasi, and L.J. Guibas, *The earth mover's distance as a metric for image retrieval*, International Journal of Computer Vision **40** (2000), no. 2, 99–121.

[Vil09] C. Villani, *Optimal transport: old and new*, Springer Verlag, 2009.