Semi-discrete optimal transport — the case $p = 1$

Valentin Hartmann* and Dominic Schuhmacher
University of Goettingen
June 22, 2017

Abstract

The Monge–Kantorovich problem of finding a minimum cost coupling between an absolutely continuous measure $\mu$ on $\mathcal{X} \subset \mathbb{R}^d$ and a finitely supported measure $\nu$ on $\mathbb{R}^d$ is considered for the special case of the Euclidean cost function. This corresponds to the natural problem of closest distance allocation of some resource that is continuously distributed in space to a finite number of processors with capacity constraints.

A systematic discussion of the choice of Euclidean cost versus squared Euclidean cost is provided from the practitioner’s point of view. We then show that the above problem has a unique solution that can be either described by a transport map $T: \mathcal{X} \to \mathbb{R}^d$ or by a partition of $\mathcal{X}$. We provide an algorithm for computing this optimal transport partition, adapting the approach by Aurenhammer, Hoffmann and Aronov (1998) and Mérigot (2011) to the Euclidean cost. We give two types of numerical applications to demonstrate the use of the algorithm and to illustrate aspects of the semi-discrete problem and the choice of the Euclidean cost.

MSC 2010: Primary 65D18; Secondary 51N20, 90B06.

Key words and phrases: Monge–Kantorovich problem; spatial resource allocation; Wasserstein metric; weighted Voronoi tessellation.

1 Introduction

Optimal transport and Wasserstein metrics are nowadays among the major tools for analyzing complex data. Theoretical advances in the last decades characterize existence, uniqueness, representation and smoothness properties of optimal transport plans in a variety of different settings. Recent algorithmic advances make it possible to compute exact transport plans and Wasserstein distances between discrete measures on regular grids of tens of thousands of support points, see e.g. [31, Section 6], and to approximate such distances (to some extent) on larger and/or irregular structures, see [1] and references therein. Applications are abundant throughout all of the applied sciences, including biomedical sciences (e.g. microscopy or tomography images [5, 14]), geography (e.g. remote sensing [10, 15]), and computer science (e.g. image processing and computer graphics [24, 35]). In brief: whenever data of a sufficiently complex structure that can be thought of as a mass distribution is available, optimal transport offers an effective, intuitively reasonable and robust tool for analysis.

*Work partially supported by DFG RTG 2088.
More formally, for measures $\mu$ and $\nu$ on $\mathbb{R}^d$ with $\mu(\mathbb{R}^d) = \nu(\mathbb{R}^d) < \infty$ the Wasserstein distance of order $p \geq 1$ is defined as

$$W_p(\mu, \nu) = \left( \min_{\pi} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \pi(dx, dy) \right)^{1/p},$$

where the minimum is taken over all transport plans (couplings) $\pi$ between $\mu$ and $\nu$, i.e. measures $\pi$ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals

$$\pi(A \times \mathbb{R}^d) = \mu(A) \quad \text{and} \quad \pi(\mathbb{R}^d \times A) = \nu(A)$$

for every Borel set $A \subset \mathbb{R}^d$. The minimum exists by [36, Theorem 4.1] and it is readily verified, see e.g. [36, after Example 6.3], that the map $W_p$ is a $[0, \infty)$-valued metric on the space of measures with fixed finite mass. The constraint linear minimization problem \([1]\) is known as Monge–Kantorovich problem [17, 36]. From an intuitive point of view, a minimizing $\pi$ describes how the mass of $\mu$ is to be associated with the mass of $\nu$ in order to make the overall transport cost minimal.

A transport map from $\mu$ to $\nu$ is a measurable map $T: \mathbb{R}^d \to \mathbb{R}^d$ satisfying $T_\#\mu = \nu$, where $T_\#$ denotes the push-forward, i.e. $(T_\#\mu)(A) = \mu(T^{-1}(A))$ for every Borel set $A \subset \mathbb{R}^d$. We say that $T$ induces the coupling $\pi = \pi_T$ if

$$\pi_T(A \times B) = \mu(A \cap T^{-1}(B))$$

for all Borel sets $A, B \subset \mathbb{R}^d$, and call the coupling $\pi$ deterministic in that case. It is easily seen that the support of $\pi_T$ is contained in the graph of $T$. Intuitively speaking, we associate with each location in the domain of the measure $\mu$ exactly one location in the domain of the measure $\nu$ to which positive mass is moved, i.e. the mass of $\mu$ is not split.

The generally more difficult (non-linear) problem of finding (the $p$-th root of)

$$\inf_T \int_{\mathbb{R}^d} \|x - T(x)\|^p \mu(dx) = \inf_T \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \pi_T(dx, dy),$$

where the infima are taken over all transport maps $T$ from $\mu$ to $\nu$ (and are in general not attained) is known as Mongè’s problem [23, 36].

In practical applications, based on discrete measurement and/or storage procedures, we often face discrete measures $\mu = \sum_{i=1}^m \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$, where $\{x_1, \ldots, x_m\}$, $\{y_1, \ldots, y_n\}$ are finite collections of support points, e.g. grids of pixel centers in a grayscale image. The Monge–Kantorovich problem \([1]\) is then simply the discrete transport problem from classical linear programming [20]:

$$W_p(\mu, \nu) = \left( \min_{(\pi_{ij})} \sum_{i=1}^m \sum_{j=1}^n d_{ij} \pi_{ij} \right)^{1/p},$$

where $d_{ij} = \|x_i - y_j\|^p$ and any measure $\pi = \sum_{i=1}^m \sum_{j=1}^n \pi_{ij} \delta_{(x_i, y_j)}$ is represented by the $m \times n$ matrix $(\pi_{ij})_{i,j}$ with nonnegative entries $\pi_{ij}$ satisfying

$$\sum_{j=1}^n \pi_{ij} = \mu_i \quad \text{and} \quad \sum_{i=1}^m \pi_{ij} = \nu_j.$$

Due to the sheer size of $m$ and $n$ in typical applications this is still computationally a very challenging problem; we have e.g. $m = n = 10^6$ for $1000 \times 1000$ grayscale images, which is far beyond the performance of a standard transportation simplex or primal-dual
algorithm. Recently many dedicated algorithms have been developed, such as [31], which
can give enormous speed-ups mainly if \( p = 2 \) and can compute exact solutions for discrete
transportation problems with \( 10^5 \) support points in seconds to a few minutes, but still
cannot deal with \( 10^6 \) or more points.

The main advantage of using \( p = 2 \) is that we can decompose the cost function as
\[
\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2x'y
\]
and hence formulate the Monge–Kantorovich problem equivalently as (the square root of twice)
\[
\max_{\pi} \int_{\mathbb{R}^d \times \mathbb{R}^d} x'y \pi(dx,dy).
\]
For the discrete problem [3] this decomposition is used in [31] to construct particularly simple so-called
shielding neighborhoods. But also if one or both of \( \mu \) and \( \nu \) are assumed absolutely
continuous with respect to Lebesgue measure, this decomposition for \( p = 2 \) has clear
computational advantages. For example if the measures \( \mu \) and \( \nu \) are assumed to have
densities \( f \) and \( g \), respectively, the celebrated Brenier’s theorem, which yields an optimal
transport map that is the gradient of a convex function \( u \) [21], allows to solve Monge’s
problem by finding a numerical solution \( u \) to the Monge-Ampère equation
\[
\det(D^2u(x)) = f(x)/g(\nabla u(x));
\]
see [30, Section 6.3] and the references given there.

Another well-known case for \( p = 2 \) is the so-called semi-discrete setting, where the
source measure \( \mu \) is absolutely continuous and the target measure \( \nu \) is discrete with finite
support; see e.g. [37], [19] or [3], where this terminology was recently used. It was shown
in [4] that a solution to Monge’s problem is given by a power diagram, a.k.a. Laguerre
tesselation, which can be found numerically by solving a (typically high dimensional)
unconstrained convex optimization problem [4, 22].

In the present article we investigate theory, algorithm and computational examples in
the semi-discrete setting for the case \( p = 1 \). This setting appears naturally in problems
of allocating some resource that is continuously distributed in space to a finite number of
centers. Suppose for example that a fast-food chain decides to introduce a home delivery
service to its customers. For each potential order the management wants to fix in advance
which branch to commission with the execution in such a way that the expected total
Euclidean travel distance to the customers is minimal. We make here the simplifying
assumption that each branch has a fixed capacity and that the overall capacity matches
the total number of orders (peak time scenario). A continuous distribution of the order
locations could be estimated, e.g. by kernel density estimation, from an existing client data
base in the ideal case or from census data otherwise. For actual applications capacities are
not fixed in the long run of course, and decisions to increase capacities in certain branches
or build new branches in certain areas would be derived based on such assignments. We
take up this example in Section 6.

Apart from the resource allocation case, we may also use the semi-discrete setting as
approximation of both the fully continuous or the fully discrete setting, as we will see in
Section 3.

We would like to emphasize the fact that especially in the semi-discrete setting \( p = 1 \),
i.e. penalizing transport by the actual Euclidean distance that has to be covered, is usually
more intuitive (although computationally more difficult) than \( p = 2 \). This is due to the
fact that certain unpleasant behaviour known for \( p = 1 \) from the fully discrete or fully
continuous setting disappear. Most notably, we will see in Section 3 that the minimal
transport cost is realized by a unique coupling \( \pi \), which is always deterministic, i.e. there
exists a \( \mu \)-a.s. unique transport map \( T^* \) from \( \mu \) to \( \nu \) such that
\[
W_1(\mu, \nu) = \int_{\mathbb{R}^d} \|x - T^*(x)\| \mu(dx).
\]
To some extent we advocate the use of \( W_1 \) over \( W_2 \) also in the fully discrete and continuous
cases, although in applications it often turns out that either choice is fine, as long as some
basic considerations are taken into account.

As pointed out above, the semi-discrete setting is well-understood for \( p = 2 \) and the AHA-algorithm developed in [11, 22] is rather fast. The setting has also been investigated to some extent for general \( p \geq 1 \) in [13]. Compared to the latter article the contributions of the present paper are additional theoretical results for the case \( p = 1 \) (most importantly Theorem 3 and part of Theorem 1), a concrete algorithm that draws on ideas from the algorithm for \( p = 2 \) in [22] and an implementation with numerical examples for this algorithm.

The present paper is a verbally condensed but mathematically extended version of the thesis [16], to which we refer at several points in the text for more detailed information. At the time of finishing the present paper, we noted that Theorem 2.1 of [19], which is for very general cost functions including the Euclidean distance (although the remainder of the paper is not), has a rather large overlap with our Theorem 3 except that it requires \( \mathcal{X} \) to be compact and \( \mu \) to have bounded density. We have retained our original proof, because we believe it is accessible to a wider audience and it is more clearly visible that the additional restrictions on \( \mathcal{X} \) and \( \mu \) are in fact not required. In addition, [19] does not contain our statement (c).

The plan of the paper is as follows. In Section 2 we give a basic comparison between \( W_1 \) and \( W_2 \) for the practitioner. While our discussion is not exhaustive, mainly because more detailed considerations depend a great deal on the scenario at hand, we feel that a systematic discussion about the choice of \( p \) has been missing from the literature. Section 3 deals with the general theory of semi-discrete optimal transport for \( p = 1 \), and Section 4 considers the special representation of transport maps as transport partition for this case and formulates an equivalent unconstraint minimization problem. Section 5 describes a concrete algorithm for computing the semi-discrete optimal transport map, which extends ideas from the well-known AHA algorithm [11, 22]. It also provides the essential details of our implementation. Section 6 contains some numerical experiments and Section 7 provides discussion and an outlook.

2 \( W_1 \) versus \( W_2 \) from an applied point of view

Intuitively it seems that imposing as cost the actual distance between locations is the most natural choice, at least in applications where we have mass distributions with some real spatial meaning. However, the majority of applied analyses work with \( p = 2 \), often based on some fundamental believe that computation of Wasserstein distances is faster and more theoretically sound in this case. Apart from the fact that it is rarely discussed whether the results that this belief is based on are relevant for the analysis at hand, there is a number of other considerations that should be taken into account that might favour the use of \( p = 1 \). Note that the following discussion is general and does not particularly refer to the semi-discrete case.

2.1 Meandering express ways and redistribution of mass along structures

One downside of using \( W_2 \) is that mass can be transported on crooked paths that are very unintuitive from an applied point of view. Even worse, such paths between source and target locations may vitally depend on how mass between and around them is distributed, in ways that go beyond mere questions of competing for mass. An artificial example that takes things to an extreme, but illustrates these points rather nicely, is given in Figure 1.
Each point along the incomplete circle denotes the location of one unit of mass of \( \mu \) (blue x-points) and/or \( \nu \) (red o-points). The presence of the shared locations along the circle is responsible for the fact that the (unique) \( L_2 \)-optimal transport (i.e. the transport plan taking the minimum in (1) for \( p = 2 \)) is along the circle, while there is always an \( L_1 \)-optimal transport that just takes the individual x-point to the individual o-point regardless of any other mass that is shared by both measures. This transport is unique if no other mass lies on the connecting line. Note that this example does not vitally depend on the fact that the other locations are shared exactly nor on the fact that the two measures are discrete.

While this behavior in the case \( p = 2 \) is unintuitive and typically unpleasant for many applications, it may also be desirable to some extent. Say we consider biochemical imagery and our interest lies in meandering structures that due to the measurement procedure appear with somewhat different mass distributions at different times, but essentially keep their shape and position. Then the \( L_2 \)-optimal transport identifies corresponding structures more easily with one another, whereas in the \( L_1 \)-optimal transport plan mass of one structure gets more quickly associated with another structure, lying e.g. in a bend of the first one.

2.2 Crossing of transportation paths and motions of rigid objects

A well-known property of the \( L_1 \)-optimal transport is that transportation paths must not form proper crossings (but may run parallelly on top of one another). This is true for any optimal transport that is based on a metric cost function, which follows by a combination of the triangle inequality and cyclical monotonicity; see [36, beginning of Chapter 8] for the precise argument. For the \( L_2 \)-optimal transport, crossings of transport paths are typically abundant. This is illustrated by an artificial example in Figure 2. Note that the angle between the two line segments of support points has been chosen slightly above 90 degrees so that the solution for \( p = 2 \) is unique.

While the crossing of transport paths seems counterintuitive and is clearly undesirable in some applications, there is also the desirable aspect that it allows to better capture motion in time of more or less rigid objects. Assume now that Figure 2 shows (discrete approximations of) “before” and “after” pictures of a stick that has been turned around a corner rather than flipped at a bisecting line. This movement may appear more natural to
the human observer, especially if the angle between “before” and “after” is more obtuse and the stick is not perfectly symmetric. While the $L_1$-optimal transport would keep insisting that the stick was flipped up to an angle of almost $\pi$, the $L_2$-optimal transport captures the turning of the stick much better. This toy example is an expression of the well-known finding in [6] that identifies the $L_2$-optimal transport as the result of a pressureless potential flow in the context of fluid mechanics.

![Figure 2: Optimal transport map from blue-x to red-o measure with unit mass at each point; left: $p = 1$, right: $p = 2$.](image)

### 2.3 Formulae for affine transformations of measures

We have the following relations when adding a common measure or multiplying by a common nonnegative scalar. The proof easily extends to a complete separable metric space instead of $\mathbb{R}^d$ equipped with the Euclidean metric.

**Lemma 1.** Let $\mu, \nu, \alpha$ be finite measures on $\mathbb{R}^d$ satisfying $\mu(\mathbb{R}^d) = \nu(\mathbb{R}^d)$. For $p \geq 1$ and $c > 0$, we have

\[
W_p(\alpha + \mu, \alpha + \nu) \leq W_p(\mu, \nu), \quad (4)
\]

\[
W_1(\alpha + \mu, \alpha + \nu) = W_1(\mu, \nu), \quad (5)
\]

\[
W_p(c\mu, c\nu) = c^{1/p} W_p(\mu, \nu), \quad (6)
\]

where we assume for (5) that $W_1(\mu, \nu) < \infty$.

**Proof.** Write $\Delta = \{(x, x) \mid x \in \mathbb{R}^d\}$. Denote by $\alpha_\Delta$ the push-forward of $\alpha$ under the map $[\mathbb{R}^d \rightarrow \mathbb{R}^d \times \mathbb{R}^d, x \mapsto (x, x)]$. Let $\pi_*$ be an optimal transport plan for the computation of $W_p(\mu, \nu)$. Then $\pi_* + \alpha_\Delta$ is a feasible transport plan for $W_p(\alpha + \mu, \alpha + \nu)$ that generates the same total cost as $\pi_*$. Thus

\[
W_p(\alpha + \mu, \alpha + \nu) \leq W_p(\mu, \nu).
\]

Likewise $c\pi_*$ is a feasible transport plan for $W_p(c\mu, c\nu)$ that generates $c^{1/p}$ times the cost of $\pi_*$ for the integral in (4). Thus

\[
W_p(c\mu, c\nu) \leq c^{1/p} W_p(\mu, \nu).
\]

Replacing $c$ by $1/c$, as well as $\mu$ by $c\mu$ and $\nu$ by $c\nu$, we obtain (6).
It remains to show \( W_1(\alpha + \mu, \alpha + \nu) \geq W_1(\mu, \nu) \). For this we use that a transport plan \( \pi \) between \( \mu \) and \( \nu \) is optimal if and only if it is cyclical monotone, meaning that for all \( N \in \mathbb{N} \) and all \( (x_1, y_1), \ldots, (x_N, y_N) \in \text{supp}(\pi) \), we have

\[
\sum_{i=1}^{N} \|x_i - y_i\| \leq \sum_{i=1}^{N} \|x_i - y_{i+1}\|,
\]

where \( y_{N+1} = y_1 \); see [36] Theorem 5.10(ii) and Definition 5.1.

Letting \( \pi_* \) be an optimal transport plan for the computation of \( W_1(\mu, \nu) \), we show optimality of \( \pi_* + \alpha \Delta \) for the computation of \( W_1(\mu + \alpha, \nu + \alpha) \). We know that \( \pi_* \) is cyclical monotone. Let \( N \in \mathbb{N} \) and \( (x_1, y_1), \ldots, (x_N, y_N) \in \text{supp}(\pi_* + \alpha \Delta) \subset \text{supp}(\pi_*) \cup \Delta \). Denote by \( 1 \leq i_1 < \ldots < i_k \leq N \), where \( k \in \{0, \ldots, N\} \), the indices of all pairs with \( x_{i_j} \neq y_{i_j} \), and hence \( (x_{i_j}, y_{i_j}) \in \text{supp}(\pi_*) \). By the cyclical monotonicity of \( \pi_* \) (writing \( i_{k+1} = i_1 \)) and the triangle inequality, we obtain

\[
\sum_{i=1}^{N} \|x_i - y_i\| = \sum_{j=1}^{k} \|x_{i_j} - y_{i_j}\| \leq \sum_{j=1}^{k} \|x_{i_j} - y_{i_{j+1}}\| \leq \sum_{i=1}^{N} \|x_i - y_{i+1}\|.
\]

Thus \( \pi_* + \alpha \Delta \) is cyclical monotone and since it is a feasible transport plan between \( \mu + \alpha \) and \( \nu + \alpha \), it is optimal for the computation of \( W_1(\mu + \alpha, \nu + \alpha) \), which concludes the proof.

**Remark 1.** Equation \([5] \) is not generally true for any \( p > 1 \). To see this consider the case \( d = 1, \mu = \delta_0, \nu = \delta_1 \) and \( \alpha = b \text{Leb} \cdot [0,1] \), where \( b \geq 1 \). Clearly \( W_p(\mu, \nu) = 1 \) for all \( p \geq 1 \). Denote by \( F \) and \( G \) the cumulative distribution functions (CDFs) of \( \mu + \alpha \) and \( \nu + \alpha \), respectively, i.e. \( F(x) = \mu((-\infty, x]) \) and \( G(x) = \nu((-\infty, x]) \) for all \( x \in \mathbb{R} \). Thus

\[
F(x) = G(x) = 0 \quad \text{if } x < 0, \\
F(x) = 1 + bx, \ G(x) = bx \quad \text{if } x \in [0,1), \\
F(x) = G(x) = b + 1 \quad \text{if } x \geq 1.
\]

We then even obtain

\[
W_p^p(\alpha + \mu, \alpha + \nu) = \int_0^{b+1} |F^{-1}(t) - G^{-1}(t)|^p \, dt \\
= 2 \int_0^1 \frac{t^p}{b^p} \, dt + \frac{1}{b^p}(b - 1) = \frac{1}{b^p} \left( b - 1 + \frac{2}{p+1} \right) \to 0
\]
as \( b \to \infty \) if \( p > 1 \). For the first equality we used the representation of \( W_p \) in terms of (generalized) inverses of their CDFs; see Equation (2) in [29] and the references given there and note that the generalization from the result for probability measures is immediate by \([6] \).

Considering in particular the case \( p = 2 \), we see that the behavior of the Wasserstein metric, when adding the same measure \( \alpha \) to the measures \( \mu \) and \( \nu \) or when multiplying the same constant to \( \mu \) and \( \nu \) is not very intuitive. Extrapolating somewhat from the above example, when adding \( \alpha \) with huge, well-distributed total mass, any information about a discrepancy between \( \mu \) and \( \nu \) gets essentially lost. When multiplying a constant, the Wasserstein distance scales as a square root of the constant. Such behavior must be taken care of, e.g. when adapting brightness and contrast in two images that are to be compared in \( W_2 \).

In contrast \( W_1 \) is not influenced by adding a common measure and scales linearly in the mass factor.
2.4 Algorithmic considerations

As pointed out in the introduction, the special form of the cost function if \( p = 2 \) allows for efficient algorithms, such as [31] in the discrete and [4, 22] in the semi-discrete setting.

For \( p = 1 \) the cost function is not decomposable, but the Monge–Kantorovich problem still offers considerable advantages because of the particularly simple form of its dual problem [35, Particular Case 5.16] and as a consequence the equivalence to Beckmann’s problem; see [30, Theorem 4.6] for this result or [32, Section 4] for a complete description of a (more general) algorithm derived in the discrete case from this fact.

In the present paper we describe an algorithm for the semi-discrete case and \( p = 1 \) that is reasonably fast, but cannot quite reach the performance of the algorithm for \( p = 2 \) in [22]. This is again mainly due to the nice decomposition property of the cost function for \( p = 2 \).

3 Semi-discrete optimal transport

We first concretize the semi-discrete setting and introduce some additional notation. Let now \( \mathcal{X} \) and \( \mathcal{Y} \) be Borel subsets of \( \mathbb{R}^d \) and let \( \mu \) and \( \nu \) be probability measures on \( \mathcal{X} \) and \( \mathcal{Y} \), respectively. This is just for notational convenience and does not change the set of admissible measures in an essential way: We may always set \( \mathcal{X} = \mathcal{Y} = \mathbb{R}^d \) and any statement about \( \mu \) and \( \nu \) we make can be easily recovered for \( c\mu \) and \( c\nu \) for arbitrary \( c > 0 \).

For the rest of the article it is tacitly assumed that \( d \geq 2 \) to avoid certain pathologies of the one-dimensional case that would lead to a somewhat tedious distinction of cases in various results for a case that is well-understood anyway. Moreover, we always require \( \mu \) to be absolutely continuous with density \( \varrho \) with respect to \( d \)-dimensional Lebesgue measure \( \text{Leb}^d \) and to satisfy

\[
\int_{\mathcal{X}} \|x\| \, \mu(dx) < \infty. \tag{7}
\]

We assume further that \( \nu = \sum_{j=1}^{n} \nu_j \delta_{y_j} \), where \( n \in \mathbb{N} \), \( y_1, \ldots, y_n \in \mathcal{Y} \) and \( \nu_1, \ldots, \nu_n \in (0,1] \). Condition (7) guarantees that

\[
W_1(\mu, \nu) \leq \int_{\mathcal{X}} \|x\| \, \mu(dx) + \int_{\mathcal{Y}} \|y\| \, \nu(dy) =: C < \infty, \tag{8}
\]

which simplifies certain arguments.

The set of Borel subsets of \( \mathcal{X} \) is denoted by \( \mathcal{B}_X \). Lebesgue mass is denoted by absolute value bars, i.e. \( |A| = \text{Leb}^d(A) \) for every \( A \in \mathcal{B}_X \).

A first pleasant realization for the semi-discrete case is that we will be able to work entirely with transport maps or — equivalently, but more conveniently — with transport partitions. A transport partition from \( \mu \) to \( \nu \) is any partition \( \mathcal{C} = (C_j)_{1 \leq j \leq n} \) of \( \mathcal{X} \) into Borel sets that satisfy \( \mu(C_j) = \nu_j \). From such a partition \( \mathcal{C} \) the associated transport map \( T_\mathcal{C} \) is given as \( T_\mathcal{C}(x) = \sum_{j=1}^{n} y_j 1\{x \in C_j\} \). Conversely, based on an arbitrary transport map \( T \) from \( \mu \) to \( \nu \), the associated transport partition is obtained as \( \mathcal{C}_T = (T^{-1}(y_j))_{1 \leq j \leq n} \).

Monge’s problem for \( p = 1 \) can then be equivalently formulated as finding

\[
\inf_{\mathcal{C}} \int_{\mathcal{X}} \|x - T_\mathcal{C}(x)\| \, \mu(dx) = \inf_{\mathcal{C}} \sum_{j=1}^{n} \int_{C_j} \|x - y_j\| \, \mu(dx), \tag{9}
\]

where the infima are taken over all transport partitions \( \mathcal{C} = (C_j)_{1 \leq j \leq n} \) from \( \mu \) to \( \nu \). Contrary to the difficulties encountered for more general measures \( \mu \) and \( \nu \) when considering Monge’s problem with Euclidean costs, the situation for semi-discrete optimal transport is clear-cut.
Theorem 1. In the semi-discrete setting with Euclidean costs (always including $d \geq 2$ and (7)) there is a $\mu$-a.s. unique solution $T_*$ to Monge’s problem. The induced coupling $\pi_{T_*}$ is the unique solution to the Monge–Kantorovich problem, yielding

$$ W_1(\mu, \nu) = \int_X \|x - T_*(x)\| \, \mu(dx). \tag{10} $$

Proof. The part concerning Monge’s problem is a consequence of the concrete construction in Section 4; see Theorem 2.

Clearly $\pi_{T_*}$ is an admissible transport plan for the Monge–Kantorovich problem. Since $\mu$ is non-atomic and the Euclidean cost function is continuous, Theorem B in [27] implies that the minimum in the Monge–Kantorovich problem is equal to the infimum in the Monge problem, so $\pi_{T_*}$ must be optimal.

For the uniqueness of $\pi_{T_*}$ in the Monge–Kantorovich problem, let $\pi$ be an arbitrary optimal transport plan. Define the measures $\tilde{\pi}_i$ on $X$ by $\tilde{\pi}_i(A) := \pi(A \times \{y_i\})$ for all $A \in \mathcal{B}_X$ and $1 \leq i \leq n$. Since $\sum \pi_i = \mu$, all $\pi_i$ are absolutely continuous with respect to Leb with densities $\tilde{\rho}_i$ satisfying $\sum \tilde{\rho}_i = \rho$. Set then $S_i := \{ x \in X | \tilde{\rho}_i > 0 \}.$

Assume first that there exist $i, j \in \{1, \ldots, n\}$, $i \neq j$, such that $|S_i \cap S_j| > 0$. Define $H^{ij}_\leq(q) := \{ x \in X | \|x - y_i\| < \|x - y_j\| + q \}$ and $H_{\leq}^{ij}(q)$ analogously. There exists a $q \in \mathbb{R}$ for which both $S_i \cap S_j \cap H_{\leq}^{ij}(q)$ and $S_i \cap S_j \cap H^{ij}(q)$ have positive Lebesgue measure: Choose $q_1, q_2 \in \mathbb{R}$ such that $|S_i \cap S_j \cap H^{ij}(q_1)| > 0$ and $|S_i \cap S_j \cap H_{\leq}^{ij}(q_2)| > 0$; using binary search between $q_1$ and $q_2$, we find the desired $q$ in finitely many steps, because otherwise there would have to exist a $q_0$ such that $|S_i \cap S_j \cap H_{\leq}^{ij}(q_0)| > 0$, which is not possible. By the definition of $S_i$ and $S_j$, we thus have $\alpha = \pi_i(S_i \cap S_j \cap H^{ij}(q)) > 0$ and $\beta = \pi_j(S_i \cap S_j \cap H_{\leq}^{ij}(q)) > 0$. Switching $i$ and $j$ if necessary, we may assume $\alpha \leq \beta$. Define then

$$ \pi_i' = \pi_i - \pi_i|_{S_i \cap S_j \cap H^{ij}_\leq(q)} + \frac{\alpha}{\beta} \pi_j|_{S_i \cap S_j \cap H_{\leq}^{ij}(q)}, $$

$$ \pi_j' = \pi_j + \pi_i|_{S_i \cap S_j \cap H^{ij}_\leq(q)} - \frac{\alpha}{\beta} \pi_j|_{S_i \cap S_j \cap H_{\leq}^{ij}(q)} $$

and $\pi_k' = \pi_k$ for $k \notin \{i, j\}$. It can be checked immediately that the measure $\pi'$ given by $\pi'(A \times \{y_i\}) = \pi_i'(A)$ for all $A \in \mathcal{B}_X$ and all $i \in \{1, 2, \ldots, n\}$ is a transport plan from $\mu$ to $\nu$ again. It satisfies

$$ \int_{X \times Y} \|x - y\| \, \pi'(dx, dy) = \int_{X \times Y} \|x - y\| \, \pi(dx, dy) $$

$$ = \int_{S_i \cap S_j \cap H^{ij}_\leq(q)} (-\|x - y_i\| + \|x - y_j\|) \pi_i(dx) + \frac{\alpha}{\beta} \int_{S_i \cap S_j \cap H_{\leq}^{ij}(q)} \left(\|x - y_i\| - \|x - y_j\|\right) \pi_j(dx) $$

$$ < 0, $$

because the integrands are strictly negative on the sets over which we integrate. But this contradicts the optimality of $\pi$.

We thus have proved that $|S_i \cap S_j| = 0$ for all pairs with $i \neq j$. This implies that we can define a transport map $T$ inducing $\pi$ in the following way. If $x \in S_i \setminus (\cup_{j \neq i} S_j)$ for some $i$, set $T(x) := y_i$. Since the intersections $S_i \cap S_j$ are Lebesgue null sets, the value of $T$ on them does not matter. So we can for example set $T(x) := y_1$ or $T(x) := y_{i_0}$ for $x \in \cap_{i \notin I} S_i \setminus \cap_{i \in I} S_i$, where $I \subseteq \{1, \ldots, n\}$ contains at least to elements and $i_0 = \min(I)$. It follows that $\pi_T = \pi$. But by the optimality of $\pi$ and Theorem 2 we obtain $T = T_* \mu$-almost surely, which implies $\pi = \pi_T = \pi_{T_*}$. \hfill \Box

It will be desirable to know in what way we may approximate the continuous and discrete Monge–Kantorovich problems by the semi-discrete problem we investigate here.
In the fully continuous case, we have a measure $\tilde{\nu}$ on $\mathcal{Y}$ with density $\tilde{\rho}$ with respect to $\text{Leb}^d$ instead of the discrete measure $\nu$. In the fully discrete case, we have a discrete measure $\tilde{\mu} = \sum_{i=1}^m \tilde{\mu}_i \delta_{x_i}$ instead of the absolutely continuous measure $\mu$, where $m \in \mathbb{N}$, $x_1, \ldots, x_m \in \mathcal{X}$ and $\tilde{\mu}_1, \ldots, \tilde{\mu}_m \in (0,1]$. In both cases existence of an optimal transport plan is still guaranteed by \cite{[36]}, Theorem 4.1, however we lose to some extent the uniqueness property.

One reason for this is that mass transported within the same line segment can be reassigned at no extra cost; see the discussion on transport rays in Section 6 of \cite{[2]}. In the continuous case this is the only reason, and uniqueness can be restored by minimizing a secondary functional (e.g. total cost with respect to $p > 1$) over all optimal transport plans; see Theorem 7.2 in \cite{[2]}.

In the discrete case non-uniqueness depends strongly on the geometry of the support points of $\tilde{\mu}$ and $\nu$. If both supports form regular grids or if several points of one support are equidistantly placed around points of the other support, mass may be rerouted in many different ways without changing the total cost. However, if support points are in sufficiently general position, the optimal transport plan is still unique.

When approximating the continuous problem with measures $\mu$ and $\tilde{\nu}$ by a semi-discrete problem, we quantize the measure $\tilde{\nu}$ to obtain a discrete measure $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$, where $\nu_j = \tilde{\nu}(N_j)$ for a partition $(N_j)$ of $\text{supp}(\tilde{\nu})$. The error we commit in Wasserstein distance by discretization of $\tilde{\nu}$ is bounded by the quantization error, i.e.

$$W_1(\mu, \tilde{\nu}) - W_1(\mu, \nu) \leq W_1(\tilde{\nu}, \nu) \leq \sum_{j=1}^n \int_{N_j} \|y - y_j\| \tilde{\nu}(dy). \tag{11}$$

We can compute $W_1(\tilde{\nu}, \nu)$ exactly by solving another semi-discrete transport problem, using the algorithm described in Section 5 to compute an optimal partition $(N_j)$ for the second inequality above. However, choosing $\nu$ for given $n$ in such a way that $W_1(\tilde{\nu}, \nu)$ is minimal is usually practically infeasible. So we would use an algorithm that makes $W_1(\tilde{\nu}, \nu)$ reasonably small, such as a suitable version of Lloyd’s algorithm; see Subsection 5.1 below.

When approximating the discrete problem with measures $\tilde{\mu}$ and $\nu$ by a semi-discrete problem, we blur each mass $\tilde{\mu}_i$ of $\tilde{\mu}$ over a neighborhood of $x_i$ using a probability density $f_i$, to obtain a measure $\mu$ with density $\varrho(x) = \sum_{i=1}^m \tilde{\mu}_i f_i(x_i)$. Typical examples use $f_i(x) = \frac{1}{h} \varphi\left(\frac{x-x_i}{h}\right)$, where $\varphi$ is the standard normal density and the bandwidth $h > 0$ is reasonably small, or $f_i(x) = \frac{1}{|M_i|} \mathbb{1}_{M_i}(x)$, where $M_i$ is some small neighborhood of $x_i$. In practice, discrete measures are often available in the form of images, where the support points $x_i$ form a fine rectangular grid; then the latter choice of $f_i$s is very natural, where the $M_i$s are just adjacent squares, each with an $x_i$ at the center. There are similar considerations for the approximation error as in the fully continuous case above. In particular the error we commit in Wasserstein distance is bounded by the blurring error

$$W_1(\tilde{\mu}, \nu) - W_1(\mu, \nu) \leq W_1(\tilde{\mu}, \mu) \leq \sum_{i=1}^m \tilde{\mu}_i \int_{\mathbb{R}^d} \|x - x_i\| f(x) \, dx. \tag{12}$$

The right hand side is typically straightforward to compute exactly, e.g. in the normal density and grid cases described above. It can be made small by choosing the bandwidth $h$ very small or picking sets $M_i$ of small radius $r = \sup_{x \in M_i} \|x - x_i\|$.

What about the approximation properties of the optimal transport plans obtained by the semi-discrete setting? Theorem 5.20 in \cite{[36]} implies for $\nu^{(k)} \rightharpoonup \tilde{\nu}$ weakly and $\mu^{(k)} \rightharpoonup \tilde{\mu}$ weakly that every subsequence of the sequence of optimal transport plans $\pi^{(k)}$ between $\mu^{(k)}$ and $\nu^{(k)}$ has a further subsequence that converges weakly to an optimal transport
plan $\pi_*$ between $\mu$ and $\nu$. This implies that for every $\varepsilon > 0$ there is a $k_0 \in \mathbb{N}$ such that for any $k \geq k_0$ the plan $\pi^{(k)}$ is within distance $\varepsilon$ (in any fixed metrization of the weak topology) of some optimal transport plan between $\mu$ and $\nu$, which is the best we could have hoped for in view of the non-uniqueness of optimal transport plans we have in general. If (in the discrete setting) there is a unique optimal transport plan $\pi^*$, this yields that $\pi^{(k)}_* \to \pi_*$ weakly.

## 4 Optimal transport maps via weighted Voronoi tessellations

As shown for bounded $\mathcal{X}$ in [13], the solution to the semi-discrete transport problem has a nice geometrical interpretation, which is similar to the well-known result in [4]: We elaborate below that the sets $C_j^*$ of the optimal transport partition are the cells of an additively weighted Voronoi tessellation of $\mathcal{X}$ around the support points of $\nu$.

For the finite set of points $\{y_1, \ldots, y_n\}$ and a vector $w \in \mathbb{R}^n$ that assigns to each $y_j$ a weight $w_j$, the additively weighted Voronoi tessellation is the set of cells

$$\text{Vor}_w(j) = \{x \in \mathcal{X} \mid \|x - y_j\| - w_j \leq \|x - y_k\| - w_k \quad \text{for all } k \neq j\}, \quad j = 1, \ldots, n.$$ 

Note that adjacent cells $\text{Vor}_w(j)$ and $\text{Vor}_w(k)$ have disjoint interiors. The intersection of their boundaries is a subset of $H = \{x \in \mathcal{X} \mid \|x - y_j\| - \|x - y_k\| = w_j - w_k\}$, which is easily seen to have Lebesgue measure (and hence $\mu$-measure) zero. If $d = 2$, the set $H$ is a branch of a hyperbola with foci at $y_j$ and $y_k$. It may also be interpreted as the set of points that have the same distance from the spheres $S(y_j, w_j)$ and $S(y_k, w_k)$, where $S(y, w) = \{x \in \mathcal{X} \mid \|x - y\| = w\}$. See Figure 3 for an illustration of these properties.

Of course not all weighted Voronoi tessellations are valid transport partitions from $\mu$ to $\nu$. But suppose we can find a weight vector $w$ such that the resulting Voronoi tessellation satisfies indeed $\mu(\text{Vor}_w(j)) = \nu_j$ for every $j \in \{1, \ldots, n\}$; we call such a $w$ adapted to $(\mu, \nu)$. Then this partition is automatically optimal.

**Theorem 2.** If $w \in \mathbb{R}^n$ is adapted to $(\mu, \nu)$, then $(\text{Vor}_w(j))_{1 \leq j \leq n}$ is the $\mu$-almost surely unique optimal transport partition from $\mu$ to $\nu$. 

Figure 3: An additively weighted Voronoi tessellation with 25 cells.
A proof was given in [13, Theorem 2] for more general distance functions, but required $X$ to be bounded. For the Euclidean distance we consider here, we can easily extend it to unbounded $X$; see [16, Theorem 3.2].

Having identified this class of optimal transport partitions, it remains to show that for each pair $(\mu, \nu)$ we can find an adapted weight vector. We adapt the approach of [4] to the case $p = 1$, which gives us a constructive proof that forms the basis for the algorithm in Section 5. Our key tool is the function $\Phi$ defined below.

**Theorem 3.** Let $\Phi : \mathbb{R}^n \to \mathbb{R}$,

$$
\Phi(w) = \sum_{j=1}^n \left( -\nu_j w_j - \int_{\text{Vor}_w(j)} (\|x - y_j\| - w_j) \, \mu(dx) \right).
$$

Then

a) $\Phi$ is convex;

b) $\Phi$ is continuously differentiable with partial derivatives

$$
\frac{\partial \Phi}{\partial w_j}(w) = -\nu_j + \mu(\text{Vor}_w(j));
$$

c) $\Phi$ takes a minimum in $\mathbb{R}^n$.

**Remark 2.** Let $w^* \in \mathbb{R}^n$ be a minimizer of $\Phi$. Then by Theorem 3b)

$$
\mu(\text{Vor}_{w^*}(j)) - \nu_j = \frac{\partial \Phi}{\partial w_j}(w^*) = 0 \quad \text{for every } j \in \{1, \ldots, n\},
$$

i.e. $w_*$ is adapted to $(\mu, \nu)$. Theorem 2 yields that $(\text{Vor}_{w^*}(j))_{1 \leq j \leq n}$ is the $\mu$-almost surely unique optimal transport partition from $\mu$ to $\nu$.

**Proof** of Theorem 3. We take a few shortcuts; for full technical details see Chapter 3 of [16].

Part (a) relies on the observation that $\Phi$ can be written as

$$
\Phi(w) = \sum_j (-\nu_j w_j) - \Psi(w)
$$

where

$$
\Psi(w) = \int_X (\|x - T^w(x)\| - w_{T^w(x)}) \, \mu(dx),
$$

$T^w$ denotes the transport map induced by the Voronoi tesselation with weight vector $w$ and we write $w_{y_j}$ instead of $w_j$ for convenience. By definition of the weighted Voronoi tesselation $\Psi$ is the infimum of the affine functions

$$
\Psi_f : \mathbb{R}^n \to \mathbb{R}, \quad w \mapsto \int_X (\|x - f(x)\| - w_{f(x)}) \, \mu(dx)
$$

over all measurable maps $f$ from $X$ to $Y$. Since pointwise infima of affine functions are concave and the first summand of $\Phi$ is linear, we see that $\Phi$ convex.

By geometric arguments it can be shown that $[w \mapsto \mu(\text{Vor}_w(j))]$ is continuous; see [16, Lemma 3.3]. A short computation involving the representation $\Psi(w) = \inf_f \Psi_f(w)$ used
above yields for the difference quotient of $\Psi$, writing $e_j$ for the $j$-th standard basis vector and letting $h \neq 0$,
\[
\left| \frac{\Psi(w + he_j) - \Psi(w)}{h} + \mu(\text{Vor}_w(j)) \right| \leq \left| -\mu(\text{Vor}_w(j)) + \mu(\text{Vor}_{w+h e_j}(j)) \right| \rightarrow 0
\]
as $h \to 0$. This implies that $\Psi$ is $C^1$ with (continuous) $j$-th partial derivative $-\mu(\text{Vor}_w(j))$ and hence statement [b] follows.

Finally, for the existence of a minimizer of $\Phi$ we consider an arbitrary sequence $(w^{(k)})_{k \in \mathbb{N}}$ of weight vectors in $\mathbb{R}^n$ with
\[
\lim_{k \to \infty} \Phi(w^{(k)}) = \inf_{w \in \mathbb{R}^n} \Phi(w).
\]
We show below that a suitably shifted version of $(w^{(k)})_{k \in \mathbb{N}}$ that has the same $\Phi$-values contains a bounded subsequence. This subsequence then has a further subsequence $(u^{(k)})$ which converges towards some $u \in \mathbb{R}^n$. Continuity of $\Phi$ yields
\[
\Phi(u) = \lim_{k \to \infty} \Phi(u^{(k)}) = \inf_{w \in \mathbb{R}^n} \Phi(w)
\]
and thus statement [c].

To obtain the bounded subsequence, note first that adding to each weight the same constant neither affects the Voronoi tessellation nor the value of $\Phi$. We may therefore assume $w_j^{(k)} \geq 0$, $1 \leq j \leq n$, for all $k \in \mathbb{N}$. Choosing an entry $i$ and an infinite set $K \subset \mathbb{N}$ appropriately leaves us with a sequence $(w^{(k)})_{k \in K}$ satisfying $w_i^{(k)} \geq w_j^{(k)}$ for all $j$ and $k$. Taking a further subsequence $(w^{(l)})_{l \in \mathbb{N}}$ for some infinite $L \subset K$ allows the choice of an $R \geq 0$ and the partitioning of $\{1, \ldots, n\}$ into two sets $A$ and $B$ such that for every $l \in \mathbb{N}$

i) $0 \leq w_i^{(l)} - w_j^{(l)} \leq R$ if $j \in A$,

ii) $w_i^{(l)} - w_j^{(l)} \geq \text{index}(l)$ if $j \in B$,

where $\text{index}(l)$ denotes the rank of $l$ in $L$.

Assume that $B \neq \emptyset$. The Voronoi cells with indices in $B$ will at some point be shrunk to measure zero, meaning there exists an $N \in \mathbb{N}$ such that
\[
\sum_{j \in A} \mu(\text{Vor}_{w^{(l)}}(j)) = 1 \text{ for all } l \geq N.
\]
Write
\[
w_A^{(l)} = \min_{i \in A} w_i^{(l)} \quad \text{and} \quad w_B^{(l)} = \max_{i \in B} w_i^{(l)},
\]
and recall the constant $C$ from [8], which may clearly serve as an upper bound for the transport cost under an arbitrary plan. We then obtain for every $l \geq N$
\[
\Phi(w^{(l)}) = \sum_{j=1}^n \left( -\nu_j w_j^{(l)} - \int_{\text{Vor}_{w^{(l)}}(j)} (\|x - y_j\| - w_j^{(l)}) \mu(dx) \right) \\
\geq -C + \sum_{j=1}^n w_j^{(l)} \left( \mu(\text{Vor}_{w^{(l)}}(j)) - \nu_j \right) \\
= -C + \sum_{j \in A} w_j^{(l)} \left( \mu(\text{Vor}_{w^{(l)}}(j)) - \nu_j \right) - \sum_{j \in B} w_j^{(l)} \nu_j \\
\geq -C - R + w_A^{(l)} \left( 1 - \sum_{j \in A} \nu_j \right) - w_B^{(l)} \sum_{j \in B} \nu_j \\
\geq -C - 2R + \text{index}(l),
\]
which contradicts the statement $\lim_{k \to \infty} \Phi(w^{(k)}) = \inf_{w \in \mathbb{R}^n} \Phi(w) < \infty$. Thus we have $B = \emptyset$.

We can then simply turn $(w^{(l)})_{l \in L}$ into a bounded sequence by substracting the minimal entry $w^{(l)} = \min_{1 \leq i \leq n} w_i^{(l)}$ from each $w_j^{(l)}$ for all $l \in L$. □

5 The algorithm

The previous section provides the theory needed to compute the optimal transport partition. It is sufficient to find a vector $w^*$ where $\Phi$ is locally optimal. By convexity, $w^*$ is then a global minimizer of $\Phi$ and Remark 2 identifies the $\mu$-a.s. unique optimal transport partition as $(\text{Vor}_{w^*}(j))_{1 \leq j \leq n}$.

For the optimization process we can choose from a variety of methods thanks to knowing the gradient $\nabla \Phi$ of $\Phi$ analytically from Theorem 3. We consider iterative methods that start at an initial weight vector $w^{(0)}$ and apply steps of the form

$$w^{(k+1)} = w^{(k)} + t_k \Delta w^{(k)}, \quad k \geq 0,$$

where $\Delta w^{(k)}$ denotes the search direction and $t_k$ the step size.

Newton’s method would use $\Delta w^{(k)} = -D^2\Phi(w^{(k)})^{-1} \nabla \Phi(w^{(k)})$, but the Hessian matrix $D^2\Phi(w^{(k)})$ is not available to us. We therefore use a quasi-Newton method that makes use of the gradient. Just like [22] for the case $p = 2$, we have obtained many good results using L-BFGS [25], the limited-memory variant of the Broyden–Fletcher–Goldfarb–Shanno algorithm, which uses the value of the gradient at the current as well as at preceding steps for approximating the Hessian. The limited-memory variant works without storing the whole Hessian of size $n \times n$, which is important since in applications our $n$ is typically large.

To determine a suitable step size $t_k$ we use the strong Wolfe condition [38], which has proven to be well suited for our problem. It considers different values for $t_k$ until it arrives at one that sufficiently decreases both $\Phi$ and $\nabla \Phi$.

5.1 Multiscale approach to determine starting value

To find a good starting value $w^{(0)}$ we use a multiscale method similar to the one proposed by [22]. We first create a decomposition of $\nu$, i.e. a sequence $\nu = \nu^{(0)}, \ldots, \nu^{(L)}$ of measures with decreasing cardinality of the support. Here $\nu^{(l)}$ is obtained as a coarsening of $\nu^{(l-1)}$ by merging the masses of several points into one point.

It seems intuitively reasonable to choose $\nu^{(l)}$ in such a way that $W_1(\nu^{(l)}, \nu^{(l-1)})$ is as small as possible, since the latter bounds $|W_1(\mu, \nu^{(l)}) - W_1(\mu, \nu^{(l-1)})|$. This comes down to a capacitated location-allocation problem, which is NP-hard even in the one-dimensional case; see [34]. Out of speed concerns and since we only need a reasonably good starting value for our algorithm, we decided to content ourselves with the same weighted $K$-means clustering algorithm used by [22] (referred to as Lloyd’s algorithm), which iteratively improves an initial aggregation of the support of $\nu^{(l-1)}$ into $|\text{supp}(\nu^{(l)})|$ clusters towards local optimality with respect to the squared Euclidean distance. The resulting $\nu^{(l)}$ is then the discrete measure with the cluster centers as its support points and as weights the summed up weights of the points of $\nu^{(l-1)}$ contained in each cluster; see Algorithm 3 in [16]. The corresponding weighted $K$-median clustering algorithm, based on alternating between assignment of points to clusters and recomputation of cluster centers as the median of all weighted points in the cluster, should intuitively give a $\nu^{(l)}$ based
on which we obtain a better starting solution. This may sometimes compensate for the much longer time needed for performing $K$-median clustering.

Having created the decomposition $\nu = \nu^{(0)}, \ldots, \nu^{(L)}$, we minimize $\Phi$ along the sequence of these coarsened measures, beginning at $\nu^{(L)}$ with the initial weight vector $w^{(L,0)} = 0 \in \mathbb{R}^{\left|\text{supp}(\nu^{(L)})\right|}$ and computing the optimal weight vector $w^{(L,*)}$ for the transport from $\mu$ to $\nu^{(L)}$. Every time we pass to a finer measure $\nu^{(l-1)}$ from the coarser measure $\nu^{(l)}$, we generate the initial weight vector $w^{(l-1,0)}$ from the last optimal weight vector $w^{(l,*)}$ by assigning the weight of each support point of $\nu^{(l)}$ to all the support points of $\nu^{(l-1)}$ from whose merging the point of $\nu^{(l)}$ originated; see also Algorithm 2 in [16].

### 5.2 Numerical computation of $\Phi$ and $\nabla \Phi$

For practical computation we assume here that $\mathcal{X}$ is a bounded rectangle in $\mathbb{R}^2$ and that the density of the measure $\mu$ is of the form

$$\varrho(x) = \sum_{i \in I} a_i 1_{Q_i}(x)$$

for $x \in \mathcal{X}$, where we assume that $I$ is a finite index set and $(Q_i)_{i \in I}$ is a partition of the domain $\mathcal{X}$ into (small) squares, called pixels, of equal side length. This is natural if $\varrho$ is given as a grayscale image and we would then typically index the pixels $Q_i$ by their centers $i \in I \subset \mathbb{Z}^2$. It may also serve as an approximation for arbitrary $\varrho$. It is however easy enough to adapt the following considerations to more general (not necessarily congruent) tiles and to obtain better approximations if the function $\varrho$ is specified more generally than piecewise constant.

The optimization procedure requires the non-trivial evaluation of $\Phi$ at a given weight vector $w$. This includes the integration over Voronoi cells and therefore the construction of a weighted Voronoi diagram. The latter task is solved by the package 2D Apollonius Graphs as part of the Computational Geometry Algorithms Library [8]. The integrals we need to compute are

$$\int_{\text{Vor}_w(j)} \rho(x) \, dx \quad \text{and} \quad \int_{\text{Vor}_w(j)} \|x - y_j\| \rho(x) \, dx.$$  

By definition the boundary of a Voronoi cell $\text{Vor}_w(j)$ is made up of hyperbola segments, each between $y_j$ and one of the other support points of $\nu$. The integration could be performed by drawing lines from $y_j$ to the end points of those segments and integrating over the resulting triangle-shaped areas separately. This would be executed by applying an affinely-linear transformation that moves the hyperbola segment onto the hyperbola $y = 1/x$ to both the area and the function we want to integrate. The required transformation can be found in [16, Section 5.6].

However, we take a somewhat more crude, but also more efficient path here, because it is a quite time-consuming task to decide which pixels intersect which weighted Voronoi cells and then to compute the (areas of the) intersections. We therefore approximate the intersections by splitting the pixels into a number of subpixels (unless the former are already very small) and assuming that each of them is completely contained in the Voronoi cell in which its center lies. This reduces the problem from computing intersections to determining the corresponding cell for each center, which the data structure used for storing the Voronoi diagram enables us to do in roughly $O(\log n)$ time; see [18]. The operation can be performed even more efficiently: When considering a subpixel other than the very first one, we already know the cell that one of the neighboring subpixel’s
center belongs to. Hence, we can begin our search at this cell which is either already the cell we are looking for or lies very close to it.

The downside of this approximation is that it can make the L-BFGS algorithm follow search directions along which the value of $\Phi$ cannot be decreased even though there exist different directions that allow a decrease. This usually only happens near a minimizing weight vector and can therefore be controlled by choosing a not too strict stopping criterion, see the next subsection.

5.3 Our implementation

Implementing the algorithm described in this section requires two technical choices: The number of subpixels every pixel is being split into and the stopping criterion for the minimization of $\Phi$. We found that choosing the number of subpixels to be the smallest integer such that their total number is larger than or equal to $1000n$ gives a good compromise between performance and precision.

The stopping criterion is implemented as follows: We terminate the optimization process once $\|\nabla \Phi(w)\|/2 \leq \epsilon$ for some $\epsilon > 0$. Due to Theorem 3c) this criterion yields an intuitive interpretation: $\|\nabla \Phi(w)\|/2$ is the amount of mass that is being mistransported, i.e. the total amount of mass missing or being in surplus at some $\nu$-location $y_i$ when transporting according to the current tesselation. In our experience this mass is typically rather proportionally distributed among the different cells and tends to be assigned in a close neighbourhood of the correct cell rather than far away. So even with somewhat larger $\epsilon$, the computed Wasserstein distance and the overall visual impression of the optimal transport partition remain mostly the same. In the numerical examples in Section 6 we choose the value of $\epsilon = 0.05$.

We implemented the algorithm in C++ and make it available on GitHub\textsuperscript{1} under the MIT license. Our implementation uses libLBFGS \cite{26} for the L-BFGS procedure and the geometry library CGAL \cite{8} for the construction and querying of weighted Voronoi tesselations. The repository also contains a Matlab script to visualize such tesselations.

6 Numerical examples

We investigate two concrete problem settings in order to better understand the workings and performance of our algorithm as well as to illustrate various theoretical aspects pointed out in the remainder of the paper. The first setting concerns the comparison of two bivariate normal distributions, the second setting a specific resource allocation problem in a geographic region.

6.1 Optimal transport between two normal distributions

We consider two bivariate normal distributions $\mu = \text{MVN}_2(a, \sigma^2 I_2)$ and $\nu = \text{MVN}_2(b, \sigma^2 I_2)$, where $a = 0.8 \cdot 1$, $b = 2.2 \cdot 1$ and $\sigma^2 = 0.1$, i.e. they both have the same spherical covariance matrix such that one distribution is just a displacement of the other. For computations we have truncated both measures to the set $\mathcal{X} = [0, 3]^2$. By discretization (quantization) a measure $\tilde{\nu}$ is obtained from $\nu$. We then compute the transport tesselations and the Wasserstein distances between $\mu$ and $\tilde{\nu}$ for both $p = 1$ and $p = 2$. Computations and

\textsuperscript{1}https://github.com/valentin-hartmann-research/semi-discrete-transport
plots for $p = 2$ are obtained with the package transport \cite{33} for the statistical computing environment R \cite{25}. For $p = 1$ we use our implementation presented in the previous section.

Note that for the original problem of optimal transport from $\mu$ to $\nu$ the solution is known exactly, so we can use this example to investigate the correct working of our implementation. In fact, for any probability measure $\mu'$ on $\mathbb{R}^d$ and its displacement $\nu' = T \circ \mu'$, where $T: \mathbb{R}^2 \to \mathbb{R}^2, x \mapsto x + (b - a)$ for some vector $b - a \in \mathbb{R}^d$, it is immediately clear that the translation $T$ induces an optimal transport plan for $\mu'$ and that $W_p(\mu',\nu') = \|b - a\|$ for arbitrary $p \geq 1$. This holds because we obtain by Jensen’s inequality $\mathbb{E}(\|X - Y\|^p)^{\frac{1}{p}} \geq \mathbb{E}(\|X - Y\|) = \|b - a\|$ for $X \sim \mu', Y \sim \nu'$; therefore $W_p(\mu',\nu') \geq \|b - a\|$ and $T$ is clearly a transport map from $\mu'$ to $\nu'$ that achieves this lower bound. For $p = 2$ Theorem 9.4 in \cite{30} yields that $T$ is the unique optimal transport map and the induced plan $\pi_T$ is the unique optimal transport plan. In the case $p = 1$ neither of these objects is unique due to the possibility to rearrange mass transported within the same line segment at no extra cost.

Discretization was performed by applying the weighted $K$-means algorithm based on the discretization of $\mu$ to a fine grid and an initial configuration of cluster centers drawn independently from distribution $\nu$ and equipped with the corresponding density values of $\nu$ as weights. The number of cluster centers was kept to $n = 300$ for better visibility in the plots below. We write $\hat{\nu} = \sum_{i=1}^{n} \delta_{y_i}$ for the discretized measure. The discretization error can be computed numerically by solving another semi-discrete transport problem, see the third column of Table 1 below.

The first column of Figure 4 depicts the measures $\mu$ and $\hat{\nu}$ and the resulting optimal transport partitions for $p = 1$ and $p = 2$. In the case $p = 1$ the nuclei of the weighted Voronoi tesselation are always contained in their cells, whereas for $p = 2$ this need not be the case. We therefore indicate the relation by a gray arrow pointing from the centroid of the cell to its nucleus whenever the nucleus is outside the cell. The theory for the case $p = 2$, see e.g. \cite{22} Section 2], identifies the tesselation as a Laguerre tesselation (or power diagram), which consists of convex polygons.

The partitions obtained for $p = 1$ and $p = 2$ look very different, but they both capture optimal transports along the direction $b - a$ very closely. For $p = 2$ we clearly see a close approximation of the optimal transport map $T$ introduced above. For $p = 1$ we see an approximation of an optimal transport plan $\pi$ that collects the mass for any $y \in \mathcal{Y}$ somewhere along the way in the direction $b - a$.

The second column of Table 1 gives the Wasserstein distances computed numerically based on these partitions. Both of them are very close to the theoretical value of $\|b - a\| = \sqrt{2} \cdot 1.4 \approx 1.979899$, and in particular they are well inside the boundaries set by the approximation error.

|                | MVN vs. MVN | MVN + Leb vs. MVN + Leb |
|----------------|-------------|--------------------------|
|                | theoretical | computed | discr. error | theoretical | computed | discr. error |
| $p=1$          | 1.979899    | 1.965988 | 0.030962     | 1.979899    | 2.164697 | 0.653370     |
| $p=2$          | 1.979899    | 1.965753 | 0.034454     | unknown     | 0.827809 | 0.220176     |

Table 1: Theoretical continuous and computed semidiscrete Wasserstein distances, together with the discretization error.

We also investigate the effect of adding a common measure to both $\mu$ and $\nu$: Let $\alpha = \text{Leb}^d|_\mathcal{X}$ and proceed in the same way as above for the measures $\mu' = \mu + \alpha$ and $\nu' = \nu + \alpha$, calling the discretized measure $\hat{\nu}'$. Note that the discretization error (sixth
The second column of Table 1 is considerably higher, on the one hand due to the fact that the $n = 300$ support points of $\nu'$ have to be spread far wider, on the other hand because the total mass of each measure is 10 now compared to 1 before.

The second column of Figure 4 depicts the measures $\mu'$ and $\nu'$ and the resulting optimal transport partitions for $p = 1$ and $p = 2$. Both partitions look very different from their counterparts when no $\alpha$ is added. However the partition for $p = 1$ clearly approximates a transport plan along the direction of $b - a$ again. Note that the movement of mass is much more local, meaning the approximated optimal transport plan is not just $\pi + \alpha\Delta$, where $\pi$ is the optimal transport plan approximated in Figure 4(c), but a substantial amount of mass available from $\alpha$ is moved as well. Furthermore, Figure 4(d) gives the impression of a slightly curved movement of mass. We attribute this to a combination of a boundary effect from trimming the Lebesgue measure to $\mathcal{X}$ and numerical error based on the coarse discretization and a small amount of mistransported mass.

The computed $W_1$-value for this new example (last column of Table 1) lies in the vicinity of the theoretical value again if one allows for the rather large discretization error.

The case $p = 2$ exhibits the distinctive curved behaviour that is often seen in $L_2$-optimal transport maps, cf. Subsection 2.1. Various of the other points mentioned in Section 2 can be observed as well, e.g. the numerically computed Wasserstein distance is much smaller than for $p = 1$, which is consistent with (4) and (6), and seems plausible in view of Remark 1.

### 6.2 A practical resource allocation problem

We revisit the delivery problem described in the introduction. A fast-food delivery service has 32 branches throughout a city area, depicted by the black dots on the map in Figure 5. For simplicity of representation we assume that most branches have the same fixed production capacity and a few individual ones (marked by an extra circle around the dot) have twice that capacity. We assume further that the expected orders at peak times have a spatial distribution as indicated by the heatmap (where yellow to white means higher number of orders) and a total volume that matches the total capacity of the branches. The task of the fast-food chain is to partition the map into 32 delivery zones, matching expected orders in each zone with the capacity of the branches, in such a way that the expected cost in form of the travel distance between branch and customer is minimal. We assume here the Euclidean distance, either because of a street layout that comes close to it, see e.g. [7], or because the deliveries are performed by UAVs (drones). The desired partition, computed by our implementation described in Subsection 5.3, is also displayed in Figure 5. A number of elongated cells in the western and central parts of the city area suggest that future expansions of the fast-food chain should concentrate on the city center in the north.

To convey an impression of the speed of our algorithm, we have generated 100 times 250, 500 and 1000 branch locations on the map of Figure 5 and capacities in $[0, 1]$, both uniformly at random. Table 2 shows the average runtimes for computing the optimal transport partition from the $\mu$-measure in Figure 5 (at a resolution of $646 \times 345$ pixels) to the $\nu$-measures of randomly generated branches. Computations were performed on one core of a mobile Intel Core i7. For some (rare) instances the target $\varepsilon$-value of 0.05 could not be reached, in which case the algorithm was restarted at a new value until the target was attained. The extra time was added to the total for those instances.
Figure 4: Left column: Semi-discrete transport between a bivariate normal distribution $\mu$ and a discretized bivariate normal distribution $\tilde{\nu}$. Right column: Same with Lebesgue measures added to both distributions (before discretization). Panels (a) and (b) illustrate the measures. The densities of the continuous measures $\mu$ and $\mu'$ are displayed as gray level images, the point masses of the discrete measures $\nu$ and $\nu'$ are shown as small discs with areas proportional to the masses placed there. Panels (c) to (f) show the optimal transport partitions.
7 Discussion and outlook

We have given a comprehensive account on semi-discrete optimal transport for the Euclidean cost function, arguing that there are sometimes good reasons to prefer Euclidean over squared-Euclidean cost and showing that for the Euclidean case the semi-discrete setting is particularly nice because we obtain a unique solution to the Monge–Kantorovich problem that is induced by a transport map. We have provided a reasonably fast algorithm that is similar to the AHA-algorithm described in detail in [22] but adapted in various aspects to the current situation of \( p = 1 \).

Apart from minor numerical problems detailed earlier, our algorithm converges towards the optimal partition subject to the convergence conditions for the L-BFGS algorithm; see e.g. [25]. Very loosely, such conditions state that we start in a region around the minimizer where the objective function \( \Phi \) shows to some extent quadratic behaviour. Similar to the AHA-algorithm in [22], a proof of such conditions is not available. In practice, the algorithm has converged in all the examples given in the present paper, but a few of the extremer \( \nu \)-measures obtained by simulation when compiling Table 2 necessitated one or more restarts of the algorithm; see Subsection 6.2.

There are several avenues for further research, both with regard to improving speed and robustness of the algorithm and for solving more complicated problems where our algorithm may be useful. Some of them are:

- As mentioned earlier, it may well be that the choice of our starting value is too simplistic and that faster convergence is obtained more often if the sequence \( \nu = \nu(0), \ldots, \nu(L) \) of coarsenings is e.g. based on the K-median algorithm or a similar method. The difficulty lies in finding \( \nu(l-1) \) that makes \( W_1(\nu(0), \nu(l-1)) \) substantially smaller without investing too much time in its computation.
• We currently keep the threshold \( \varepsilon \) in the stopping criterion of the multiscale approach in Subsection 5.1 fixed. Another alleviation of the computational burden may be obtained by choosing a suitable sequence \( \varepsilon_L, \ldots, \varepsilon_0 \) of thresholds for the various scales. It seems particularly attractive to use for the threshold at the coarser scale a value \( \varepsilon_l > 0 \) that is smaller than the value \( \varepsilon_{l-1} \) at the finer scale, especially for the last step, where \( l = 1 \). The rationale is that at the coarser scale we do not run easily into numerical problems and still reach the stricter \( \varepsilon_l \)-target efficiently. The obtained weight vector is expected to result in a better starting solution for the finer problem that reaches the more relaxed threshold \( \varepsilon_{l-1} \) more quickly than a starting solution stemming from an \( \varepsilon_{l-1} \)-target at the coarser scale.

• The L-BFGS algorithm used for the optimization process may be custom-tailored to our discretization of \( \mu \) in order to reach the theoretically maximal numerical precision that the discretization allows. It could e.g. use simple gradient descent from the point on where L-BFGS cannot minimize \( \Phi \) any further since even in the discretized case the gradient always points in a descending direction.

• The approximation of \( \mu \) by a fine-grained discrete measure has shown good results. However, as mentioned earlier, higher numerical stability and precision at the expense of complexity and possibly runtime could be obtained by a version of our algorithm that computes the intersections between the Voronoi cells and the pixels of \( \mu \) exactly.

• It appears to be a largely open question under what conditions on a sequence \( (\nu^{(k)}) \) of discrete measures that converges weakly towards a continuous measure \( \nu \), we obtain weak convergence of the optimal transport plans \( \pi^*_k \) from \( \mu \) to \( \nu^{(k)} \) towards one of the optimal transport plans \( \pi_* \) from \( \mu \) to \( \nu \) and how to characterize the \( \pi_* \) obtained. One special \( \pi_* \) which in view of the examples in Subsection 6.1 seems well within reach is the ray-monotone optimal transport plan; see [30], Subsection 3.1.5 and especially the open question on the stability of ray-monotone optimal transport on p. 104. It would be particularly desirable to have conditions on \( (\nu^{(k)}) \) under which \( \pi^{(k)}_* \) approximates the ray-monotone optimal transport from \( \mu \) to \( \nu \) as \( k \to \infty \).

• Semi-discrete optimal transport may be used as an auxiliary step in a number of algorithms for more complicated problems. The most natural example is a simple alternating scheme for the capacitated location-allocation (or transportation-location) problem; see [2]. Suppose that our fast-food chain from Subsection 6.2 has not entered the market yet and would like to open \( n \) branches anywhere in the city and divide up the area into delivery zones in such a way that (previously known) capacity constraints of the branches are met and the expected cost in terms of travel distance is minimized again. A natural heuristic algorithm would start with a random placement of \( n \) branches and alternate between capacitated allocation of expected orders (the continuous measure \( \mu \)) using our algorithm described in Section 5 and the relocation of branches to the spatial medians of the zones. The latter can be computed by discrete approximation, see e.g. [11], and possibly by continuous techniques, see [12] for a vantage point.
References

[1] Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. Preprint. https://arxiv.org/abs/1705.09634v1

[2] Ambrosio, L. and Pratelli, A. (2003). Existence and stability results in the $L^1$ theory of optimal transportation. In Optimal transportation and applications (Martina Franca, 2001), volume 1813 of Lecture Notes in Math., pages 123–160. Springer, Berlin.

[3] Aude, G., Cuturi, M., Peyré, G., and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. Preprint. https://arxiv.org/abs/1605.08527v1

[4] Aurenhammer, F., Hoffmann, F., and Aronov, B. (1998). Minkowski-type theorems and least-squares clustering. Algorithmica, 20(1):61–76.

[5] Basua, S., Kolouria, S., and Rohde, G. K. (2014). Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. PNAS, 111(9):3448—3453.

[6] Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numer. Math., 84:375–393.

[7] Boscoe, F. P., Henry, K. A., and Zdeb, M. S. (2012). A nationwide comparison of driving distance versus straight-line distance to hospitals. Prof. Geogr., 64(2):188–196.

[8] CGAL2015 (2015). CGAL, Computational Geometry Algorithms Library (Version 4.6.1). http://www.cgal.org

[9] Cooper, L. (1972). The transportation-location problem. Oper. Res., 20(1):94–108.

[10] Courty, N., Flamary, R., Tuia, D., and Corpetti, T. (2016). Optimal transport for data fusion in remote sensing. In 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pages 3571–3574.

[11] Croux, C., Filzmoser, P., and Fritz, H. (2012). A comparison of algorithms for the multivariate L1-median. Computational Statistics, 27(3):393–410.

[12] Fekete, S. P., Mitchell, J. S. B., and Beurer, K. (2005). On the continuous Fermat-Weber problem. Operations Research, 53(1):61–76.

[13] Geiß, D., Klein, R., Penninger, R., and Rote, G. (2013). Optimally solving a transportation problem using Voronoi diagrams. Comput. Geom., 46(8):1009–1016.

[14] Gramfort, A., Peyré, G., and Cuturi, M. (2015). Fast optimal transport averaging of neuroimaging data. In 24th International Conference on Information Processing in Medical Imaging (IPMI 2015), volume 9123 of Lecture Notes in Computer Science, pages 123–160.

[15] Guo, J., Pan, Z., Lei, B., and Ding, C. (2017). Automatic color correction for multisource remote sensing images with Wasserstein CNN. Rem. Sens., 9(5):1–16 (electronic).

[16] Hartmann, V. (2016). A geometry-based approach for solving the transportation problem with Euclidean cost. Bachelor’s thesis, Institute of Mathematical Stochastics, University of Göttingen. https://arxiv.org/abs/1706.07403
[17] Kantorovich, L. (1942). On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.),* 37:199–201.

[18] Karavelas, M. I. and Yvinec, M. (2002). Dynamic additively weighted Voronoi diagrams in 2D. In *Algorithms – ESA 2002*, pages 586–598. Springer.

[19] Kitagawa, J., Mérigot, Q., and Thibert, B. (2017). Convergence of a Newton algorithm for semi-discrete optimal transport. *J. Eur. Math. Soc., to appear.* [https://arxiv.org/abs/1603.05579v2](https://arxiv.org/abs/1603.05579v2)

[20] Luenberger, D. G. and Ye, Y. (2008). *Linear and nonlinear programming.* International Series in Operations Research & Management Science, 116. Springer, New York, third edition.

[21] McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.,* 80(2):309–323.

[22] Mérigot, Q. (2011). A multiscale approach to optimal transport. *Comput. Graph. Forum,* 30(5):1583–1592.

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

[24] Nicolas, P. (2016). Optimal transport for image processing. Habilitation thesis, Signal and Image Processing, Université de Bordeaux. [https://hal.archives-ouvertes.fr/tel-01246096v6](https://hal.archives-ouvertes.fr/tel-01246096v6).

[25] Nocedal, J. (1980). Updating quasi-Newton matrices with limited storage. *Mathematics of computation,* 35(151):773–782.

[26] Okazaki, N. and Nocedal, J. (2010). libLBFGS (Version 1.10). [http://www.chokkan.org/software/liblbfgs/](http://www.chokkan.org/software/liblbfgs/)

[27] Pratelli, A. (2007). On the equality between Monge’s infimum and Kantorovich’s minimum in optimal mass transportation. *Ann. Inst. H. Poincaré Probab. Statist.,* 43(1):1–13.

[28] R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. Version 3.3.0. [https://www.R-project.org/](https://www.R-project.org/)

[29] Rippl, T., Munk, A., and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivar. Anal.,* 151:90–109.

[30] Santambrogio, F. (2015). *Optimal transport for applied mathematicians,* volume 87 of *Progress in Nonlinear Differential Equations and their Applications.* Birkhäuser/Springer, Cham. Calculus of variations, PDEs, and modeling.

[31] Schmitzer, B. (2016). A sparse multiscale algorithm for dense optimal transport. *J Math Imaging Vis,* 56(2):238–259.

[32] Schmitzer, B. and Wirth, B. (2017). A framework for Wasserstein-1-type metrics. *Preprint.* [https://arxiv.org/abs/1701.01945v1](https://arxiv.org/abs/1701.01945v1)
[33] Schuhmacher, D., Bähre, B., Gottschlich, C., and Heinemann, F. (2017). *transport: Optimal Transport in Various Forms*. R package version 0.8-2. [https://cran.r-project.org/package=transport](https://cran.r-project.org/package=transport).

[34] Sherali, H. D. and Nordai, F. L. (1988). NP-hard, capacitated, balanced p-median problems on a chain graph with a continuum of link demands. *Math. Oper. Res.*, 13(1):32–49.

[35] Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11.

[36] Villani, C. (2009). *Optimal transport, old and new*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

[37] Wolansky, G. (2015). Semi-discrete approximation of optimal mass transport. Preprint. [https://arxiv.org/abs/1502.04309v1](https://arxiv.org/abs/1502.04309v1).

[38] Wolfe, P. (1969). Convergence conditions for ascent methods. *SIAM Rev.*, 11:226–235.

---

Valentin Hartmann  
Institute for Mathematical Stochastics  
University of Goettingen  
Goldschmidtstr. 7  
37077 Göttingen, Germany  
*Present address:*  
Department of Mathematics  
Technical University of Munich  
Boltzmannstr. 3  
85747 Garching bei München, Germany  
*E-mail:* v.hartmann.research@gmail.com

Dominic Schuhmacher  
Institute for Mathematical Stochastics  
University of Goettingen  
Goldschmidtstr. 7  
37077 Göttingen, Germany  
*E-mail:* schuhmacher@math.uni-goettingen.de