Mathematical Methods of Operations Research
https://doi.org/10.1007/s00186-020-00703-z

ORIGINAL ARTICLE

Semi-discrete optimal transport: a solution procedure for the unsquared Euclidean distance case

Valentin Hartmann[1,2] · **Dominic Schuhmacher**[1]

Received: 22 May 2019 / Revised: 20 December 2019

Abstract
We consider the problem of finding an optimal transport plan between an absolutely continuous measure and a finitely supported measure of the same total mass when the transport cost is the unsquared Euclidean distance. We may think of this problem as closest distance allocation of some resource continuously distributed over Euclidean space to a finite number of processing sites with capacity constraints. This article gives a detailed discussion of the problem, including a comparison with the much better studied case of squared Euclidean cost. We present an algorithm for computing the optimal transport plan, which is similar to the approach for the squared Euclidean cost by Aurenhammer et al. (Algorithmica 20(1):61–76, 1998) and Mérigot (Comput Graph Forum 30(5):1583–1592, 2011). We show the necessary results to make the approach work for the Euclidean cost, evaluate its performance on a set of test cases, and give a number of applications. The later include goodness-of-fit partitions, a novel visual tool for assessing whether a finite sample is consistent with a posited probability density.

Keywords Monge–Kantorovich problem · Spatial resource allocation · Wasserstein metric · Weighted Voronoi tessellation

Mathematics Subject Classification Primary 65D18; Secondary 51N20 · 62-09

VH was partially supported by Deutsche Forschungsgemeinschaft RTG 2088. We thank Marcel Klatt for helpful discussions and three anonymous referees for comments that led to an improvement of the paper.

✉ Dominic Schuhmacher
schuhmacher@math.uni-goettingen.de

Valentin Hartmann
valentin.hartmann@epfl.ch

[1] Institute for Mathematical Stochastics, University of Goettingen, Goldschmidtstr. 7, 37077 Goettingen, Germany

[2] Present Address: IC IINFCOM DLAB, EPFL, Station 14, 1015 Lausanne, Switzerland

Published online: 12 February 2020

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 (Peyré and Cuturi 2018) 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. Schmitzer (2016, Sect. 6), and to approximate such distances (to some extent) on larger and/or irregular structures, see Altschuler et al. (2017) and references therein. The development of new methodology for data analysis based on optimal transport is a booming research topic in statistics and machine learning, see e.g. Sommerfeld and Munk (2018), Schmitz et al. (2018), Arjovsky et al. (2017), Genevay et al. (2018), and Flamary et al. (2018). Applications are abundant throughout all of the applied sciences, including biomedical sciences (e.g. microscopy or tomography images; Basua et al. 2014, Gramfort et al. 2015), geography (e.g. remote sensing; Courty et al. 2016, Guo et al. 2017), and computer science (e.g. image processing and computer graphics; Nicolas 2016, Solomon et al. 2015). 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.

More formally, for measures μ and ν 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

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

where the minimum is taken over all *transport plans (couplings)* π between μ and ν, i.e. measures π 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 Villani (2009, Theorem 4.1) and it is readily verified, see e.g. Villani (2009, 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* (Kantorovich 1942; Villani 2009). From an intuitive point of view, a minimizing π describes how the mass of μ is to be associated with the mass of ν in order to make the overall transport cost minimal.

A *transport map* from μ to ν 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 π *deterministic* in that case. It is easily seen that the support of π_T is contained in the graph of T. Intuitively speaking, we associate with each location in the domain of the measure μ exactly one location in the domain of the measure ν to which positive mass is moved, i.e. the mass of μ is not split.

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

$$\inf_T \int_{\mathbb{R}^d} \|x - T(x)\|^p \, \mu(dx) = \inf_T \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, \pi_T(dx, dy), \tag{2}$$

where the infima are taken over all transport maps T from μ to ν (and are in general not attained) is known as *Monge's problem* (Monge 1781; Villani 2009).

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 (Luenberger and Ye 2008):

$$W_p(\mu, \nu) = \left(\min_{(\pi_{ij})} \sum_{i=1}^m \sum_{j=1}^n d_{ij} \pi_{ij} \right)^{1/p}, \tag{3}$$

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 π_{ij} satisfying

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

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 1000×1000 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 (Schmitzer 2016), 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. Approximative solutions can be computed for this order of magnitude and $p = 2$ by variants of the celebrated Sinkhorn algorithm (Cuturi 2013; Schmitzer 2019; Altschuler et al. 2017), but it has been observed that these approximations have their limitations (Schmitzer 2019; Klatt et al. 2019).

The main advantage of using $p = 2$ is that we can decompose the cost function as $\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2x^\top y$ and hence formulate the Monge–Kantorovich problem equivalently as $\max_\pi \int_{\mathbb{R}^d \times \mathbb{R}^d} x^\top y \; \pi(dx, dy)$. For the discrete problem (3) this decomposition is used in Schmitzer (2016) to construct particularly simple so-called shielding neighborhoods. But also if one or both of μ and ν are assumed absolutely continuous with respect to Lebesgue measure, this decomposition for $p = 2$ has clear computational advantages. For example if the measures μ and ν 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 (McCann 1995), 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 Santambrogio (2015, Sect. 6.3) and the references given there.

In the rest of this article we focus on the semi-discrete setting, meaning here that the measure μ is absolutely continuous with respect to Lebesgue measure and the measure ν has finite support. This terminology was recently used in Wolansky (2015), Kitagawa et al. (2019), Genevay et al. (2016) and Bourne et al. (2018) among others. In the semi-discrete setting we can represent a solution to Monge's problem as a partition of $\mathbb{R}^d$, where each cell is the pre-image of a support point of ν under the optimal transport map. We refer to such a partition as *optimal transport partition*.

In the case $p = 2$ this setting is well studied. It was shown in Aurenhammer et al. (1998) that an optimal transport partition always exists, is essentially unique, and takes the form of a Laguerre tessellation, a.k.a. power diagram. The authors proved further that the right tessellation can be found numerically by solving a (typically high dimensional) unconstrained convex optimization problem. Since Laguerre tessellations are composed of convex polytopes, the evaluation of the objective function can be done very precisely and efficiently. Mérigot (2011) elaborates details of this algorithm and combines it with a powerful multiscale idea. In Kitagawa et al. (2019) a damped Newton algorithm is presented for the same objective function and the authors are able to show convergence with optimal rates.

In this article we present the corresponding theory for the case $p = 1$. It is shown in Sect. 2.3 of Crippa et al. (2009) and independently in Geiß et al. (2013), which both treat more general cost functions, that an optimal transport partition always exists, is essentially unique and takes the form of a weighted Voronoi tessellation, or more precisely an Apollonius diagram. We extend this result somewhat within the case $p = 1$ in Theorems 1 and 2 below. We prove then in Theorem 3 that the right tessellation can be found by optimizing an objective function corresponding to that from the case $p = 2$. Since the cell boundaries in an Apollonius diagram in 2d are segments of hyperbolas, computations are more involved and we use a new strategy for computing integrals over cells and for performing line search in the optimization method. Details of the algorithm are given in Sect. 4 and the complete implementation can be downloaded from Github[1] and is included in the latest version of the `transport`-package (Schuhmacher et al. 2019) for the statistical computing environment R (R Core Team 2017). Up to Sect. 4 the present paper is a condensed version of the thesis (Hartmann 2016), to which we refer from time to time for more details. In the remainder we evaluate the performance of our algorithm on a set of test cases (Sect. 5), give a number of applications (Sect. 6), and provide a discussion and open questions for further research (Sect. 7).

At the time of finishing the present paper, it has come to our attention that Theorem 2.1 of Kitagawa et al. (2019), which is for very general cost funtions including the Euclidean distance (although the remainder of the paper is not), has a rather large overlap with our Theorem 3. Within the case of Euclidean cost it assumes somewhat

[1] https://github.com/valentin-hartmann-research/semi-discrete-transport.

stronger conditons than our Theorem 3, namely a compact domain $\mathscr{X}$ and a bounded density for μ. In addition the statement is somewhat weaker as it does not contain our statement (c). We also believe that due to the simpler setting of $p = 1$ our proof is accessible to a wider audience and it is more clearly visible that the additional restrictions on $\mathscr{X}$ and μ are in fact not needed.

We end this introduction by providing some first motivation for studying the semi-discrete setting for $p = 1$. This will be further substantiated in the application Sect. 6.

1.1 Why semi-discrete?

The semi-discrete setting appears naturally in problems of allocating a continuously distributed resource to a finite number of sites. Suppose for example that a fast-food chain introduces a home delivery service. Based on a density map of expected orders (the "resource"), the management would like to establish delivery zones for each branch (the "sites"). We assume that each branch has a fixed capacity (at least in the short run), that the overall capacity matches the total number of orders (peak time scenario), and that the branches are not too densely distributed, so that the Euclidean distance is actually a reasonable approximation to the actual travel distance; see Boscoe et al. (2012). We take up this example in Sect. 6.2. A somewhat different model that adds waiting time costs to the distance-based costs instead of using capacity constraints was studied theoretically in Crippa et al. (2009).

An important general class that builds on resource allocation are location-allocation problems: where to position a number of sites (branches, service stations, etc.) in such a way that the sum of the resource allocation cost plus maybe further costs for installation, maintainance and waiting times is minimized, possibly under capacity and/or further constraints. See e.g. Mallozzi et al. (2019) for a flexible model, which was algorithmically solved via discretizing the continuous domain. Positioning of sites can also be competitive, involving different agents (firms), such as in Núñez and Scarsini (2016).

A special case of location-allocation is the quantization problem, which consists in finding positions and capacities of sites that minimize the resulting resource allocation cost. See Bourne et al. (2018, Sect. 4) for a recent discussion using incomplete transport and $p = 2$.

As a further application we propose in Sect. 6.3 optimal transport partitions as a simple visual tool for investigating local deviations from a continuous probability distribution based on a finite sample.

Since the computation of the semi-discrete optimal transport is linear in the resolution at which we consider the continuous measure (for computational purposes), it can also be attractive to use the semi-discrete setting as an approximation of either the fully continuous setting (if ν is sufficiently simple) or the fully discrete setting (if μ has a large number of support points). This will be further discussed in Sect. 2.

1.2 Why $p = 1$?

The following discussion highlights some of the strengths of optimal transport based on an unsquared Euclidean distance ($p = 1$), especially in the semi-discrete setting, and contrasts $p = 1$ with $p = 2$.

From a computational point of view the case $p = 2$ can often be treated more efficiently, mainly due to the earlier mentioned decomposability, leading e.g. to the algorithms in Schmitzer (2016) in the discrete and Aurenhammer et al. (1998), Mérigot (2011) in the semi-discrete setting. The case $p = 1$ has the advantage that the Monge–Kantorovich problem has a particularly simple dual (Villani 2009, Particular Case 5.16), which is equivalent to Beckmann's problem (Beckmann 1952; Santambrogio 2015, Theorem 4.6). If we discretize the measures (if necessary) to a common mesh of n points, the latter is an optimization problem in n variables rather than the n^2 variables needed for the general discrete transport formulation (3). Algorithms that make use of this reduction have been described in Solomon et al. (2014) (for general discrete surfaces) and in Schmitzer and Wirth (2019, Sect. 4) (for general incomplete transport), but their performance in a standard situation, e.g. complete optimal transport on a regular grid in $\mathbb{R}^d$, remains unclear. In particular we are not aware of any performance comparisons between $p = 1$ and $p = 2$.

In the present paper we do not make use of this reduction, but keep the source measure μ truly continuous except for an integral approximation that we perform for numerical purposes. We describe an algorithm for the semi-discrete problem with $p = 1$ that is reasonably fast, but cannot quite reach the performance of the algorithm for $p = 2$ in Mérigot (2011). This is again mainly due to the nice decomposition property of the cost function for $p = 2$ or, more blatantly, the fact that we minimize for $p = 2$ over partitions formed by line rather than hyperbola segments.

From an intuitive point of view $p = 1$ and $p = 2$ have both nice interpretations and depending on the application setting either the one or the other may be more justified. The difference is between thinking in terms of transportation logistics or in terms of fluid mechanics. If $p = 1$, the optimal transport plan minimizes the cumulative *distance* by which mass is transported. This is (up to a factor that would not change the transport plan) the natural cost in the absence of fixed costs or any other savings on long-distance transportation. If $p = 2$, the optimal transport plan is determined by a pressureless potential flow from μ to ν as seen from the kinetic energy minimization formulation of Benamou and Brenier (2000), Villani (2009, Chapter 7).

The different behaviors in the two cases can be illustrated by the discrete toy example in Fig. 1. Each point along the incomplete circle denotes the location of one unit of mass of μ (blue x-points) and/or ν (red o-points). The unique solution for $p = 1$ moves one unit of mass from one end of the circular structur to the other. This is how we would go about carrying boxes around to get from the blue scenario to the red scenario. The unique solution for $p = 2$ on the other hand is to transport each unit a tiny bit further to the next one, corresponding to a (discretized) flow along the circle. It is straightforward to adapt this toy example for the semi-discrete or the continuous setting. A more complex semi-discrete example is given in Sect. 6.1.

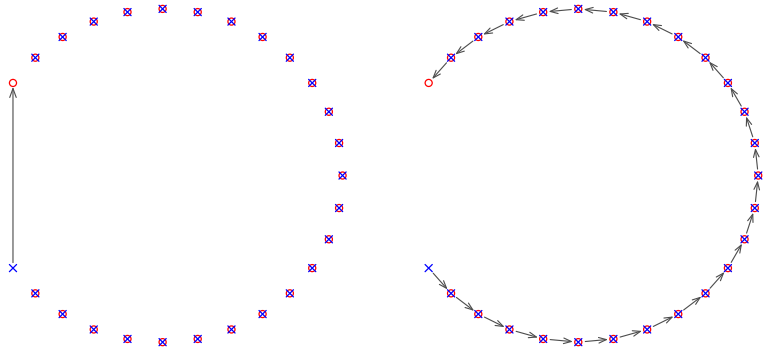


Fig. 1 Optimal transport maps from blue-x to red-o measure with unit mass at each point. Left: the transportation logistics solution ($p = 1$); right: the fluid mechanics solution ($p = 2$). (Color figure online)

One argument in favour of the metric W_1 is its nice invariant properties that are not shared by the other W_p. In particular, considering finite measures μ, ν, α on $\mathbb{R}^d$ satisfying $\mu(\mathbb{R}^d) = \nu(\mathbb{R}^d)$, $p \geq 1$ and $c > 0$, we have

$$W_1(\alpha + \mu, \alpha + \nu) = W_1(\mu, \nu), \tag{4}$$

$$W_1(c\mu, c\nu) = c\, W_1(\mu, \nu). \tag{5}$$

The first result is in general not true for any other p, the second result holds with a factor $c^{1/p}$ on the right hand side. We prove these statements in the appendix. These invariance properties have important implications for image analysis, where it is quite common to adjust for differing levels of brightness (in grayscale images) by affine transformations. While the above equalities show that it is safe to do so for $p = 1$, it may change the resulting Wasserstein distance and the optimal transport plan dramatically for other p; see Appendix and Sect. 6.1.

It is sometimes considered problematic that optimal transport plans for $p = 1$ are in general not unique. But this is not so in the semi-discrete case, as we will see in Sect. 2: the minimal transport cost in (1) is realized by a unique coupling π, which is always deterministic. The same is true for $p = 2$. A major difference in the case $p = 1$ is that for $d > 1$ each cell of the optimal transport partition contains the support point of the target measure ν that it assigns its mass to. This can be seen as a consequence of cyclical monotonicity (Villani 2009, beginning of Chapter 8). In contrast, for $p = 2$, optimal transport cells can be separated by many other cells from their support points, which can make the resulting partition hard to interpret without drawing corresponding arrows for the assignment; see the bottom panels of Fig. 5. For this reason we prefer to use $p = 1$ for the goodness-of-fit partitions considered in Sect. 6.3.

2 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 μ and ν be *probability* measures on

$\mathscr{X}$ and $\mathscr{Y}$, respectively. This is just for notational convenience and does not change the set of admissible measures in an essential way: we may always set $\mathscr{X} = \mathscr{Y} = \mathbb{R}^d$ and any statement about μ and ν 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 μ to be absolutely continuous with density ϱ with respect to d-dimensional Lebesgue measure Leb^d and to satisfy

$$\int_{\mathscr{X}} \|x\| \, \mu(dx) < \infty. \tag{6}$$

We assume further that $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$, where $n \in \mathbb{N}$, $y_1, \ldots, y_n \in \mathscr{Y}$ and $\nu_1, \ldots \nu_n \in (0, 1]$. Condition (6) guarantees that

$$W_1(\mu, \nu) \leq \int_{\mathscr{X}} \|x\| \, \mu(dx) + \int_{\mathscr{Y}} \|y\| \, \nu(dy) =: C < \infty, \tag{7}$$

which simplifies certain arguments.

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

We call a partition $\mathfrak{C} = (C_j)_{1 \leq j \leq n}$ of $\mathscr{X}$ into Borel sets satisfying $\mu(C_j) = \nu_j$ for every j a *transport partition* from μ to ν. Any such partition characterizes a transport map T from μ to ν, where we set $T_{\mathfrak{C}}(x) = \sum_{j=1}^n y_j 1\{x \in C_j\}$ for a given transport partition $\mathfrak{C} = (C_j)_{1 \leq j \leq n}$ and $\mathfrak{C}_T = (T^{-1}(y_j))_{1 \leq j \leq n}$ for a given transport map T. Monge's problem for $p = 1$ can then be equivalently formulated as finding

$$\inf_{\mathfrak{C}} \int_{\mathscr{X}} \|x - T_{\mathfrak{C}}(x)\| \, \mu(dx) = \inf_{\mathfrak{C}} \sum_{j=1}^n \int_{C_j} \|x - y_j\| \, \mu(dx), \tag{8}$$

where the infima are taken over all transport partitions $\mathfrak{C} = (C_j)_{1 \leq j \leq n}$ from μ to ν. Contrary to the difficulties encountered for more general measures μ and ν when considering Monge's problem with Euclidean costs, we can give a clear-cut existence and uniqueness theorem in the semi-discrete case, without any further restrictions.

Theorem 1 *In the semi-discrete setting with Euclidean costs (always including $d \geq 2$ and (6)) there is a μ-a.e. unique solution T_* to Monge's problem. The induced coupling π_{T_*} is the unique solution to the Monge–Kantorovich problem, yielding*

$$W_1(\mu, \nu) = \int_{\mathscr{X}} \|x - T_*(x)\| \, \mu(dx). \tag{9}$$

Proof The part concerning Monge's problem is a consequence of the concrete construction in Sect. 3; see Theorem 2.

Clearly π_{T_*} is an admissible transport plan for the Monge–Kantorovich problem. Since μ is non-atomic and the Euclidean cost function is continuous, Theorem B in Pratelli (2007) implies that the minimum in the Monge–Kantorovich problem is equal to the infimum in the Monge problem, so π_{T*} must be optimal.

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

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

$$\pi_i' = \pi_i - \pi_i|_{S_i \cap S_j \cap H^{i,j}_>(q)} + \frac{\alpha}{\beta}\pi_j|_{S_i \cap S_j \cap H^{i,j}_<(q)},$$
$$\pi_j' = \pi_j + \pi_i|_{S_i \cap S_j \cap H^{i,j}_>(q)} - \frac{\alpha}{\beta}\pi_j|_{S_i \cap S_j \cap H^{i,j}_<(q)}$$

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

$$\begin{aligned}
&\int_{\mathscr{X}\times\mathscr{Y}} \|x - y\| \, \pi'(dx, dy) - \int_{\mathscr{X}\times\mathscr{Y}} \|x - y\| \, \pi(dx, dy) \\
&\qquad = \int_{S_i \cap S_j \cap H^{i,j}_>(q)} \left(-\|x - y_i\| + \|x - y_j\|\right) \pi_i(dx) \\
&\qquad\quad + \frac{\alpha}{\beta} \int_{S_i \cap S_j \cap H^{i,j}_<(q)} \left(\|x - y_i\| - \|x - y_j\|\right) \pi_j(dx) \\
&\qquad < 0,
\end{aligned}$$

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

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 π 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 $x \in \bigcap_{i \in I} S_i \setminus \bigcap_{i \in I^c} S_i$, where $I \subset \{1, \ldots, n\}$ contains at least two elements and $i_0 = \min(I)$. It follows that $\pi_T = \pi$. But by the optimality of π and Theorem 2 we obtain $T = T_*$ μ-almost surely, which implies $\pi = \pi_T = \pi_{T_*}$. □

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 $\mathscr{Y}$ with density $\tilde{\varrho}$ with respect to Leb^d instead of the discrete measure ν. 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 μ, where $m \in \mathbb{N}$, $x_1, \ldots, x_m \in \mathscr{X}$ and $\tilde{\mu}_1, \ldots \tilde{\mu}_m \in (0, 1]$. In both cases existence of an optimal transport plan is still guaranteed by Villani (2009, 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 Sect. 6 of Ambrosio and Pratelli (2003). 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 Ambrosio and Pratelli (2003).

In the discrete case uniqueness depends strongly on the geometry of the support points of $\tilde{\mu}$ and ν. In addition to collinearity of support points, equality of interpoint distances can also lead to non-unique solutions. While uniqueness can typically be achieved when the support points are in sufficiently general position, we are not aware of any precise result to this effect.

When approximating the continuous problem with measures μ and $\tilde{\nu}$ by a semi-discrete problem, we quantize the measure $\tilde{\nu}$ into 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 $\mathrm{supp}(\tilde{\nu})$. The error we commit in Wasserstein distance by discretization of $\tilde{\nu}$ is bounded by the quantization error, i.e.

$$\left|W_1(\mu, \tilde{\nu}) - W_1(\mu, \nu)\right| \leq W_1(\tilde{\nu}, \nu) \leq \sum_{j=1}^n \int_{N_j} \|y - y_j\| \, \tilde{\nu}(dy). \tag{10}$$

We can compute $W_1(\tilde{\nu}, \nu)$ exactly by solving another semi-discrete transport problem, using the algorithm described in Sect. 4 to compute an optimal partition (N_j) for the second inequality above. However, choosing ν 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 Sect. 4.1 below.

When approximating the discrete problem with measures $\tilde{\mu}$ and ν 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 μ with density $\varrho(x) = \sum_{i=1}^m \tilde{\mu}_i f_i(x)$. Typical examples use $f_i(x) = \frac{1}{h^d} \varphi\left(\frac{x - x_i}{h}\right)$, where φ 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_is is very natural, where the M_is 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

$$\left|W_1(\tilde{\mu}, \nu) - W_1(\mu, \nu)\right| \leq W_1(\tilde{\mu}, \mu) \leq \sum_{i=1}^{m} \tilde{\mu}_i \int_{\mathbb{R}^d} \|x - x_i\| f_i(x)\, dx. \tag{11}$$

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 Villani (2009) implies for $\nu^{(k)} \to \tilde{\nu}$ weakly and $\mu^{(k)} \to \tilde{\mu}$ 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 π_* between μ and ν. 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 ε (in any fixed metrization of the weak topology) of *some* optimal transport plan between μ and ν, 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 π_*, this yields that $\pi_*^{(k)} \to \pi_*$ weakly.

3 Optimal transport maps via weighted Voronoi tessellations

As shown for bounded $\mathscr{X}$ in Geiß et al. (2013), the solution to the semi-discrete transport problem has a nice geometrical interpretation, which is similar to the well-known result in Aurenhammer et al. (1998): we elaborate below that the sets C_j^* of the optimal transport partition are the cells of an additively weighted Voronoi tessellation of $\mathscr{X}$ around the support points of ν.

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

$$\begin{aligned} \mathrm{Vor}_w(j) = \{x \in \mathscr{X} \,|\, &\|x - y_j\| - w_j \\ &\leq \|x - y_k\| - w_k \quad \text{for all } k \neq j\}, \quad j = 1, \ldots, n. \end{aligned}$$

Note that adjacent cells $\mathrm{Vor}_w(j)$ and $\mathrm{Vor}_w(k)$ have disjoint interiors. The intersection of their boundaries is a subset of $H = \{x \in \mathscr{X} \,|\, \|x - y_j\| - \|x - y_k\| = w_j - w_k\}$, which is easily seen to have Lebesgue measure (and hence μ-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 $S(y, w) = \{x \in \mathscr{X} \,|\, \|x - y\| = w\}$. See Fig. 2 for an illustration of these properties.

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

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

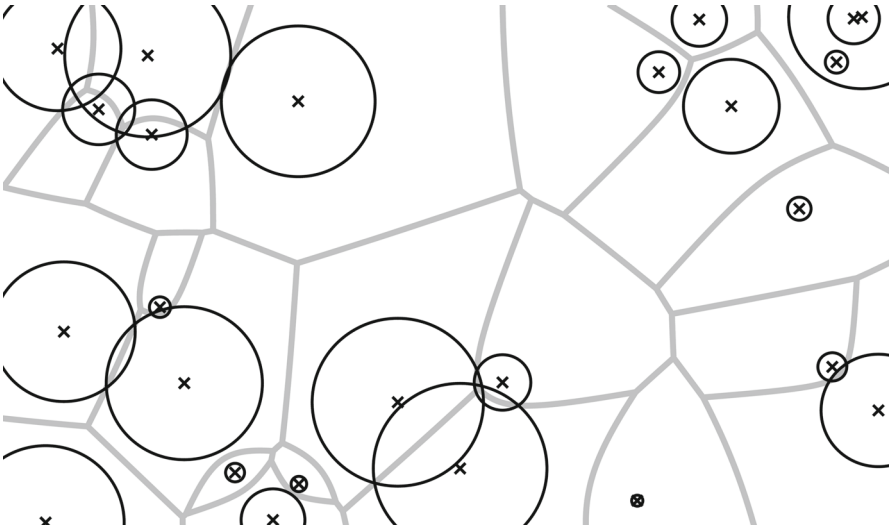


Fig. 2 An additively weighted Voronoi tessellation with 25 cells

A proof was given in Geiß et al. (2013), Theorem 2 for more general distance functions, but required $\mathscr{X}$ to be bounded. For the Euclidean distance we consider here, we can easily extend it to unbounded $\mathscr{X}$; see Hartmann (2016, Theorem 3.2).

Having identified this class of optimal transport partitions, it remains to show that for each pair (μ, ν) we can find an adapted weight vector. We adapt the approach of Aurenhammer et al. (1998) to the case $p = 1$, which gives us a constructive proof that forms the basis for the algorithm in Sect. 4. Our key tool is the function Φ defined below.

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

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

Then

a. Φ *is convex;*
b. Φ *is continuously differentiable with partial derivatives*

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

c. Φ *takes a minimum in* $\mathbb{R}^n$.

Remark 1 Let $w^* \in \mathbb{R}^n$ be a minimizer of Φ. Then by Theorem 3b)

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

i.e. w_* is adapted to (μ, ν). Theorem 2 yields that $(Vor_{w^*}(j))_{1\le j\le n}$ is the μ-almost surely unique optimal transport partition from μ to ν.

Proof (of Theorem 3) We take a few shortcuts; for full technical details see Chapter 3 of Hartmann (2016).

Part a) relies on the observation that Φ can be written as

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

where

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

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

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

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

By geometric arguments it can be shown that $[w \mapsto \mu(\mathrm{Vor}_w(j))]$ is continuous; see Hartmann (2016, Lemma 3.3). A short computation involving the representation $\Psi(w) = \inf_f \Psi_f(w)$ used above yields for the difference quotient of Ψ, writing e_j for the j-th standard basis vector and letting $h \neq 0$,

$$\left| \frac{\Psi(w + he_j) - \Psi(w)}{h} + \mu(\mathrm{Vor}_w(j)) \right| \le \left| -\mu(\mathrm{Vor}_{w+he_j}(j)) + \mu(\mathrm{Vor}_w(j)) \right| \longrightarrow 0$$

as $h \to 0$. This implies that Ψ is differentiable with continuous j-th partial derivative $-\mu(\mathrm{Vor}_w(j))$ and hence statement b) follows.

Finally, for the existence of a minimizer of Φ 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 Φ-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 Φ 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 Φ. 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 $(w^{(l)})_{l\in L}$ 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 $l \in L$

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, in the sense that l is the $\text{index}(l)$-th smallest element of 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 $N \in L$ such that

$$\sum_{j\in A} \mu\big(\text{Vor}_{w^{(l)}}(j)\big) = 1 \quad \text{for all } l \geq N.$$

Write

$$\underline{w}_A^{(l)} = \min_{i\in A} w_i^{(l)} \quad \text{and} \quad \overline{w}_B^{(l)} = \max_{i\in B} w_i^{(l)},$$

and recall the constant C from (7), which may clearly serve as an upper bound for the transport cost under an *arbitrary* plan. We then obtain for every $l \geq N$

$$\begin{aligned}
\Phi(w^{(l)}) &= \sum_{j=1}^{n} \left(-\nu_j w_j^{(l)} - \int_{\text{Vor}_{w^{(l)}}(j)} \big(\|x - y_j\| - w_j^{(l)}\big)\, \mu(dx) \right) \\
&\geq -C + \sum_{j=1}^{n} w_j^{(l)} \Big(\mu\big(\text{Vor}_{w^{(l)}}(j)\big) - \nu_j \Big) \\
&= -C + \sum_{j\in A} w_j^{(l)} \Big(\mu\big(\text{Vor}_{w^{(l)}}(j)\big) - \nu_j \Big) - \sum_{j\in B} w_j^{(l)} \nu_j \\
&\geq -C - R + \underline{w}_A^{(l)} \Big(1 - \sum_{j\in A} \nu_j \Big) - \overline{w}_B^{(l)} \sum_{j\in B} \nu_j \\
&\geq -C - 2R + \text{index}(l),
\end{aligned}$$

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 $\underline{w}^{(l)} = \min_{1\leq i\leq n} w_i^{(l)}$ from each $w_j^{(l)}$ for all $l \in L$. □

4 The algorithm

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

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

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

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

Newton's method would use $\Delta w^{(k)} = -\big(D^2\Phi(w^{(k)})\big)^{-1}\nabla\Phi(w^{(k)})$, 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 Mérigot (2011) for the case $p = 2$, we have obtained many good results using L-BFGS (Nocedal 1980), 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 for L-BFGS, we use the Armijo rule (Armijo 1966), 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 $\Phi(w^{(k)})$: the step size t_k needs to fulfill $\Phi(w^{(k)} + t_k\Delta w^{(k)}) \le \Phi(w^{(k)}) + c\,t_k\nabla\Phi(w^{(k)})^T\Delta w^{(k)}$ for a small fixed c with $0 < c < 1$. We use the default value $c = 10^{-4}$ of the L-BFGS library (Okazaki and Nocedal 2010) employed by our implementation, which is also given as an example by Nocedal and Wright (1999). An alternative that could be investigated is to use a non-monotone line search such as the one proposed in Grippo et al. (1986). There the above condition is relaxed by admitting a step whenever it sufficiently decreases a function value from one of the previous K iterations, for some $K \ge 1$. This might lead to fewer function evaluations and also to convergence in fewer steps. We also considered replacing the Armijo rule with the strong Wolfe conditions (1969, 1971) as done in Mérigot (2011), which contain an additional decrease requirement on the gradient. In our case, however, this requirement could often not be fulfilled because of the pixel splitting method used for computing the gradient (cf. Sect. 4.2), which made it less suited.

4.1 Multiscale approach to determine starting value

To find a good starting value $w^{(0)}$ we use a multiscale method similar to the one proposed in Mérigot (2011). We first create a decomposition of ν, 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 Sherali and Nordai (1988). 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 Mérigot (2011) (referred to as Lloyd's algorithm), which iteratively improves an initial aggregation of the support of $\nu^{(l-1)}$ into $|\operatorname{supp}(\nu^{(l)})|$ 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 Hartmann (2016). 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 Φ along the sequence of these coarsened measures, beginning at $\nu^{(L)}$ with the initial weight vector $w^{(L,0)} = 0 \in \mathbb{R}^{|\operatorname{supp}(\nu^{(L)})|}$ and computing the optimal weight vector $w^{(L,*)}$ for the transport from μ 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 Hartmann (2016).

4.2 Numerical computation of Φ and $\nabla\Phi$

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

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

for $x \in \mathscr{X}$, where we assume that I is a finite index set and $(Q_i)_{i \in I}$ is a partition of the domain $\mathscr{X}$ into (small) squares, called *pixels*, of equal side length. This is natural if ϱ 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 ϱ. 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 ϱ is specified more generally than piecewise constant.

The optimization procedure requires the non-trivial evaluation of Φ 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* (CGAL 2015). The integrals we need to compute are

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

By definition the boundary of a Voronoi cell $\mathrm{Vor}_w(j)$ is made up of hyperbola segments, each between y_j and one of the other support points of ν. 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 Hartmann (2016, Sect. 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 quadratic 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 $\mathcal{O}(\log n)$ time; see Karavelas and Yvinec (2002). 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 Φ cannot be sufficiently 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 for a given subpixel resolution, see the next subsection.

4.3 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 Φ. We found that choosing the number of subpixels to be the smallest square number 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 $\|\nabla\Phi(w)\|_1/2 \leq \varepsilon$ for some $\varepsilon > 0$. Due to Theorem 3b) this criterion yields an intuitive interpretation: $\|\nabla\Phi(w)\|_1/2$ is the amount of mass that is being mistransported, i.e. the total amount of mass missing or being in surplus at some ν-location y_i when transporting according to the current tessellation. In our experience this mass is typically rather proportionally distributed among the different cells and tends to be assigned in a close neighborhood of the correct cell rather than far away. So even with somewhat larger ε, the computed Wasserstein distance and the overall visual impression of the optimal transport partition remain mostly the same. In the numerical examples in Sects. 5 and 6 we choose the value of $\varepsilon = 0.05$.

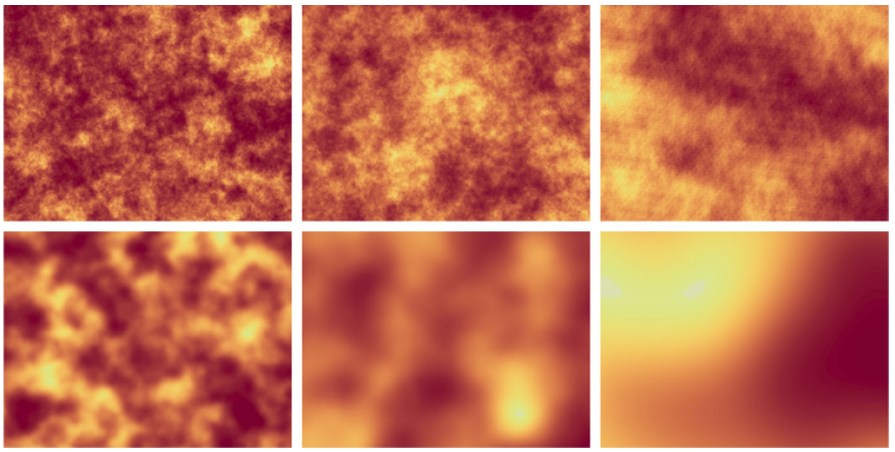


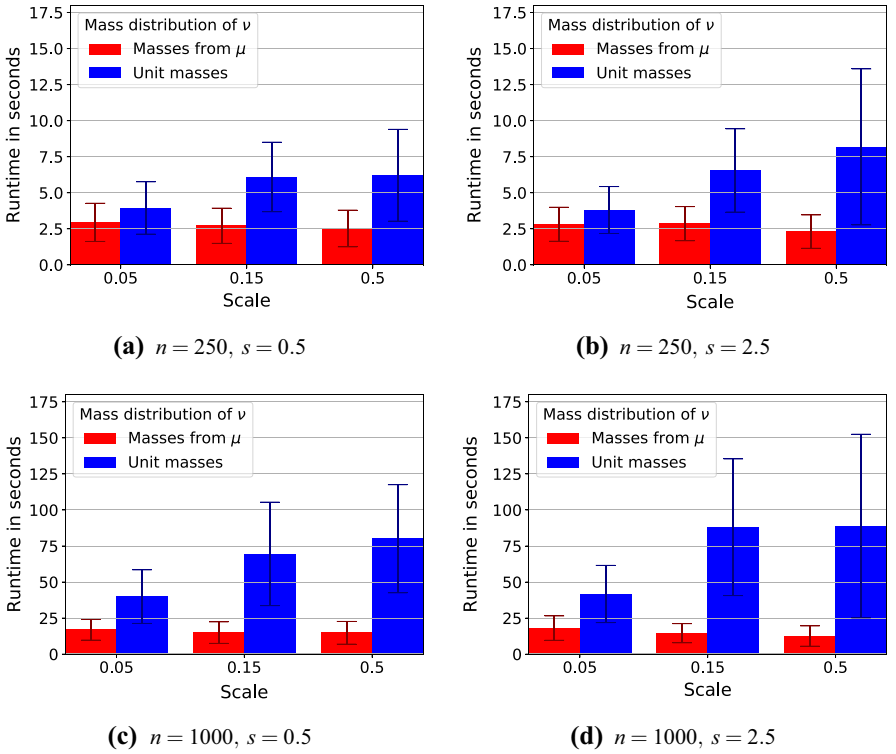
Fig. 3 Realizations of the measure μ for all six parameter combinations in Sect. 5. First row: smoothness $s = 0.5$; second row: smoothness $s = 2.5$. The correlation scale γ is 0.05, 0.15 and 0.5 (from left to right)

We implemented the algorithm in C++ and make it available on GitHub[2] under the MIT license. Our implementation uses libLBFGS (Okazaki and Nocedal 2010) for the L-BFGS procedure and the geometry library CGAL (CGAL 2015) for the construction and querying of weighted Voronoi tessellations. The repository also contains a Matlab script to visualize such tessellations. Our implementation is also included in the latest version of the `transport`-package (Schuhmacher et al. 2019) for the statistical computing environment R (R Core Team 2017).

5 Performance evaluation

We evaluated the performance of our algorithm by randomly generating measures μ and ν with varying features and computing the optimal transport partitions between them. The measure μ was generated by simulating its density ϱ as a Gaussian random field with Matérn covariance function on the rectangle $[0, 1] \times [0, 0.75]$, applying a quadratic function and normalizing the result to a probability density. Corresponding images were produced at resolution 256×196 pixels and were further divided into 25 subpixels each to compute integrals over Voronoi cells. In addition to a variance parameter, which we kept fixed, the Matérn covariance function has parameters for the scale γ of the correlations, which we varied among 0.05, 0.15 and 0.5, and the smoothness s of the generated surface, which we varied between 0.5 and 2.5 corresponding to a continuous surface and a C^2-surface, respectively. The simulation mechanism is similar to the ones for classes 2–5 in the benchmark DOTmark proposed in Schrieber et al. (2017), but allows to investigate the influence of individual parameters more directly. Figure 3 shows one realization for each parameter combination. For the performance evaluation we generated 10 realizations each.

[2] https://github.com/valentin-hartmann-research/semi-discrete-transport.



(a) $n = 250,\ s = 0.5$ **(b)** $n = 250,\ s = 2.5$

(c) $n = 1000,\ s = 0.5$ **(d)** $n = 1000,\ s = 2.5$

Fig. 4 Runtimes of the experiments of Sect. 5. Bars and lines indicate means and standard deviations over 200 experiments, combining 10 realizations of μ with 20 realizations of ν. The measures μ are based on Gaussian random fields with Matérn covariance function; see Fig. 3. The measures ν are based on support points picked uniformly at random with unit masses (blue) or masses picked from the corresponding μ (red). Rows: ν with $n = 250$ versus $n = 1000$ support points. Columns: smoothness parameter $s = 0.5$ versus $s = 2.5$. Note the different scaling. (Color figure online)

The measures ν have n support points generated uniformly at random on $[0, 1] \times [0, 0.75]$, where we used $n = 250$ and $n = 1000$. We then assigned either mass 1 or mass $\varrho(x)$ to each point x and normalized to obtain probability measures. We generated 20 independent ν-measures of the first kind (unit mass) and computed the optimal transport from each of the 10×6 μ-measures for each of the 6 parameter combinations. We further generated for each of the 10×6 μ-measures 20 corresponding ν-measures of the second kind (masses from μ) and computed again the corresponding optimal transports. The stopping criterion for the optimization was an amount of ≤ 0.05 of mistransported mass.

The results for $n = 250$ support points of ν are shown in Fig. 4a, b, those for $n = 1000$ support points in Fig. 4c, d. Each bar shows the mean of the runtimes on one core of a mobile Intel Core i7 across the 200 experiments for the respective parameter combination; the blue bars are for the ν measures with uniform masses, the red bars for the measures with masses selected from the corresponding μ measure. The lines indicate the standard deviations.

We observe that computation times stay more or less the same between parameter choices (with some sampling variation) if the ν-masses are taken from the corresponding μ-measure. In this case mass can typically be assigned (very) locally, and slightly more so if ρ has fewer local fluctuations (higher γ and/or s).

This seems a plausible explanation for the relatively small computation times.

In contrast, if all ν-masses are the same, the computation times are considerably higher and increase substantially with increasing γ and somewhat with increasing smoothness. This seems consistent with the hypothesis that the more the optimal transport problem can be solved by assigning mass locally the lower the computation times. For larger scales many of the support points of ν compete strongly for the assignment of mass and a solution can only be found globally. A lower smoothness may alleviate the problem somewhat, because it creates locally more variation in the available mass.

In addition to the runtimes, we also recorded how many update steps for the weight vector w were performed until convergence. We only investigate the update steps for the transport to the original measure ν, not the coarsenings ν^l, $l > 0$, because the former dominates the runtime, and also has a different dimensionality than the coarsenings. We have computed the Pearson and Spearman correlation coefficients between the numbers of update steps and the runtimes. Both for $n = 250$ and $n = 1000$ support points of ν, these correlation coefficients are larger than 0.99, indicating very high correlation. This strongly suggests that the differences in runtimes are not due to intricacies of the line search procedure or Voronoi cell computations, but rather due to differences in the structures of the simulated problem instances.

We would like to note that to the best of our knowledge the present implementation is the first one for computing the optimal transport in the semi-discrete setting for the case $p = 1$, which means that fair performance comparisons with other algorithms are not easily possible.

6 Applications

We investigate three concrete problem settings in order to better understand the workings and performance of our algorithm as well as to illustrate various theoretical and practical aspects pointed out in the paper.

6.1 Optimal transport between two normal distributions

We consider the two bivariate normal distributions $\mu = \mathrm{MVN}_2(a, \sigma^2 \mathrm{I}_2)$ and $\nu = \mathrm{MVN}_2(b, \sigma^2 \mathrm{I}_2)$, where $a = 0.8 \cdot \mathbb{1}$, $b = 2.2 \cdot \mathbb{1}$ 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 $\mathscr{X} = [0, 3]^2$. By discretization (quantization) a measure $\tilde{\nu}$ is obtained from ν. We then compute the optimal transport partition and the Wasserstein distances between μ and $\tilde{\nu}$ for both $p = 1$ and $p = 2$. Computations and plots for $p = 2$ are obtained with the package `transport` (Schuhmacher et al. 2019) for the statistical computing environment R

Table 1 Theoretical continuous and computed semi-discrete Wasserstein distances, together with the discretization error

MVN versus MVN			MVN + Leb versus 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

(R Core Team 2017). For $p = 1$ we use our implementation presented in the previous section.

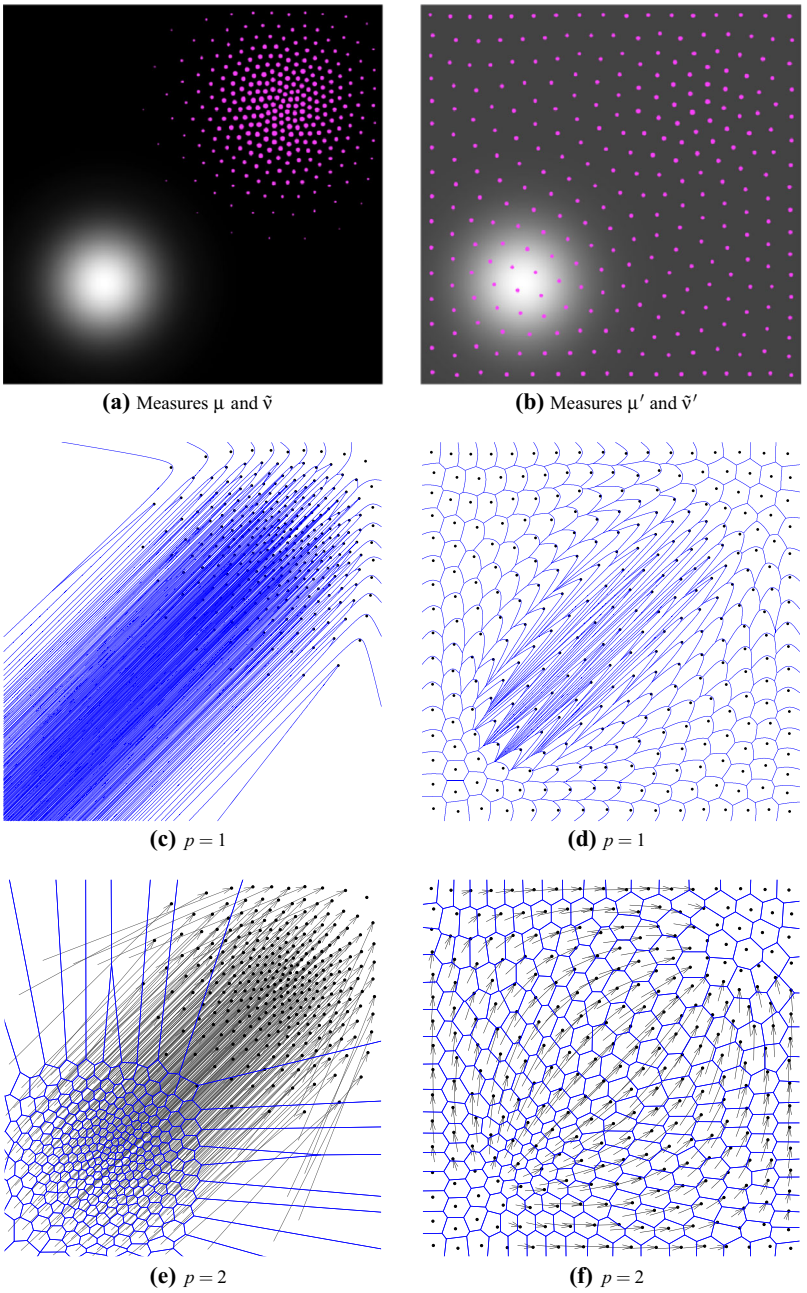
Note that for the original problem of optimal transport from μ to ν the solution is known exactly, so we can use this example to investigate the correct working of our implementation. In fact, for any probability measure μ' on $\mathbb{R}^d$ and its displacement $\nu' = T_{\#}\mu'$, where $T : \mathbb{R}^2 \to \mathbb{R}^2, x \mapsto x + (b - a)$ for some vector $b - a \in \mathbb{R}^d$, it is immediately clear that the translation T induces an optimal transport plan for (1) and that $W_p(\mu', \nu') = \|b - a\|$ for arbitrary $p \geq 1$. This holds because we obtain by Jensen's inequality $(\mathbb{E}\|X - Y\|^p)^{1/p} \geq \|\mathbb{E}(X - Y)\| = \|b - a\|$ for $X \sim \mu', Y \sim \nu'$; therefore $W_p(\mu', \nu') \geq \|b - a\|$ and T is clearly a transport map from μ' to ν' that achieves this lower bound. For $p = 2$ Theorem 9.4 in Villani (2009) yields that T is the unique optimal transport map and the induced plan π_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 μ to a fine grid and an initial configuration of cluster centers drawn independently from distribution ν and equipped with the corresponding density values of ν as weights. The number of cluster centers was kept to $n = 300$ for better visibility in the plots below. We write $\tilde{\nu} = \sum_{i=1}^{n} \delta_{y_i}$ 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 Fig. 5 depicts the measures μ and $\tilde{\nu}$ and the resulting optimal transport partitions for $p = 1$ and $p = 2$. In the case $p = 1$ the nuclei of the weighted Voronoi tessellation 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. Merigot (2011, Sect. 2), identifies the tessellation as a Laguerre tessellation (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 π that collects the mass for any $y \in \mathscr{Y}$ 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



(a) Measures μ and $\tilde{\nu}$ **(b)** Measures μ' and $\tilde{\nu}'$

(c) $p = 1$ **(d)** $p = 1$

(e) $p = 2$ **(f)** $p = 2$

Fig. 5 Left column: semi-discrete transport between a bivariate normal distribution μ 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 μ and μ' are displayed as gray level images, the point masses of the discrete measures ν and ν' are shown as small discs with areas proportional to the masses placed there. Panels **c** to **f** show the optimal transport partitions

$\|b-a\| = \sqrt{2} \cdot 1.4 \approx 1.979899$, and in particular they are well inside the boundaries set by the approximation error.

We also investigate the effect of adding a common measure to both μ and ν: let $\alpha = \mathrm{Leb}^d|_{\mathscr{X}}$ and proceed in the same way as above for the measures $\mu' = \mu + \alpha$ and $\nu' = \nu + \alpha$, calling the discretized measure $\tilde{\nu}'$. Note that the discretization error (sixth column of Table 1) is considerably higher, on the one hand due to the fact that the $n = 300$ support points of $\tilde{\nu}'$ 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 Fig. 5 depicts the measures μ' and $\tilde{\nu}'$ and the resulting optimal transport partitions for $p = 1$ and $p = 2$. Both partitions look very different from their counterparts when no α 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 now, meaning the approximated optimal transport plan is not just obtained by keeping measure α in place and moving the remaining measure μ according to the optimal transport plan π approximated in Fig. 5c, but a substantial amount of mass available from α is moved as well. Furthermore, Fig. 5d 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 $\mathscr{X}$ 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 behavior that goes with the fluid mechanics interpretation discussed in Sect. 1.2, see also Fig. 1. Various of the other points mentioned in Sect. 1.2 can be observed as well, e.g. the numerically computed Wasserstein distance is much smaller than for $p = 1$, which illustrates the lack of invariance and seems plausible in view of the example in Remark 2 in the appendix.

6.2 A practical resource allocation problem

We revisit the delivery problem mentioned in the introduction. A fast-food delivery service has 32 branches throughout a city area, depicted by the black dots on the map in Fig. 6. 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. Boscoe et al. (2012), or because the deliveries are performed by drones. The desired partition, computed by our implementation described in Sect. 4.3, is also displayed in Fig. 6. A number

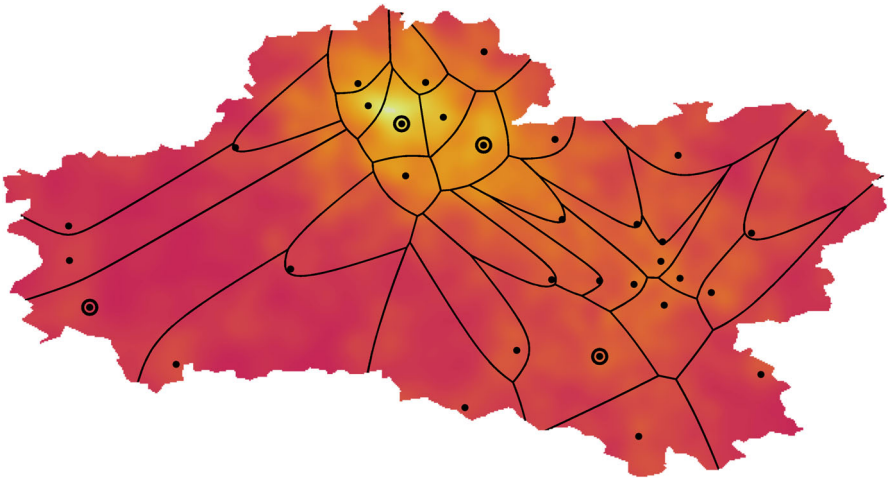


Fig. 6 The optimal partition of the city area for the delivery example

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.

6.3 A visual tool for detecting deviations from a density map

Very recently, asymptotic theory has been developed that allows, among other things, to test based on the Wasserstein metric W_p whether a sample in $\mathbb{R}^d$ comes from a given multivariate probability distribution Q. More precisely, assuming independent and identically distributed random vectors $X_1, \ldots, X_n$ with distribution P, limiting distributions have been derived for suitable standardizations of $W_p(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}, Q)$ both if $P = Q$ and if $P \neq Q$. Based on an observed value $W_p(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, Q)$, where $x_1, \ldots, x_n \in \mathbb{R}^d$, these distributions allow to assign a level of statistical certainty (p-value) to statements of $P = Q$ and $P \neq Q$, respectively. See Sommerfeld and Munk (2018), which uses general $p \geq 1$, but requires discrete distributions P and Q; and del Barrio and Loubes (2018), which is constraint to $p = 2$, but allows for quite general distributions (P and Q not both discrete).

We propose here the optimal transport partition between an absolutely continuous Q and $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ as a simple but useful tool for assessing the hypothesis $P = Q$. We refer to this tool as *goodness-of-fit (GOF) partition*. If $d = 2$, relevant information may be gained from a simple plot of this partition in a similar way as residual plots are used for assessing the fit of linear models. As a general rule of thumb the partition is consistent with the hypothesis $P = Q$ if it consists of many "round" cells that contain their respective P-points roughly in their middle. The size of cells may vary according to local densities and there are bound to be some elongated cells due to sampling error (i.e. the fact that we can only sample from P and do not know it exactly), but a local accumulation of many elongated cells should give rise to the suspicion that $P = Q$ may be violated in a specific way. Thus GOF partitions provide the data scientist both with a global impression for the plausibility of $P = Q$ and with detailed local

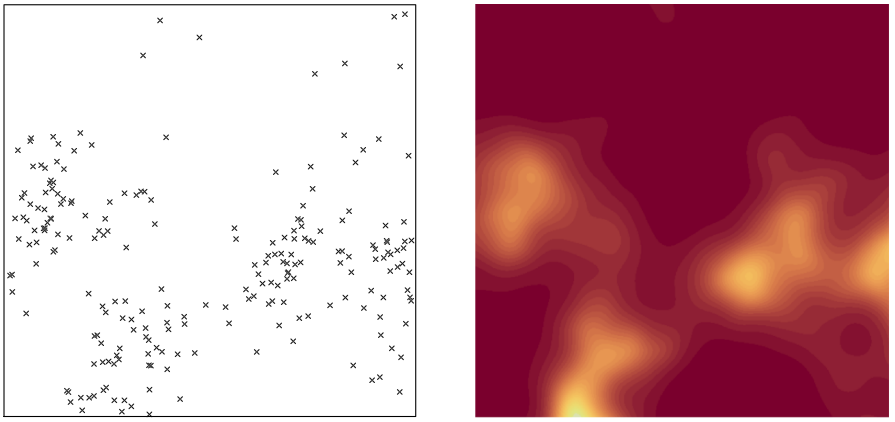


Fig. 7 A data example and a continuous density to compare it to

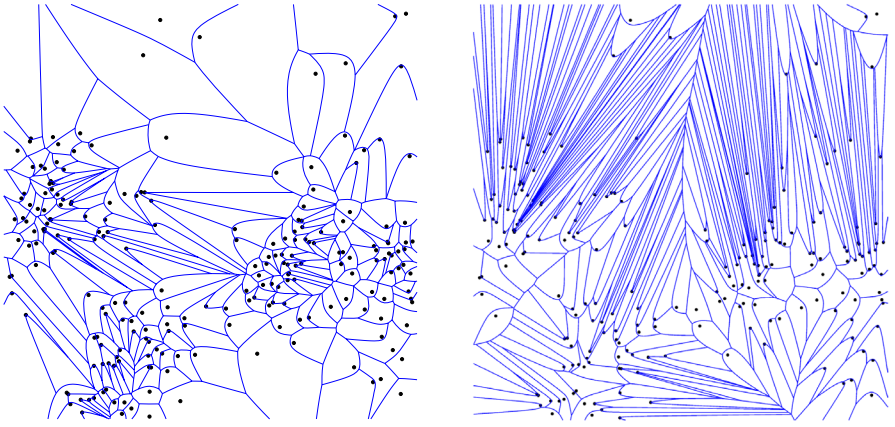


Fig. 8 Goodness-of-fit partitions for the data points in the left panel of Fig. 7 compared (on the left) with the density in the right panel of Fig. 7 and (on the right) with the uniform density on the square

information about the nature of potential deviations of P from Q. Of course they are a purely explorative tool and do not give any quantitative guarantees.

We give here an example for illustration. Suppose we have data as given in the left panel of Fig. 7 and a distribution Q as represented by the heat map in the right panel. Fig. 8 shows the optimal transport partition for this situation on the left hand side. The partition indicates that the global fit of the data is quite good. However it also points out some deviations that might be spurious, but might also well be worth further investigation: one is the absence of points close to the two highest peaks in the density map, another one that there are some points too many in the cluster on the very left of the plot. Both of them are quite clearly visible as accumulations of elongated cells.

As an example of a globally bad fit we show in the right panel of Fig. 8 the GOF partition when taking as Q the uniform measure on the square.

For larger d direct visual inspection becomes impossible. However, a substantial amount of information may still be extracted, either by considering statistics of the

GOF partition in d dimensions that are able to detect local regions of common orientation and high eccentricity of cells, or by applying dimension reduction methods, such as (Flamary et al. 2018), before applying the GOF partition.

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 Mérigot (2011) but adapted in various aspects to the current situation of $p = 1$.

Our algorithm converges towards the optimal partition subject to the convergence conditions for the L-BFGS algorithm; see e.g. Nocedal (1980). Very loosely, such conditions state that we start in a region around the minimizer where the objective function Φ shows to some extent quadratic behavior. Similar to the AHA-algorithm in Mérigot (2011), a proof of such conditions is not available. In practice, the algorithm has converged in all the experiments and examples given in the present paper.

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 $W_1(\nu^{(l)}, \nu^{(l-1)})$ substantially smaller without investing too much time in its computation.
- We currently keep the threshold ε in the stopping criterion of the multiscale approach in Sect. 4.1 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 ε_{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 ε_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 ε_{l-1} more quickly than a starting solution stemming from an ε_{l-1}-target at the coarser scale.
- The L-BFGS algorithm used for the optimization process may be custom-tailored to our discretization of μ 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 Φ any further since even in the discretized case the gradient always points in a descending direction.
- Approximating the intersections of μ-pixels with weighted Voronoi cells by splitting pixels into very small subpixels has shown good results. However, as

mentioned in Sect. 4.2, higher numerical stability and precision could be obtained by computing the intersections between the Voronoi cells and the pixels of μ exactly. Currently we are only able to do this at the expense of a large increase in the overall computation time. It is of considerable interest to have a more efficient method at hand.

– One of the reviewers pointed out that there are recent formulae available for the Hessian of the function Φ in Theorem 3. Indeed, based on Theorem 1 in De Gournay et al. (2019) we formally obtain in our setting a Hessian matrix with entries

$$\frac{\partial^2 \Phi}{\partial w_j \, \partial w_k}(w) = -\int_{\mathrm{Vor}_w(j)\cap \mathrm{Vor}_w(k)} \left\| \frac{x-y_j}{\|x-y_j\|} - \frac{x-y_k}{\|x-y_k\|} \right\|^{-1} \varrho(x)\, \sigma_{d-1}(dx) \tag{12}$$

for $j \neq k$, where σ_{d-1} denotes $(d-1)$-dimensional Hausdorff measure, and

$$\frac{\partial^2 \Phi}{\partial w_j^2}(w) = -\sum_{k \neq j} \frac{\partial^2 \Phi}{\partial w_j \, \partial w_k}(w).$$

Unfortunately, condition (Diff-2-a) required for this theorem is not satisfied for the unsquared Euclidean cost, since the norm term in (12) (without taking the inverse) goes to 0 as $x \to \infty$ along the boundary set $H = \{x \in \mathbb{R}^d \,|\, \|x-y_j\| - \|x-y_k\| = w_j - w_k\}$. We conjecture that the second derivative of Φ at w still exists and is of the above form if the integrals in (12) are finite (maybe under mild additional conditions).

If this can be established, we may in principle use a Newton method (with appropriate step size correction) for optimizing Φ. It remains to be seen, however, if the advantage from using the Hessian rather than performing a quasi Newton method outweighs the considerably higher computational cost due to computing the above integrals numerically. Another goal could be to establish global convergence of such a Newton algorithm under similar conditions as in Theorem 1.5 in Kitagawa et al. (2019), which is quite general, but requires higher regularity of the cost function.

– 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 Cooper (1972). Suppose that our fast-food chain from Sect. 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 μ) using our algorithm described in Sect. 4 and the relocation of branches to the spatial medians of the zones. The latter can be computed by discrete approximation, see e.g. Croux et al. (2012), and possibly by continuous techniques, see Fekete et al. (2005) for a vantage point.

Acknowledgements Open Access funding provided by Projekt DEAL.

Appendix: 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 μ, ν, α 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*

$$W_p(\alpha + \mu, \alpha + \nu) \leq W_p(\mu, \nu), \tag{13}$$

$$W_1(\alpha + \mu, \alpha + \nu) = W_1(\mu, \nu), \tag{14}$$

$$W_p(c\mu, c\nu) = c^{1/p} W_p(\mu, \nu), \tag{15}$$

where we assume for (14) that $W_1(\mu, \nu) < \infty$.

Proof Write $\Delta = \{(x, x) | x \in \mathbb{R}^d\}$. Denote by α_Δ the push-forward of α under the map $[\mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d, x \mapsto (x, x)]$. Let π_* 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 π_*. 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 π_* for the integral in (1). Thus

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

Replacing c by $1/c$, as well as μ by $c\mu$ and ν by $c\nu$, we obtain (15).

It remains to show $W_1(\alpha + \mu, \alpha + \nu) \geq W_1(\mu, \nu)$. For this we use that a transport plan π between μ and ν is optimal if and only if it is cyclical monotone, meaning that for all $N \in \mathbb{N}$ and all $(x_1, y_1), \ldots, (x_N, y_N) \in \operatorname{supp}(\pi)$, 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 Villani (2009, Theorem 5.10(ii) and Definition 5.1).

Letting π_* 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 π_* is cyclical monotone. Let $N \in \mathbb{N}$ and $(x_1, y_1), \ldots, (x_N, y_N) \in \operatorname{supp}(\pi_* + \alpha_\Delta) \subset \operatorname{supp}(\pi_*) \cup \Delta$. 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 $(x_{i_j}, y_{i_j}) \in \operatorname{supp}(\pi_*)$. By the cyclical monotonicity of π_* (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 2 Equation (14) is not generally true for any $p > 1$. To see this consider the case $d = 1$, $\mu = \delta_0$, $\nu = \delta_1$ and $\alpha = b \operatorname{Leb}|_{[0,1]}$, 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

$$\begin{cases} F(x) = G(x) = 0 & \text{if } x < 0, \\ F(x) = 1 + bx, \ G(x) = bx & \text{if } x \in [0, 1), \\ F(x) = G(x) = b + 1 & \text{if } x \geq 1. \end{cases}$$

We then even obtain

$$\begin{aligned} W_p^p(\alpha + \mu, \alpha + \nu) &= \int_0^{b+1} |F^{-1}(t) - G^{-1}(t)|^p \, dt \\ &= 2 \int_0^1 \frac{t^p}{b^p} \, dt + \frac{1}{b^p}(b - 1) = \frac{1}{b^p}\Big(b - 1 + \frac{2}{p+1}\Big) \longrightarrow 0 \end{aligned}$$

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 Eq. (2) in Rippl et al. (2016) and the references given there and note that the generalization from the result for probability measures is immediate by (15).

References

Altschuler J, Weed J, Rigollet P (2017) Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In: Proceedings of NIPS 2017, pp 1961–1971

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

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: Proceedings of the 34th international conference on machine learning, PMLR, vol. 70. Sydney, Australia (2017)

Armijo L (1966) Minimization of functions having Lipschitz continuous first partial derivatives. Pac J Math 16(1):1–3

Aurenhammer F, Hoffmann F, Aronov B (1998) Minkowski-type theorems and least-squares clustering. Algorithmica 20(1):61–76
Basua S, Kolouria S, Rohde GK (2014) Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. PNAS 111(9):3448–3453
Beckmann M (1952) A continuous model of transportation. Econometrica 20:643–660
Benamou JD, Brenier Y (2000) A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numer Math 84:375–393
Boscoe FP, Henry KA, Zdeb MS (2012) A nationwide comparison of driving distance versus straight-line distance to hospitals. Prof Geogr 64(2):188–196
Bourne DP, Schmitzer B, Wirth B (2018) Semi-discrete unbalanced optimal transport and quantization. Preprint. arXiv:1808.01962
CGAL (2015) Computational geometry algorithms library (version 4.6.1). http://www.cgal.org
Cooper L (1972) The transportation-location problem. Oper Res 20(1):94–108
Courty N, Flamary R, Tuia D, Corpetti T (2016) Optimal transport for data fusion in remote sensing. In: 2016 IEEE international geoscience and remote sensing symposium (IGARSS), pp 3571–3574
Crippa G, Jimenez C, Pratelli A (2009) Optimum and equilibrium in a transport problem with queue penalization effect. Adv Calc Var 2(3):207–246
Croux C, Filzmoser P, Fritz H (2012) A comparison of algorithms for the multivariate L_1-median. Comput Stat 27(3):393–410
Cuturi M (2013) Sinkhorn distances: lightspeed computation of optimal transport. Proc NIPS 2013:2292–2300
De Gournay F, Kahn J, Lebrat L (2019) Differentiation and regularity of semi-discrete optimal transport with respect to the parameters of the discrete measure. Numer Math 141(2):429–453
del Barrio E, Loubes JM (2018) Central limit theorems for empirical transportation cost in general dimension. Ann Probab 47(2):926–951
Fekete SP, Mitchell JSB, Beurer K (2005) On the continuous Fermat–Weber problem. Oper Res 53(1):61–76
Flamary R, Cuturi M, Courty N, Rakotomamonjy A (2018) Wasserstein discriminant analysis. Mach Learn 107(12):1923–1945
Geiß D, Klein R, Penninger R, Rote G (2013) Optimally solving a transportation problem using Voronoi diagrams. Comput Geom 46(8):1009–1016
Genevay A, Cuturi M, Peyré G, Bach F (2016) Stochastic optimization for large-scale optimal transport. Proc NIPS 2016:3432–3440
Genevay A, Peyré G, Cuturi M (2018) Learning generative models with Sinkhorn divergences. In: Proceedings of the 21st international conference on artificial intelligence and statistics, PMLR, vol 84. Lanzarote, Spain
Gramfort A, Peyré G, Cuturi M (2015) Fast optimal transport averaging of neuroimaging data. In: 24th International conference on information processing in medical imaging (IPMI 2015), lecture notes in computer science, vol 9123, pp 123–160
Grippo L, Lampariello F, Lucidi S (1986) A nonmonotone line search technique for Newton's method. SIAM J Numer Anal 23(4):707–716
Guo J, Pan Z, Lei B, Ding C (2017) Automatic color correction for multisource remote sensing images with Wasserstein CNN. Rem Sens 9(5):1–16 (electronic)
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. arXiv:1706.07403
Kantorovich L (1942) On the translocation of masses. C R (Doklady) Acad Sci URSS (NS) **37**, 199–201
Karavelas MI, Yvinec M (2002) Dynamic additively weighted Voronoi diagrams in 2D. In: Algorithms—ESA 2002. Springer, Berlin, pp 586–598
Kitagawa J, Mérigot Q, Thibert B (2019) Convergence of a Newton algorithm for semi-discrete optimal transport. J Eur Math Soc 21:2603–2651
Klatt M, Tameling C, Munk A (2019) Empirical regularized optimal transport: statistical theory and applications. Preprint. arXiv:1810.09880
Luenberger DG, Ye Y (2008) Linear and nonlinear programming, third edn. International series in operations research and management science, 116. Springer, New York
Mallozzi L, Puerto J, Rodríguez-Madrena M (2019) On location-allocation problems for dimensional facilities. J Optim Theory Appl 182(2):730–767
McCann RJ (1995) Existence and uniqueness of monotone measure-preserving maps. Duke Math J 80(2):309–323

Mérigot Q (2011) A multiscale approach to optimal transport. Comput Graph. Forum 30(5):1583–1592
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, pp 666–704
Nicolas P (2016) Optimal transport for image processing. Habilitation thesis, Signal and Image Processing, Université de Bordeaux. https://hal.archives-ouvertes.fr/tel-01246096v6
Nocedal J (1980) Updating quasi-Newton matrices with limited storage. Math Comput 35(151):773–782
Nocedal J, Wright S (1999) Numerical optimization. Springer Sci 35(67–68):7
Núñez M, Scarsini M (2016) Competing over a finite number of locations. Econ Theory Bull 4(2):125–136
Okazaki N, Nocedal J (2010) libLBFGS (Version 1.10). http://www.chokkan.org/software/liblbfgs/
Peyré G, Cuturi M (2018) Computational optimal transport. now Publishers. arXiv:1803.00567
Pratelli A (2007) On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. Ann Inst H Poincaré Probab Stat 43(1):1–13
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/
Rippl T, Munk A, Sturm A (2016) Limit laws of the empirical Wasserstein distance: Gaussian distributions. J Multivar Anal 151:90–109
Santambrogio F (2015) Optimal transport for applied mathematicians, Progress in nonlinear differential equations and their applications, vol 87. Birkhäuser/Springer, Cham
Schmitz MA, Heitz M, Bonneel N, Ngolè F, Coeurjolly D, Cuturi M, Peyré G, Starck JL (2018) Wasserstein dictionary learning: optimal transport-based unsupervised nonlinear dictionary learning. SIAM J Imaging Sci 11(1):643–678
Schmitzer B (2016) A sparse multiscale algorithm for dense optimal transport. J Math Imaging Vis 56(2):238–259
Schmitzer B (2019) Stabilized sparse scaling algorithms for entropy regularized transport problems. SIAM J Sci Comput 41(3):A1443–A1481
Schmitzer B, Wirth B (2019) A framework for Wasserstein-1-type metrics. J Convex Anal 26(2):353–396
Schrieber J, Schuhmacher D, Gottschlich C (2017) DOTmark: a benchmark for discrete optimal transport. IEEE Access, 5
Schuhmacher D, Bähre B, Gottschlich C, Hartmann V, Heinemann F, Schmitzer B, Schrieber J (2019) Transport: computation of optimal transport plans and Wasserstein distances. R package version 0.11-1. https://cran.r-project.org/package=transport
Sherali HD, Nordai FL (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
Solomon J, de Goes F, Peyré G, Cuturi M, Butscher A, Nguyen A, Du T, Guibas L (2015) Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. ACM Trans Graph 34(4): 66:1–66:11
Solomon J, Rustamov R, Guibas L, Butscher A (2014) Earth mover's distances on discrete surfaces. ACM Trans Graph 33(4): 67:1–67:12
Sommerfeld M, Munk A (2018) Inference for empirical Wasserstein distances on finite spaces. J R Stat Soc: Ser B (Statistical Methodology) 80(1):219–238
Villani C (2009) Optimal transport, old and new, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol 338. Springer, Berlin
Wolansky G (2015) Semi-discrete approximation of optimal mass transport. Preprint. arXiv:1502.04309v1
Wolfe P (1969) Convergence conditions for ascent methods. SIAM Rev 11:226–235
Wolfe P (1971) Convergence conditions for ascent methods. II. Some corrections. SIAM Rev 13:185–188