Wasserstein Distance on Finite Spaces: Statistical Inference and Algorithms

Dissertation zur Erlangung
des mathematisch-naturwissenschaftlichen Doktorgrades
„Doctor rerum naturalium“
der Georg-August-Universität Göttingen
im Promotionsstudiengang „Mathematical Sciences“
der Georg-August University School of Science (GAUSS)

vorgelegt von
Max Sommerfeld
aus Hannover

Göttingen, September 2017

Betreuungsausschuss
Prof. Dr. Axel Munk, Institut für Mathematische Stochastik
Prof. Dr. Stephan Huckemann, Institut für Mathematische Stochastik

Mitglieder der Prüfungskommission
Referent:
Prof. Dr. Axel Munk, Institut für Mathematische Stochastik
Koreferent:
Prof. Dr. Stephan Huckemann, Institut für Mathematische Stochastik

Weitere Mitglieder der Prüfungskommission
Prof. Dr. Max Wardetzky, Institut für Numerische u. Angewandte Mathematik
Prof. Dr. Anja Sturm, Institut für Mathematische Stochastik
Prof. Dr. Ingo Witt, Institut für Reine Mathematik
Dr. Frank Werner, Max-Planck-Institut für biophysikalische Chemie

Tag der mündlichen Prüfung: 18.10.2017

Preface

Wasserstein distances or, more generally, distances that quantify the *optimal transport* between probability measures on metric spaces have long been established as an important tool in probability theory. More recently, it has found its way into statistical theory, applications and machine learning - not only as a theoretical tool but also as a quantity of interest in its own right. Examples include goodness-of-fit, two-sample and equivalence testing, classification and clustering, exploratory data analysis using Fréchet means and geodesics in the Wasserstein metric.

This advent of the Wasserstein distance as a statistical tool manifests two major challenges. First, knowledge on the theoretical properties of empirical, i.e. sample-based, Wasserstein distances remains incomplete, in particular as far as distributional limits on spaces other than the real line are concerned. Second, any application of the Wasserstein distance invokes massive computational challenges, leaving many practically interesting problems outside of the scope of available algorithms.

The main thesis of this work is that restricting ourselves to the *Wasserstein distance on finite spaces* offers a perspective that is able to solve or at least avoid these problems and is still general enough to include many practical problems. Indeed, this work will present comprehensive distributional limits for empirical Wasserstein distances on finite spaces, strategies to apply these limits with controllable computational burden in large-scale inference and a fast probabilistic approximation scheme for optimal transport distances.

Previous publications and joint work Large parts of this work have previously been published in Sommerfeld and Munk (2017). In fact, all of Chapter 2, except for the sections on normal limits under the alternative and the limiting distribution as a Wasserstein distance as well as the introductory part concerning distributional limits in Chapter 1 are taken from Sommerfeld and Munk (2017) with only few modifications.

The ideas and results of Chapter 3 have been published in the preprint Tameling et al. (2017).

The application to single-marker switching microscopy in Section 3.3 is joint work with Carla Tameling. The author of this dissertation and Carla Tameling contributed equally to design, implementation and evaluation of the application.

The numerical experiments on the performance of the probabilistic approximation scheme in Section 4.3 are joint work with Jörn Schrieber. The author of this dissertation and Jörn Schrieber contributed equally to design, implementation and evaluation of the experiments.

Contents

Chapter 1

Introduction

1.1 Distributional limits

The *Wasserstein distance* (Vasershtein, 1969), also known as Mallows distance (Mallows, 1972), Monge-Kantorovich-Rubinstein distance in the physical sciences (Kantorovich and Rubinstein, 1958; Rachev, 1985; Jordan et al., 1998), earth-mover's distance in computer science (Rubner et al., 2000) or optimal transport distance in optimization (Ambrosio, 2003), is one of the most fundamental metrics on the space of probability measures. Besides its prominence in probability (e.g. Dobrushin (1970); Gray (1988)) and finance (e.g. Rachev and Rüschendorf (1998)) it has deep connections to the asymptotic theory of PDEs of diffusion type (Otto (2001), Villani (2003, 2008) and references therein). In a statistical setting it has mainly been used as a tool to prove weak convergence in the context of limit laws (e.g. Bickel and Freedman (1981); Shorack and Wellner (1986); Johnson and Samworth (2005); Dümbgen et al. (2011); Dorea and Ferreira (2012)) as it metrizes weak convergence together with convergence of moments. However, recently the empirical (i.e. estimated from data) Wasserstein distance has also been recognized as a central quantity itself in many applications, among them clinical trials (Munk and Czado, 1998; Freitag et al., 2007), metagenomics (Evans and Matsen, 2012), medical imaging (Ruttenberg et al., 2013), goodness-of-fit testing (Freitag and Munk, 2005; Del Barrio et al., 1999), biomedical

engineering (Oudre et al., 2012), computer vision (Gangbo and McCann, 2000; Ni et al., 2009), cell biology (Orlova et al., 2016) and model validation (Halder and Bhattacharya, 2011). The barycenter with respect to the Wasserstein metric (Agueh and Carlier, 2011) has been shown to elicit important structure from complex data and to be a promising tool, for example in deformable models (Boissard et al., 2015; Agulló-Antolín et al., 2015). It has also been used in large-scale Bayesian inference to combine posterior distributions from subsets of the data (Srivastava et al., 2015).

Generally speaking three characteristics of the Wasserstein distance make it particularly attractive for various applications. First, it incorporates a ground distance on the space in question. This often makes it more adequate than competing metrics such as total-variation or χ^2-metrics which are oblivious to any metric or similarity structure on the ground space. As an example, the success of the Wasserstein distance in metagenomics applications can largely be attributed to this fact (see Evans and Matsen (2012) and also our application in Section 2.9.3).

Second, it has a clear and intuitive interpretation as the amount of 'work' required to transform one probability distribution into another and the resulting transport can be visualized (see Section 2.9.2). This is also interesting in applications where probability distributions are used to represent actual physical mass and spatio-temporal changes have to be tracked.

Third, it is well-established (Rubner et al., 2000) that the Wasserstein distance performs exceptionally well at capturing human perception of similarity. This motivates its popularity in computer vision and related fields.

Despite these advantages, the use of the empirical Wasserstein distance in a statistically rigorous way is severely hampered by a lack of inferential tools. We argue that this issue stems from considering too large classes of candidate distributions (e.g. those which are absolutely continuous with respect to the Lebesgue measure if the ground space has dimension ≥ 2). In this paper, we therefore discuss the Wasserstein distance on finite spaces, which allows to solve this issue. We argue that the restriction to finite spaces is not merely an approximation to the truth, but rather that this setting is sufficient for many practical situations as measures often already come naturally discretized (e.g.

two- or three-dimensional images - see also our applications in Section 2.9).

We remark that from our methodology further inferential procedures can be derived, e.g. a (M)ANOVA type of analysis and multiple comparisons of Wasserstein distances based on their p-values (see e.g. Benjamini and Hochberg (1995)). Our techniques also extend immediately to dependent samples (X_i, Y_i) with marginals $\boldsymbol{r}$ and $\boldsymbol{s}$.

Wasserstein distance Let $(\mathcal{X}, d)$ be a complete metric space with metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. The *Wasserstein distance of order* p ($p \geq 1$) between two Borel probability measures μ_1 and μ_2 on $\mathcal{X}$ is defined as

$$W_p(\mu_1, \mu_2) = \left\{ \inf_{\nu \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{X} \times \mathcal{X}} d^p(x, x') \nu(dx, dx') \right\}^{1/p},$$

where $\Pi(\mu_1, \mu_2)$ is the set of all Borel probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ_1 and μ_2, respectively.

Wasserstein distance on finite spaces If we restrict in the above definition $\mathcal{X} = \{x_1, \ldots, x_N\}$ to be a finite space, every probability measure on $\mathcal{X}$ is given by a vector $\boldsymbol{r}$ in $\mathcal{P}_{\mathcal{X}} = \left\{ \boldsymbol{r} = (r_x)_{x \in \mathcal{X}} \in \mathbb{R}^{\mathcal{X}}_{>0} : \sum_{x \in \mathcal{X}} r_x = 1 \right\}$, via $P_{\boldsymbol{r}}(\{x\}) = r_x$. We will not distinguish between the vector $\boldsymbol{r}$ and the measure it defines. The *Wasserstein distance of order* p between two finitely supported probability measures $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_{\mathcal{X}}$ then becomes

$$W_p(\boldsymbol{r}, \boldsymbol{s}) = \left\{ \min_{\boldsymbol{w} \in \Pi(\boldsymbol{r}, \boldsymbol{s})} \sum_{x, x' \in \mathcal{X}} d^p(x, x') w_{x,x'} \right\}^{1/p}, \tag{1.1}$$

where $\Pi(\boldsymbol{r}, \boldsymbol{s})$ is the set of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginal distributions $\boldsymbol{r}$ and $\boldsymbol{s}$, respectively. All our methods and results concern this Wasserstein distance on finite spaces.

1.1.1 Overview of main results

Distributional limits The basis for inferential procedures for the Wasserstein distance on finite spaces is a limit theorem for its empirical version

$W_p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)$. Here, the empirical measure generated by independent random variables $X_1, \ldots, X_n \sim \boldsymbol{r}$ is given by $\hat{\boldsymbol{r}}_n = (\hat{r}_{n,x})_{x \in \mathcal{X}}$, where $\hat{r}_{n,x} = \frac{1}{n} \# \{k : X_k = x\}$. Let $\hat{\boldsymbol{s}}_m$ be generated from i.i.d. $Y_1, \ldots, Y_m \sim \boldsymbol{s}$ in the same fashion. Under the null hypothesis $\boldsymbol{r} = \boldsymbol{s}$ we prove that

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2p}} W_p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) \Rightarrow \left\{\max_{\boldsymbol{u} \in \Phi_p^*} \langle \boldsymbol{G}, \boldsymbol{u} \rangle\right\}^{\frac{1}{p}}, \quad n, m \to \infty. \tag{1.2}$$

Here, '$\Rightarrow$' means convergence in distribution, $\boldsymbol{G}$ is a mean zero Gaussian random vector with covariance depending on $\boldsymbol{r} = \boldsymbol{s}$ and Φ_p^* is the convex set of dual solutions to the Wasserstein problem depending on the metric d only (see Theorem 1). In Section 2.9.2 we use this result to assess the statistical significance of the differences between real and synthetically generated fingerprints in the Fingerprint Verification Competition (Maio et al., 2002).

We give analogous results under the alternative $\boldsymbol{r} \neq \boldsymbol{s}$. This extends the scope of our results beyond the classical two-sample (or goodness-of-fit test) as it allows for confidence statements on $W_p(\boldsymbol{r}, \boldsymbol{s})$ when the null hypothesis of equality is likely or even *known to be false*. An example for this is given by our application to metagenomics (Section 2.9.3) where samples from the same person taken at different times are typically statistically different but our asymptotic results allow us to assert with statistical significance that inter-personal distances are larger that intra-personal ones.

Proof strategy We prove these results by showing that the Wasserstein distance is *directionally Hadamard differentiable* (Shapiro, 1990) and the right hand side of (1.2) is its derivative evaluated at the Gaussian limit of the empirical multinomial process (see Theorem 4). This notion generalizes Hadamard differentiability by allowing *non-linear* derivatives but still allows for a refined delta-method (Römisch (2004) and Theorem 3). Notably, the Wasserstein distance is not Hadamard differentiable in the usual sense.

Explicit limiting distribution for tree metrics When the space $\mathcal{X}$ are the vertices of a tree and the metric d is given by path length we give an explicit expression for the limiting distribution in (1.2) (see Theorem

5). In contrast to the general case, this explicit formula allows for fast and direct simulation of the limiting distribution. This extends a previous result of Samworth and Johnson (2004) who considered a finite number of point masses on the real line. The Wasserstein distance on trees has, to the best of our knowledge, only been considered in two papers: Kloeckner (2013) studies the geometric properties of the Wasserstein space of measures on a tree and Evans and Matsen (2012) use the Wasserstein distance on phylogenetic trees to compare microbial communities.

The bootstrap Directional Hadamard differentiability is not enough to guarantee the consistency of the naive (n out of n) bootstrap (Dümbgen, 1993; Fang and Santos, 2014) - in contrast to the usual notion of Hadamard differentiability. This implies that the bootstrap is *not* consistent for the Wasserstein distance (1.1)(see Theorem 9). In contrast, the m-out-of-n bootstrap for $m/n \to 0$ is known to be consistent in this setting (Dümbgen, 1993) and can be applied to the Wasserstein distance. Under the null hypothesis $\boldsymbol{r} = \boldsymbol{s}$, however, there is a more direct way of obtaining an approximation of the limiting distribution. In the appendix, we discuss this alternative resampling scheme based on ideas of Fang and Santos (2014), that essentially consists of plugging in a bootstrap version of the underlying empirical process in the derivative. We show that this scheme, which we will call *directional bootstrap*, is consistent for the Wasserstein distance (see Theorem 9, Section 2.7).

1.1.2 Related work

Empirical Wasserstein distances In very general terms, we study a particular case (finite spaces) of the following question and its two-sample analog: Given the empirical measure μ_n based on n i.i.d. random variables taking variables in a metric space with law μ. What can be inferred about $W_p(\mu_n, \mu_0)$ for a reference measure μ_0 which may be equal to μ?

It is a well-known and straightforward consequence of the strong law of large numbers that if the p-th moments are finite for μ and μ_0 then $W_p(\mu_n, \mu_0)$

converges to $W_p(\mu, \mu_0)$, almost surely, as the sample size n approaches infinity (Villani, 2008, Cor. 6.11). Determining the exact rate of this convergence is the subject of an impressive body of literature developed over the last decades starting with the seminal work of Ajtai et al. (1984) considering for μ_0 the uniform distribution on the unit square, followed by Talagrand (1992, 1994) for the uniform distribution in higher dimensions and Horowitz and Karandikar (1994) giving bounds on mean rates of convergence. Boissard and Gouic (2014); Fournier and Guillin (2014) gave general deviation inequalities for the empirical Wasserstein distance on metric spaces. For a discussion in the light of our distributional limit results see Section 2.10.

Distributional limits give a natural perspective for practicable inference, but despite considerable interest in the topic have remained elusive to a large extent. For measures on $\mathcal{X} = \mathbb{R}$ a rather complete theory is available (see Munk and Czado (1998); Freitag et al. (2007); Freitag and Munk (2005) for $\mu_0 \neq \mu$ and e.g. Del Barrio et al. (1999); Samworth and Johnson (2005); Del Barrio et al. (2005) for $\mu_0 = \mu$ as well as Mason (2016); Bobkov and Ledoux (2014) for recent surveys). However, for $\mathcal{X} = \mathbb{R}^d$, $d \geq 2$ there are only two distributional results known to us. The first is due to Rippl et al. (2015) for specific multivariate (elliptic) parametric classes of distributions, when the empirical measure is replaced by a parametric estimate. The second is the very recent work of Del Barrio and Loubes (2017), which considers the case of different underlying measures on $\mathbb{R}^d$ (in the case of equal measures the limiting distribution becomes degenerate) with positive Lebesgue density on their convex support. They prove their result using a Stein identity. In the context of deformable models distributional results are proven (Del Barrio et al., 2015) for specific multidimensional parametric models which factor into one-dimensional parts.

The simple reason why the Wasserstein distance is so much easier to handle in the one-dimensional case is that in this case the optimal coupling attaining the infimum in (1.1) is known explicitly. In fact, the Wasserstein distance of order p between two measures on $\mathbb{R}$ then becomes the L^p norm of the difference of their quantile functions (see Mallows (1972) for an early reference) and the analysis of empirical Wasserstein distances can be based

on quantile process theory. Beyond this case, explicit coupling results are only known for multivariate Gaussians and elliptic distributions (Gelbrich, 1990). A classical result of Ajtai et al. (1984) for the uniform distribution on $\mathcal{X} = [0,1]^2$ suggests that, even in this simple case, distributional limits will have a complicated form if they exist at all. We will elaborate on this thought in the discussion, in Section 2.10.

The Wasserstein distance on finite spaces has been considered recently by Gozlan et al. (2013) to derive entropy inequalities on graphs and by Erbar and Maas (2012) to define Ricci curvature for Markov chains on discrete spaces. To the best of our knowledge, empirical Wasserstein distances on finite spaces have only been considered by Samworth and Johnson (2004) in the special case of measures supported on $\mathbb{R}$. We will show (Section 2.4) that our results extend theirs.

Directional Hadamard differentiability We prove our distributional limit theorems using the theory of parametric programming (Bonnans and Shapiro, 2013) which investigates how the optimal value and the optimal solutions of an optimization problem change when the objective function and the constraints are changed. While differentiability properties of optimal values of linear programs are extremely well studied such results have, to the best of our knowledge, not yet been applied to the statistical analysis of Wasserstein distances.

It is well-known that under certain conditions the optimal value of a mathematical program is differentiable with respect to the constraints of the problem (Rockafellar, 1984; Gal et al., 1997). However, the derivative will typically be non-linear. The appropriate concept for this is directional Hadamard differentiability (Shapiro, 1990). The derivative of the optimal value of a mathematical program is typically again given as an extremal value.

Although the delta-method for directional Hadamard derivatives has been known for a long time (Shapiro, 1991; Dümbgen, 1993), this notion scarcely appears in the statistical context (with some exceptions, such as Römisch (2004), see also Donoho and Liu (1988)). Recently, an interest in the topic has

evolved in econometrics (see Fang and Santos (2014) and references therein).

1.2 Strategies for inference in large-scale problems

When the size N of the underlying space $\mathcal{X}$ becomes large, both the Wasserstein distance itself and the limiting distributions described above pose serious computational challenges. Frequently, the application of the distributional results to a practical problem will become computationally infeasible. In Chapter 4 we propose an algorithm to efficiently approximate the Wasserstein distance. However, this approach is often inappropriate when rigorous statistical inference is the goal as it does not provide useful statistical guarantees for the approximation error.

As an alternative approach we propose to combine a lower bound for the Wasserstein distance (based on thresholding the ground distance (Pele and Werman, 2009)) with a stochastic upper bound for the limiting distribution (based on the explicit expression for the limiting distribution for trees, Section 2.4) to obtain a conservative but fast to compute two-sample test. The lower bound can typically be computed in super-quadratic (in N) runtime, compared to super-cubic runtimes for the exact Wasserstein distance. One realization of the stochastic upper bound only even requires linear time, while a sample from the exact limiting distribution would essentially require the same computational effort as the Wasserstein distance itself.

We apply this method to validate drift correction in stochastic sub-diffraction microscopy.

1.3 Fast probabilistic approximation

The outstanding theoretical and practical performance of optimal transport distances is contrasted by its excessive computational cost. For example, optimal transport distances can be computed with an auction algorithm (Bertsekas, 1992). For two probability measures supported on N points this

algorithm has a worst case run time of $\mathcal{O}(N^3 \log N)$. Other methods like the transportation simplex have sub-cubic empirical average runtime (compare Gottschlich and Schuhmacher (2014)), but exponential worst case runtimes.

Many attempts have therefore been made to improve upon these run times. Ling and Okada (2007) proposed a specialized algorithm for L_1-ground distance and $\mathcal{X}$ a regular grid and report an empirical runtime of $\mathcal{O}(N^2)$. Gottschlich and Schuhmacher (2014) improved existing general purpose algorithms by initializing with a greedy heuristic. Their *Shortlist* algorithm achieves an empirical average runtime of the order $\mathcal{O}(N^{5/2})$. Schmitzer (2016) solves the optimal transport problem by solving a sequence of sparse problems.

Despite these efforts, many practically relevant problems remain well outside the scope of available algorithms (see Schrieber et al. (2016) for a comparison of state-of-the-art algorithms). This is true in particular for two or three dimensional images and spatio temporal imaging, which constitute an important area of potential applications. Here, N is the number of pixels or voxels and is typically very large. Naturally, this problem is aggravated when many distances have to be computed as is the case for Wasserstein barycenters (Agueh and Carlier, 2011; Cuturi and Doucet, 2014), which have become an important use case.

To bypass the computational bottleneck, many surrogates for optimal transport distances that are more amenable to fast computation have been proposed. Shirdhonkar and Jacobs (2008) proposed to use an equivalent distance based on wavelets that can be computed in linear time but cannot be calibrated to approximate the Wasserstein distance with arbitrary accuracy. Pele and Werman (2009) threshold the ground distance to reduce the complexity of the underlying linear program, obtaining a lower bound for the exact distance. Cuturi (2013) altered the optimization problem by adding an entropic penalty term in order to use faster and more stable algorithms. Bonneel et al. (2015) consider the 1-D Wasserstein distances of radial projections of the original measures, exploiting the fact that, in one dimension, computing the Wasserstein distance amounts to sorting the point masses and hence has quasi-linear computation time.

1.3.1 Contribution

We do *not* propose a new algorithm to solve the optimal transport problem. Instead, we propose a probabilistic scheme as a meta-algorithm that can use any algorithm (e.g. those mentioned above) as a black-box back-end and gives a random but fast approximation of the exact distance. This scheme

a) is extremely easy to implement and to tune towards higher accuracy or shorter computation time as desired;

b) can be used with any algorithm for transportation problems as a back-end, including general LP solvers, specialized network solvers and algorithms using entropic penalization (Cuturi, 2013);

c) comes with theoretical non-asymptotic guarantees for the approximation error - in particular, this error is independent of the size of the original problem in many important cases, including images;

d) works well in practice. For example, the Wasserstein distance between two 128^2-pixel images can typically be approximated with a relative error of less than 5% in only 1% of the time required for exact computation.

1.4 Organization of the work

This work is organized in three Chapters containing the results on distributional limits, strategies for inference in large-scale problems and probabilistic approximation of the Wasserstein distance with exact solvers, respectively. Each chapter begins with a brief overview of the results presented followed by the main body of text. The first and third chapter conclude with a discussion of the presented results and possible directions for further research. Most proofs are given in a designated section within the respective chapter.

Chapter 2

Distributional limits

This chapter gives distributional limits for empirical Wasserstein distances on finite spaces. In the first section, the main result is presented, followed by two sections outlining the notions and results required for its proof. The fourth, fifth and sixth section consider cases in which the limiting distribution has an easier form. In particular, the fourth section gives an explicit expression for the limiting distribution when the underlying metric is generated by a tree. The fifth section demonstrates that the limiting distribution under the null hypothesis of equal measures can be written as a Wasserstein distance. The sixth section gives conditions on the underlying measures under which the limiting distribution under the alternative (the true measures being different) is normal. The seventh section discussed failure of the naive bootstrap under the null hypothesis and possible alternatives. The eigth section gives an alternative, numerically more stable representation of the limiting distribution for different measures. Finally, the eighth section contains simulations assessing the speed of convergence to the limiting distribution and applications under the null hypothesis as well as the alternative.

The chapter is concluded with a discussion section and a section containing the proofs of the presented results.

2.1 Main result

In this section we give a comprehensive result on distributional limits for the Wasserstein distance when the underlying population measures are supported on finitely many points $\mathcal{X} = \{x_1, \dots, x_N\}$. We denote the inner product on the vector space $\mathbb{R}^{\mathcal{X}}$ by $\langle \boldsymbol{u}, \boldsymbol{u}' \rangle = \sum_{x \in \mathcal{X}} u_x u'_x$ for $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}^{\mathcal{X}}$.

Theorem 1. *Let $p \geq 1$, $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_{\mathcal{X}}$ and $\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m$ generated by i.i.d. samples $X_1, \dots, X_n \sim \boldsymbol{r}$ and $Y_1, \dots, Y_m \sim \boldsymbol{s}$, respectively. We define the convex sets*

$$\begin{aligned} \Phi_p^* &= \left\{ \boldsymbol{u} \in \mathbb{R}^{\mathcal{X}} : u_x - u_{x'} \leq d^p(x, x'), \quad x, x' \in \mathcal{X} \right\} \\ \Phi_p^*(\boldsymbol{r}, \boldsymbol{s}) &= \left\{ (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}} : \begin{array}{l} \langle \boldsymbol{u}, \boldsymbol{r} \rangle + \langle \boldsymbol{v}, \boldsymbol{s} \rangle = W_p^p(\boldsymbol{r}, \boldsymbol{s}), \\ u_x + v_{x'} \leq d^p(x, x'),\ x, x' \in \mathcal{X} \end{array} \right\} \end{aligned} \tag{2.1}$$

and the multinomial covariance matrix

$$\Sigma(\boldsymbol{r}) = \begin{bmatrix} r_{x_1}(1 - r_{x_1}) & -r_{x_1} r_{x_2} & \cdots & -r_{x_1} r_{x_N} \\ -r_{x_2} r_{x_1} & r_{x_2}(1 - r_{x_2}) & \cdots & -r_{x_2} r_{x_N} \\ \vdots & & \ddots & \vdots \\ -r_{x_N} r_{x_1} & -r_{x_N} r_{x_2} & \cdots & r_{x_N}(1 - r_{x_N}) \end{bmatrix} \tag{2.2}$$

such that with independent Gaussian random variables $\boldsymbol{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ and $\boldsymbol{H} \sim \mathcal{N}(0, \Sigma(\boldsymbol{s}))$ we have the following.

a) ***(One sample - Null hypothesis)*** *With the sample size n approaching infinity, we have the weak convergence*

$$n^{\frac{1}{2p}} W_p(\hat{\boldsymbol{r}}_n, \boldsymbol{r}) \Rightarrow \left\{ \max_{\boldsymbol{u} \in \Phi_p^*} \langle \boldsymbol{G}, \boldsymbol{u} \rangle \right\}^{\frac{1}{p}}. \tag{2.3}$$

b) ***(One sample - Alternative)*** *With n approaching infinity we have*

$$n^{\frac{1}{2}} \left(W_p(\hat{\boldsymbol{r}}_n, \boldsymbol{s}) - W_p(\boldsymbol{r}, \boldsymbol{s}) \right) \Rightarrow \frac{1}{p} W_p^{1-p}(\boldsymbol{r}, \boldsymbol{s}) \left\{ \max_{(\boldsymbol{u}, \boldsymbol{v}) \in \Phi_p^*(\boldsymbol{r}, \boldsymbol{s})} \langle \boldsymbol{G}, \boldsymbol{u} \rangle \right\}. \tag{2.4}$$

c) ***(Two samples - Null hypothesis)*** *Let $\rho_{n,m} = (nm/(n+m))^{1/2}$. If $\boldsymbol{r} = \boldsymbol{s}$ and n and m are approaching infinity such that $n \wedge m \to \infty$ and*

$m/(n+m) \to \lambda \in (0,1)$ *we have*

$$\rho_{n,m}^{1/p} W_p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) \Rightarrow \left\{ \max_{\boldsymbol{u} \in \Phi_p^*} \langle \boldsymbol{G}, \boldsymbol{u} \rangle \right\}^{\frac{1}{p}}. \tag{2.5}$$

d) ***(Two samples - Alternative)*** *With n and m approaching infinity such that $n \wedge m \to \infty$ and $m/(n+m) \to \lambda \in [0,1]$*

$$\begin{aligned} &\rho_{n,m} \left(W_p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) - W_p(\boldsymbol{r}, \boldsymbol{s}) \right) \Rightarrow \\ &\quad \frac{1}{p} W_p^{1-p}(\boldsymbol{r}, \boldsymbol{s}) \left\{ \max_{(\boldsymbol{u}, \boldsymbol{v}) \in \Phi_p^*(\boldsymbol{r}, \boldsymbol{s})} \sqrt{\lambda} \langle \boldsymbol{G}, \boldsymbol{u} \rangle + \sqrt{1-\lambda} \langle \boldsymbol{H}, \boldsymbol{v} \rangle \right\}. \end{aligned} \tag{2.6}$$

The sets Φ_p^* and $\Phi_p^*(\boldsymbol{r}, \boldsymbol{s})$ are (derived from) the dual solutions to the Wasserstein linear program (see Theorem 4 below). This result is valid for all probability measures with finite support, regardless of the (dimension of the) underlying space. In particular, it generalizes a result of Samworth and Johnson (2004), who considered a finite collection of point masses on the real line and $p = 2$. We will re-obtain their result as a special case in Section 2.4 when we give explicit expressions for the limit distribution when the metric d, which enters the limit law via the dual solutions Φ_p^* or $\Phi_p^*(\boldsymbol{r}, \boldsymbol{s})$, is given by a tree.

Remark 1. In our numerical experiments (see Section 2.9 we have found the representation (2.6) to be numerically unstable when used to simulate from the limiting distribution under the alternative. We therefore give an alternative representation (2.27) in the supplementary material as a one-dimensional optimization problem of a non-linear function (in contrast to a high-dimensional linear program shown here). Note that the limiting distribution under the null does not suffer from this problem and can be simulated from directly using a linear program solver.

The scaling rate in Theorem 1 depends solely on p and is completely independent of the underlying space $\mathcal{X}$. This contrasts known bounds on the rate of convergence in the continuous case. We will elaborate on the differences in the discussion. Typical choices are $p = 1, 2$. The faster scaling

rate can be a reason to favor $p = 1$. In our numerical experiments however, this advantage was frequently outweighed by larger quantiles of the limiting distribution.

Dümbgen (1993) showed that the naive n-out-of-n bootstrap is inconsistent for functionals with a non-linear Hadamard derivative, but resampling fewer than n observations leads to a consistent bootstrap. Since we will show in the following that the Wasserstein distance belongs to this class of functionals, it is a direct consequence that the naive bootstrap fails for the Wasserstein distance (see Section 2.7 in the supplementary material for details) and that the following holds.

Theorem 2. *Let $\hat{\boldsymbol{r}}_n^*$ and $\hat{\boldsymbol{s}}_m^*$ be bootstrap versions of $\hat{\boldsymbol{r}}_n$ and $\hat{\boldsymbol{s}}_m$ that are obtained via re-sampling k observations with $k/n \to 0$ and $k/m \to 0$. Then, the plug-in bootstrap with $\hat{\boldsymbol{r}}_n^*$ and $\hat{\boldsymbol{s}}_m^*$ is consistent, that is*

$$\sup_{f \in \mathrm{BL}_1(\mathbb{R})} E\left[f(\phi_p(\sqrt{k}\left\{(\hat{\boldsymbol{r}}_n^{**}, \hat{\boldsymbol{s}}_m^{**}) - (\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\}))|X_1, \ldots, X_n, Y_1, \ldots, Y_m\right] - E\left[f\left(\rho_{n,m}\left\{W_p^p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) - W_p^p(\boldsymbol{r}, \boldsymbol{s})\right\}\right)\right]$$

converges to zero in probability.

In the following we will prove our main Theorem 1 by

i) introducing Hadamard directional differentiability, which does not require the derivative to be linear but still allows for a delta-method;

ii) showing that the map $(\boldsymbol{r}, \boldsymbol{s}) \mapsto W_p(\boldsymbol{r}, \boldsymbol{s})$ is differentiable in this sense.

2.2 Hadamard directional derivatives

In this section we follow Römisch (2004). A map f defined on a subset $D_f \subset \mathbb{R}^d$ with values in $\mathbb{R}$ is called *Hadamard directionally differentiable* at $\boldsymbol{u} \in \mathbb{R}^d$ if there exists a map $f'_{\boldsymbol{u}} : \mathbb{R}^d \to \mathbb{R}$ such that

$$\lim_{n \to \infty} \frac{f(\boldsymbol{u} + t_n \boldsymbol{h}_n) - f(\boldsymbol{u})}{t_n} = f'_{\boldsymbol{u}}(\boldsymbol{h}) \tag{2.7}$$

for any $\boldsymbol{h} \in \mathbb{R}^d$ and for arbitrary sequences t_n converging to zero from above and $\boldsymbol{h}_n$ converging to $\boldsymbol{h}$ such that $\boldsymbol{u} + t_n \boldsymbol{h}_n \in D_f$ for all $n \in \mathbb{N}$. Note that in contrast to the usual notion of Hadamard differentiability (e.g. Van der Vaart and Wellner (1996)) the derivative $\boldsymbol{h} \mapsto f'_{\boldsymbol{u}}(\boldsymbol{h})$ is *not* required to be linear. A prototypical example is the absolute value $f : \mathbb{R} \to \mathbb{R}$, $t \mapsto |t|$ which is not in the usual sense Hadamard differentiable at $t = 0$ but directionally differentiable with the non-linear derivative $t \mapsto |t|$.

Theorem 3 (Römisch, 2004, Theorem 1)**.** *Let f be a function defined on a subset F of $\mathbb{R}^d$ with values in $\mathbb{R}$, such that*

1. *f is Hadamard directionally differentiable at $\boldsymbol{u} \in F$ with derivative $f'_{\boldsymbol{u}} : F \to \mathbb{R}$ and*
2. *there is a sequence of $\mathbb{R}^d$-valued random variables X_n and a sequence of non-negative numbers $\rho_n \to \infty$ such that $\rho_n(X_n - \boldsymbol{u}) \Rightarrow X$ for some random variable X taking values in F.*

Then, $\rho_n(f(X_n) - f(\boldsymbol{u})) \Rightarrow f'_{\boldsymbol{u}}(X)$.

2.3 Directional derivative of the Wasserstein distance

In this section we show that the functional $(\boldsymbol{r}, \boldsymbol{s}) \mapsto W_p^p(\boldsymbol{r}, \boldsymbol{s})$ is Hadamard directionally differentiable and give a formula for the derivative.

The *dual* program (cf. (Luenberger and Ye, 2008, Ch. 4), also Kantorovich and Rubinstein (1958)) of the linear program defining the Wasserstein distance (1.1) is given by

$$\begin{aligned} &\max_{(\boldsymbol{u},\boldsymbol{v}) \in \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}}} && \langle \boldsymbol{u}, \boldsymbol{r} \rangle + \langle \boldsymbol{s}, \boldsymbol{v} \rangle \\ &\textbf{s.t.} && u_x + v_{x'} \leq d^p(x, x') \quad \forall x, x' \in \mathcal{X}. \end{aligned} \tag{2.8}$$

As noted above, the optimal value of the primal problem is $W_p^p(\boldsymbol{r}, \boldsymbol{s})$ and by standard duality theory of linear programs (e.g. Luenberger and Ye (2008))

this is also the optimal value of the dual problem. Therefore, the set of optimal solutions to the dual problem is given by $\Phi_p^*(\boldsymbol{r}, \boldsymbol{s})$ as defined in (2.1).

Theorem 4. *The functional* $(\boldsymbol{r}, \boldsymbol{s}) \mapsto W_p^p(\boldsymbol{r}, \boldsymbol{s})$ *is directionally Hadamard differentiable at all* $(\boldsymbol{r}, \boldsymbol{s}) \in \mathcal{P}_\mathcal{X} \times \mathcal{P}_\mathcal{X}$ *with derivative*

$$(\boldsymbol{h}_1, \boldsymbol{h}_2) \mapsto \max_{(\boldsymbol{u},\boldsymbol{v}) \in \Phi_p^*(\boldsymbol{r},\boldsymbol{s})} -(\langle \boldsymbol{u}, \boldsymbol{h}_1 \rangle + \langle \boldsymbol{v}, \boldsymbol{h}_2 \rangle). \tag{2.9}$$

We can give a more explicit expression for the set $\Phi_p^*(\boldsymbol{r}, \boldsymbol{s})$ in the case $\boldsymbol{r} = \boldsymbol{s}$, when the optimal value of the primal and the dual problem is 0. Then, the condition $W_p^p(\boldsymbol{r}, \boldsymbol{s}) = \langle \boldsymbol{r}, \boldsymbol{u} \rangle + \langle \boldsymbol{s}, \boldsymbol{v} \rangle$ becomes $\langle \boldsymbol{r}, \boldsymbol{u} + \boldsymbol{v} \rangle = 0$. Since $u_x + v_{x'} \leq d^p(x, x')$ for all $x, x' \in \mathcal{X}$ implies $\boldsymbol{u} + \boldsymbol{v} \leq 0$ this yields $\boldsymbol{u} = -\boldsymbol{v}$. This gives

$$\Phi_p^*(\boldsymbol{r}, \boldsymbol{r}) = \left\{ (\boldsymbol{u}, -\boldsymbol{u}) \in \mathbb{R}^\mathcal{X} \times \mathbb{R}^\mathcal{X} : u_x - u_{x'} \leq d^p(x, x'),\ x, x' \in \mathcal{X} \right\}$$

and the following more compact representation of the dual solutions in the case $\boldsymbol{r} = \boldsymbol{s}$, independent of $\boldsymbol{r}$:

$$\Phi_p^*(\boldsymbol{r}, \boldsymbol{r}) = \Phi_p^* \times \left(-\Phi_p^* \right). \tag{2.10}$$

2.4 Explicit limiting distribution for tree metrics

Assume that the metric structure on $\mathcal{X}$ is given by a weighted tree, that is, an undirected connected graph $\mathcal{T} = (\mathcal{X}, E)$ with vertices $\mathcal{X}$ and edges $E \subset \mathcal{X} \times \mathcal{X}$ that contains no cycles. We assume the edges to be weighted by a function $w : E \to \mathbb{R}_{>0}$. For $x, x' \in \mathcal{X}$ let $e_1, \ldots, e_l \in E$ be the unique path in $\mathcal{T}$ joining x and x', then the length of this path, $d_\mathcal{T}(x, x') = \sum_{j=1}^l w(e_j)$ defines a metric $d_\mathcal{T}$ on $\mathcal{X}$. Without imposing any further restriction on $\mathcal{T}$, we assume it to be rooted at $\mathrm{root}(\mathcal{T}) \in \mathcal{X}$, say. Then, for $x \in \mathcal{X}$ and $x \neq \mathrm{root}(\mathcal{T})$ we may define $\mathrm{par}(x) \in \mathcal{X}$ as the immediate neighbor of x in the unique path connecting x and $\mathrm{root}(\mathcal{T})$. We set $\mathrm{par}(\mathrm{root}(\mathcal{T})) = \mathrm{root}(\mathcal{T})$.

We also define children(x) as the set of vertices $x' \in \mathcal{X}$ such that there exists a sequence $x' = x_1, \ldots, x_l = x \in \mathcal{X}$ with $\text{par}(x_j) = x_{j+1}$ for $j = 1, \ldots, l-1$. Note that with this definition $x \in \text{children}(x)$. Additionally, define the linear operator $S_{\mathcal{T}} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$

$$(S_{\mathcal{T}}\boldsymbol{u})_x = \sum_{x' \in \text{children}(x)} u_{x'}. \tag{2.11}$$

Theorem 5. *Let $p \geq 1$, $\boldsymbol{r} \in \mathcal{P}_{\mathcal{X}}$, defining a probability distribution on $\mathcal{X}$ and let the empirical measures $\hat{\boldsymbol{r}}_n$ and $\hat{\boldsymbol{s}}_m$ be generated by independent random variables $X_1, \ldots, X_n$ and $Y_1, \ldots Y_m$, respectively, all drawn from $\boldsymbol{r} = \boldsymbol{s}$.*

Then, with a Gaussian vector $\boldsymbol{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ as defined in (2.2) we have the following.

a) ***(One sample)*** *As $n \to \infty$,*

$$n^{\frac{1}{2p}} W_p(\hat{\boldsymbol{r}}_n, \boldsymbol{r}) \Rightarrow \left\{ \sum_{x \in \mathcal{X}} |(S_{\mathcal{T}}\boldsymbol{G})_x| d_{\mathcal{T}}(x, \text{par}(x))^p \right\}^{\frac{1}{p}} \tag{2.12}$$

b) ***(Two samples)*** *If $n \wedge m \to \infty$ and $n/(n+m) \to \lambda \in (0,1)$ we have*

$$\left(\frac{nm}{n+m} \right)^{\frac{1}{2p}} W_p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) \Rightarrow \left\{ \sum_{x \in \mathcal{X}} |(S_{\mathcal{T}}\boldsymbol{G})_x| d_{\mathcal{T}}(x, \text{par}(x))^p \right\}^{\frac{1}{p}}. \tag{2.13}$$

The proof of Theorem 5 is given in the supplementary material. The theorem includes the special case of a discrete measure on the real line, that is $\mathcal{X} \subset \mathbb{R}$, since in this case, $\mathcal{X}$ can be regarded as a simple rooted tree consisting of only one branch.

Corollary 1 (Samworth and Johnson, 2004, Theorem 2.6)**.** *Let $\mathcal{X} = \{x_1 < \cdots < x_N\} \in \mathbb{R}$, $\boldsymbol{r} \in \mathcal{P}_{\mathcal{X}}$ and $\hat{\boldsymbol{r}}_n$ the empirical measure generated by i.i.d. random variables $X_1, \ldots, X_n \sim \boldsymbol{r}$. With $\bar{r}_j = \sum_{i=1}^{j} r_{x_i}$, for $j = 1, \ldots N$ and*

B a standard Brownian bridge, we have as $n \to \infty$,

$$n^{\frac{1}{4}} W_2(\hat{\boldsymbol{r}}_n, \boldsymbol{r}) \Rightarrow \left\{ \sum_{j=1}^{N-1} |B(\bar{r}_j)|(x_{j+1} - x_j)^2 \right\}^{\frac{1}{2}}. \tag{2.14}$$

2.5 The limiting distribution as a Wasserstein distance

The limiting distribution (2.5) under the null hypothesis can be written as a transport distance between random measures. Besides its theoretical appeal, this result has practical implications. Any solver for the Wasserstein problem can also be directly used for Monte Carlo simulation of the limiting distribution.

For the sake of brevity we will in this section use the notation $W_p^p(\boldsymbol{r}, \boldsymbol{s})$ also for vectors $\boldsymbol{r}, \boldsymbol{s} \in \mathbb{R}^{\mathcal{X}}_{\geq 0}$ which are not probability measures but satisfy $\sum_x r_x = \sum_x s_x$. One may read this as

$$W_p^p(\boldsymbol{r}, \boldsymbol{s}) = \left(\sum_x r_x \right) \times W_p^p \left(\frac{\boldsymbol{r}}{\sum_x r_x}, \frac{\boldsymbol{s}}{\sum_x s_x} \right).$$

Theorem 6. *Let* $\boldsymbol{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ *as in* (2.2) *and define* $\boldsymbol{G}^+ \in \mathbb{R}^{\mathcal{X}}$

$$\boldsymbol{G}^+ = \begin{cases} G_x & \text{if } G_x > 0 \\ 0 & \text{else,} \end{cases}$$

as well as $\boldsymbol{G}^- = \boldsymbol{G} - \boldsymbol{G}^+$, *such that* $\boldsymbol{G}^{\pm}$ *have only non-negative entries and* $\boldsymbol{G} = \boldsymbol{G}^+ - \boldsymbol{G}^-$. *Further, let* $\mathbf{1} \in \mathbb{R}^{\mathcal{X}}$ *be the vector of ones, that is* $\mathbf{1}_x = 1$ *for all* $x \in \mathcal{X}$. *Then,*

$$\max_{\boldsymbol{u} \in \Phi^*} \langle \boldsymbol{G}, \boldsymbol{u} \rangle = W_p^p(\boldsymbol{G}^+ + c\mathbf{1}, \boldsymbol{G}^- + c\mathbf{1}) \tag{2.15}$$

for all $c > (\min_{x,x' \in \mathcal{X}} d^p(x, x'))^{-1} W_p^p(\boldsymbol{G}^+, \boldsymbol{G}^-)$.

Remark 2. The constant $(\min_{x,x' \in \mathcal{X}} d^p(x, x'))^{-1} W_p^p(\boldsymbol{G}^+, \boldsymbol{G}^-)$ may be upper

bounded by

$$\left(\min_{x,x'\in\mathcal{X}} d^p(x,x')\right)^{-1} (\operatorname{diam}(\mathcal{X}))^p \sum_x G_x^+$$

which can easily be computed for any given $\boldsymbol{G}$. It may become very large (e.g. when $\mathcal{X}$ is a regular grid in dimension D it will be of order $N^{1/D}$) but this has no influence on the computational burden of the right hand side in (2.15), since the size of the transport problem remains unaltered.

We suspect, that the statement of the theorem remains valid if only $c > 1$, but it appears that this is more difficult to prove.

Proof. Recall that

$$\begin{aligned}
&\max_{\boldsymbol{u}\in\Phi_p^*} \langle \boldsymbol{G}, \boldsymbol{u}\rangle = \max \langle \boldsymbol{G}, \boldsymbol{u}\rangle \\
&\textbf{s.t. } u_x - u_{x'} \le d^p(x,x') \forall x, x' \in \mathcal{X}.
\end{aligned}$$

By introducing the new variable $\boldsymbol{v} = -\boldsymbol{u}$ we can rewrite this as

$$\begin{aligned}
&\max \langle \boldsymbol{G}^+, \boldsymbol{u}\rangle + \langle \boldsymbol{G}^-, \boldsymbol{v}\rangle \\
&\textbf{s.t. } u_x - v_{x'} \le d^p(x,x') \forall x, x' \in \mathcal{X} \\
&\quad \boldsymbol{u} + \boldsymbol{v} = 0.
\end{aligned}$$

The linear programming dual (Luenberger and Ye, 2008, Ch. 4) of this is

$$\begin{aligned}
&\min \sum_{x,x'\in\mathcal{X}} w_{x,x'} d^p(x,x') \\
&\textbf{s.t. } \boldsymbol{w} \ge 0, \boldsymbol{z} \in \mathbb{R}^{\mathcal{X}} \\
&\quad \sum_{x'} w_{x,x'} - z_x = G_x^+ \\
&\quad \sum_{x} w_{x,x'} - z_{x'} = G_{x'}^-.
\end{aligned} \tag{2.16}$$

First, we note that any feasible solution must satisfy $\boldsymbol{z} \ge 0$. To see this, assume that $z_x < 0$ for some $x \in \mathcal{X}$. By definition, at least one of G_x^+ and G_x^- is zero. Without loss of generality, assume $G_x^+ = 0$, yielding $0 < \sum_{x'} w_{x,x'} - z_x = 0$, a contradiction.

Evidently, when $\boldsymbol{z} \geq 0$ the optimum of the last linear program is

$$\min_{\boldsymbol{z} \in \mathbb{R}^{\mathcal{X}}_{\geq 0}} W_p^p(\boldsymbol{G}^+ + \boldsymbol{z}, \boldsymbol{G}^- + \boldsymbol{z}).$$

We will now consider the function $\boldsymbol{z} \mapsto W_p^p(\boldsymbol{G}^+ + \boldsymbol{z}, \boldsymbol{G}^- + \boldsymbol{z})$. To this end, for $\boldsymbol{u} \in \mathbb{R}^{\mathcal{X}}$ define $\mathrm{diag}(\boldsymbol{u}) \in \mathbb{R}^{\mathcal{X} \times \mathcal{X}}$ as

$$(\mathrm{diag}(\boldsymbol{u}))_{x,x'} = \begin{cases} u_x & \text{if } x = x', \\ 0 & \text{else.} \end{cases}$$

Note that $\mathrm{diag}(\boldsymbol{u})$ plugged into the objective function of (2.16) gives zero for all $\boldsymbol{u} \in \mathbb{R}^{\mathcal{X}}$.

Assume that $\boldsymbol{z}_1 \leq \boldsymbol{z}_2$ (component-wise) and let $\boldsymbol{w}^*$ the optimal coupling of $\boldsymbol{G}^+ + \boldsymbol{z}_1$ and $\boldsymbol{G}^- + \boldsymbol{z}_1$. Then, $\boldsymbol{w}^* + \mathrm{diag}(\boldsymbol{z}_2 - \boldsymbol{z}_1)$ is a coupling of $\boldsymbol{G}^+ + \boldsymbol{z}_2$ and $\boldsymbol{G}^- + \boldsymbol{z}_2$ with cost $W_p^p(\boldsymbol{G}^+ + \boldsymbol{z}_1, \boldsymbol{G}^- + \boldsymbol{z}_1)$. Hence,

$$W_p^p(\boldsymbol{G}^+ + \boldsymbol{z}_2, \boldsymbol{G}^- + \boldsymbol{z}_2) \leq W_p^p(\boldsymbol{G}^+ + \boldsymbol{z}_1, \boldsymbol{G}^- + \boldsymbol{z}_1).$$

Now, let $c_0 = (\min_{x,x' \in \mathcal{X}} d^p(x, x'))^{-1} W_p^p(\boldsymbol{G}^+, \boldsymbol{G}^-)$ and assume that $\boldsymbol{z} \geq c_o \mathbf{1}$. Then, with $\boldsymbol{w}^*$ the optimal coupling of $\boldsymbol{G}^+ + \boldsymbol{z}$ and $\boldsymbol{G}^- + \boldsymbol{z}$ we have that for any $x \in \mathcal{X}$

$$\begin{aligned} w^*_{x,x} &= G^+_x + z_x - \sum_{x' \in \mathcal{X}, x' \neq x} w^*_{x,x'} \\ &\geq z_x - \sum_{x' \in \mathcal{X}, x' \neq x} w^*_{x,x'} \\ &\geq z_x - \left(\min_{x,x' \in \mathcal{X}} d^p(x, x') \right)^{-1} \left(\sum_{x,x'} w^*_{x,x'} d^p(x, x') \right) \\ &\geq z_x - \left(\min_{x,x' \in \mathcal{X}} d^p(x, x') \right)^{-1} W_p^p(\boldsymbol{G}^+, \boldsymbol{G}^-) \\ &\geq z_x - c_0. \end{aligned}$$

Hence,

$$\boldsymbol{w}^* + \mathrm{diag}(c_0 - \boldsymbol{z})$$

has only non-negative entries and is therefore a coupling of $\boldsymbol{G}^+ + c_0\mathbf{1}$ and $\boldsymbol{G}^- + c_0\mathbf{1}$ with cost $W_p^p(\boldsymbol{G}^+ + \boldsymbol{z}, \boldsymbol{G}^- + \boldsymbol{z})$. Therefore,

$$W_p^p(\boldsymbol{G}^+ + c_0\mathbf{1}, \boldsymbol{G}^- + c_0\mathbf{1}) \leq W_p^p(\boldsymbol{G}^+ + \boldsymbol{z}, \boldsymbol{G}^- + \boldsymbol{z}).$$

It follows that the function $z \mapsto W_p^p(\boldsymbol{G}^+ + \boldsymbol{z}, \boldsymbol{G}^- + \boldsymbol{z})$ assumes its minimum at every point $\boldsymbol{z} \geq c_0\mathbf{1}$. □

2.6 Normal limits under the alternative

Under certain conditions, the limiting distribution under the alternative $\boldsymbol{r} = \boldsymbol{s}$ is normal. We say that two measures $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_\mathcal{X}$ satisfy the *non-degenracy condition* if

$$\sum_{x \in A} r_x \neq \sum_{x' \in B} s_{x'} \quad \text{for all proper subsets } A \subsetneqq \mathcal{X} \text{ and } B \subsetneqq \mathcal{X}. \tag{2.17}$$

Theorem 7 (Theorem and Definition)**.** *If $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_\mathcal{X}$ satisfy the non-degeneracy condition* (2.17) *and $\boldsymbol{u}^*, \boldsymbol{v}^*$ is a solution to the dual transportation problem* (2.8)*, then any other solution is of the form $\boldsymbol{u}^* + c$, $\boldsymbol{v}^* - c$ for some $c \in \mathbb{R}$. Hence, the following are independent of the choice of a solution $\boldsymbol{u}^*$, $\boldsymbol{v}^*$*

$$\begin{aligned} \sigma_1^2(\boldsymbol{r}, \boldsymbol{s}) &= \sum_{x \in \mathcal{X}} (u_x^*)^2 r_x - \left(\sum_{x \in \mathcal{X}} u_x^* r_x \right)^2 \\ \sigma_2^2(\boldsymbol{r}, \boldsymbol{s}) &= \sum_{x \in \mathcal{X}} (v_x^*)^2 s_x - \left(\sum_{x \in \mathcal{X}} v_x^* s_x \right)^2 . \end{aligned} \tag{2.18}$$

If $\boldsymbol{r}, \boldsymbol{s}$ do not satisfy the non-degeneracy condition, we define $\boldsymbol{u}^*, \boldsymbol{v}^*$ to be the lexicographically smallest dual solution and define $\sigma_{1,2}^2(\boldsymbol{r}, \boldsymbol{s})$ as above.

Proof. If the condition (2.17) is satisfied, then the transport simplex

$$\left\{ \boldsymbol{w} \in \mathcal{P}_{\mathcal{X}\times\mathcal{X}} : \sum_{x'} w_{x,x'} = r_x \text{ and } \sum_{x} w_{x,x'} = s_{x'} \right\}$$

is non-degenerate in the sense of linear programming. That is, every vertex of the above transport simplex has exactly $2N-1$ non-zero entries. We refer to (Luenberger and Ye, 2008, Ch.3) for a definition of non-degeneracy in the context of linear programming and to (Klee and Witzgall, 1968, Cor. 3) and Hung et al. (1986) for the fact that in the case of a transportation problem, non-degeneracy is equivalent to (2.17).

Therefore any primal solution to the transportation problem (and such a solution always exists) will be non-degenerate (after deleting one linear constraint to make them linearly independent) and therefore the dual transportation problem has a unique solution up to an additive constant (since deleting one constraint in the primal corresponds to fixing one coordinate of the solution in the dual) (Sierksma, 2001, Thm. 4.5). Note that this additive constant will not change the value of the limiting distribution since $\sum_x G_x = 0$ whenever $\boldsymbol{G} \sim \Sigma(\boldsymbol{r})$. □

Theorem 8. *Let $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_{\mathcal{X}}$ be measures that satisfy the non-degeneracy condition* (2.17) *and $\hat{\boldsymbol{r}}_n$ and $\hat{\boldsymbol{s}}_m$ empirical versions as in Theorem 1. Further, let $\boldsymbol{G}$ and $\boldsymbol{H}$ be independent Gaussian random vectors with mean zero and covariance $\Sigma(\boldsymbol{r})$ and $\Sigma(\boldsymbol{s})$ as defined in* (2.2)*, respectively, then*

a) ***(One sample)*** *with n approaching infinity we have*

$$\frac{n^{\frac{1}{2}} \left(W_p(\hat{\boldsymbol{r}}_n, \boldsymbol{s}) - W_p(\boldsymbol{r}, \boldsymbol{s})\right)}{\frac{1}{p} W_p^{1-p}(\hat{\boldsymbol{r}}_n, \boldsymbol{s}) \sigma_1(\hat{\boldsymbol{r}}_n, \boldsymbol{s})} \Rightarrow \mathcal{N}(0,1). \tag{2.19}$$

b) ***(Two sample)*** *with n and m approaching infinity such that $n \wedge m \to \infty$ and $m/(n+m) \to \lambda \in [0,1]$,*

$$\frac{\rho_{n,m} \left(W_p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) - W_p(\boldsymbol{r}, \boldsymbol{s})\right)}{\frac{1}{p} W_p^{1-p}(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) \sqrt{\lambda \sigma_1^2(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) + (1-\lambda)\sigma_2^2(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)}} \Rightarrow \mathcal{N}(0,1). \tag{2.20}$$

Proof. We only prove the two sample case, the one sample case follows anal-

ogously. From Theorems 1 and 7 we know that

$$\begin{aligned}&\rho_{n,m}\left(W_p(\hat{\boldsymbol{r}}_n,\hat{\boldsymbol{s}}_m)-W_p(\boldsymbol{r},\boldsymbol{s})\right)\Rightarrow\\&\quad\frac{1}{p}W_p^{1-p}(\boldsymbol{r},\boldsymbol{s})\left\{\sqrt{\lambda}\langle\boldsymbol{G},\boldsymbol{u}^*\rangle+\sqrt{1-\lambda}\langle\boldsymbol{H},\boldsymbol{v}^*\rangle\right\},\end{aligned}\tag{2.21}$$

with the unique dual solutions $\boldsymbol{u}^*, \boldsymbol{v}^*$. Note that

$$\begin{aligned}\operatorname{var}\left[\langle\boldsymbol{G},\boldsymbol{u}^*\rangle\right]&=\sum_{x,x'\in\mathcal{X}}\left(\Sigma(\boldsymbol{r})\right)_{x,x'}u_x^*u_{x'}^*\\&=-\sum_{x\neq x'}u_x^*u_{x'}^*r_xr_{x'}+\sum_x(u_x^*)^2r_x(1-r_x)\\&=\sum_x(u_x^*)^2r_x-\sum_{x,x'}u_x^*u_{x'}^*r_xr_{x'}\\&=\sigma_1^2(\boldsymbol{r},\boldsymbol{s}).\end{aligned}$$

Hence, the limit in (2.21) is a mean zero normal distribution with standard deviation

$$\frac{1}{p}W_p^{1-p}(\boldsymbol{r},\boldsymbol{s})\sqrt{\lambda\sigma_1^2(\boldsymbol{r},\boldsymbol{s})+(1-\lambda)\sigma_2^2(\boldsymbol{r},\boldsymbol{s})}.$$

The statement will follow from Slutzky's Theorem if we show that this is the limit (in probability) of the empirical version of this term

$$\frac{1}{p}W_p^{1-p}(\hat{\boldsymbol{r}}_n,\hat{\boldsymbol{s}}_m)\sqrt{\lambda\sigma_1^2(\hat{\boldsymbol{r}}_n,\hat{\boldsymbol{s}}_m)+(1-\lambda)\sigma_2^2(\hat{\boldsymbol{r}}_n,\hat{\boldsymbol{s}}_m)}.$$

It is clear that $W_p(\hat{\boldsymbol{r}}_n,\hat{\boldsymbol{s}}_m)\to W_p(\boldsymbol{r},\boldsymbol{s})$ in probability. Hence, it remains to show that $\sigma_j^2(\hat{\boldsymbol{r}}_n,\hat{\boldsymbol{s}}_m)\to\sigma_j^2(\boldsymbol{r},\boldsymbol{s})$ in probability. The latter will follow from the continuous mapping theorem if we can show that the dual solutions $(\boldsymbol{u}^*,\boldsymbol{v}^*)$ are stable in the following sense: if $(\boldsymbol{r}_k,\boldsymbol{s}_k)$ is a (deterministic) sequence of measures converging to $(\boldsymbol{r},\boldsymbol{s})$ we need to show that the corresponding sequence $(\boldsymbol{u}_k^*,\boldsymbol{v}_k^*)$ of dual solutions converges to $(\boldsymbol{u}^*,\boldsymbol{v}^*)$. This stability follows, for example, from Theorem 1 of Robinson (1977), noting that the set of primal and dual solutions of the transportation problem are bounded if $\boldsymbol{r},\boldsymbol{s}$ satisfy the non-degeneracy condition. This concludes the proof. □

2.6.1 The non-degeneracy condition

In this section we study in more detail the non-degeneracy condition (2.17). In particular, we address how restrictive the condition is.

Remark 3. The problem of determining whether a given pair of measures $(\boldsymbol{r}, \boldsymbol{s}) \in \mathcal{P}_{\mathcal{X}\times\mathcal{X}}$ satisfies the non-degeneracy condition is NP-complete (Chandrasekaran et al., 1982).

It seems to be well-known in mathematical programming that a small perturbation can usually remove non-degeneracy from a linear program. In the following result we give some formal statements with regard to this, in particular, with a view towards our statistical application.

Proposition 1. *a) For fixed $N \in \mathbb{N}$ the set of pairs of measures $(\boldsymbol{r}, \boldsymbol{s}) \in \mathcal{P}_{\mathcal{X}\times\mathcal{X}}$ that satisfy the non-degeneracy condition is open and dense in $\mathcal{P}_{\mathcal{X}\times\mathcal{X}}$.*

b) If $(\boldsymbol{r}, \boldsymbol{s})$ satisfy the non-degeneracy condition and $(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)$ are consistent estimators then

$$P\left[(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) \text{ satisfy the non-degeneracy condition}\right] \to 1 \quad (n, m \to \infty).$$

In particular, the dual solutions to the transport problem with marginals $(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)$ will be unique with probability tending to one.

c) If $(\boldsymbol{r}, \boldsymbol{s})$ are drawn randomly from some distribution on $\mathcal{P}_{\mathcal{X}\times\mathcal{X}}$ which is absolutely continuous with respect to the Lebesgue measure, then

$$P\left[(\boldsymbol{r}, \boldsymbol{s}) \text{ satisfy the non-degeneracy condition}\right] = 1.$$

Proof. The set of measures satisfying the non-degeneracy condition can be written as

$$\bigcap_{A,B \subsetneq \mathcal{X}} \left\{ (\boldsymbol{r}, \boldsymbol{s}) \in \mathcal{P}_{\mathcal{X}\times\mathcal{X}} : \sum_{x\in A} r_x \neq \sum_{x'\in B} s_{x'} \right\}. \tag{2.22}$$

This is the intersection of the complements of

$$\left\{(\boldsymbol{r},\boldsymbol{s}) \in \mathcal{P}_{\mathcal{X}\times\mathcal{X}} : \sum_{x\in A} r_x = \sum_{x'\in B} s_{x'}\right\}$$

which are closed subsets of dimension one and hence Lebesgue zero-sets. Consequently, (2.22) is open and dense as the intersection of finitely many open sets with co-dimension 1 and it has measure one with respect to any measure that has a Lebesgue density. This proves the first and third part.

For the second part let $\epsilon > 0$ such that every $(\boldsymbol{r}', \boldsymbol{s}')$ with $||(\boldsymbol{r}', \boldsymbol{s}') - (\boldsymbol{r}, \boldsymbol{s})|| \leq \epsilon$ satisfies the non-degeneracy condition.

$$\begin{aligned} &P\left[(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) \text{ satisfy the non-degeneracy condition}\right] \\ &\geq P\left[||(\boldsymbol{r}', \boldsymbol{s}') - (\boldsymbol{r}, \boldsymbol{s})|| \leq \epsilon\right] \to 1. \end{aligned}$$

□

2.7 Bootstrap

In this section we discuss the bootstrap for the Wasserstein distance under the null hypothesis $\boldsymbol{r} = \boldsymbol{s}$. In addressing the usual measurability issues that arise in the formulation of consistency for the bootstrap, we follow Van der Vaart and Wellner (1996). We denote by $\hat{\boldsymbol{r}}_n^*$ and $\hat{\boldsymbol{s}}_m^*$ some bootstrapped versions of $\hat{\boldsymbol{r}}_n$ and $\hat{\boldsymbol{s}}_m$. More precisely, let $\hat{\boldsymbol{r}}_n^*$ a measurable function of $X_1, \ldots, X_n$ and random weights $W_1, \ldots, W_n$, independent of the data and analogously for $\hat{\boldsymbol{s}}_m^*$. This setting is general enough to include many common bootstrapping schemes. We say that, with the assumptions and notation of Theorem 1, the bootstrap is consistent if the limiting distribution of

$$\rho_{n,m}\left\{(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) - (\boldsymbol{r}, \boldsymbol{s})\right\} \Rightarrow (\sqrt{\lambda}\boldsymbol{G}, \sqrt{1-\lambda}\boldsymbol{H})$$

is consistently estimated by the law of

$$\rho_{n,m}\left\{(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - (\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\}.$$

To make this precise, we define for $A \subset \mathbb{R}^d$, with $d \in \mathbb{N}$, the set of bounded Lipschitz-1 functions

$$\mathrm{BL}_1(A) = \left\{ f : A \to \mathbb{R} : \sup_{x \in A} |f(x)| \leq 1, \quad |f(x_1) - f(x_2)| \leq ||x_1 - x_2|| \right\},$$

where $||\cdot||$ is the Euclidean norm. We say that the bootstrap versions $(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*)$ are consistent if

(2.23)
$$\sup_{f \in \mathrm{BL}_1(\mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}})} |E\left[f(\rho_{n,m}\left\{(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - (\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\}) | X_1, \ldots, X_n, Y_1, \ldots, Y_m\right] - E\left[f((\sqrt{\lambda}\boldsymbol{G}, \sqrt{1-\lambda}\boldsymbol{H}))\right]|$$

converges to zero in probability.

Bootstrap for directionally differentiable functions The most straightforward way to bootstrap $W_p^p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)$ is to simply plug-in $\hat{\boldsymbol{r}}_n^*$ and $\hat{\boldsymbol{s}}_m^*$. That is, trying to approximate the limiting distribution of $\rho_{n,m} W_p^p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)$ by the law of

$$\rho_{n,m}\left\{W_p^p(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - W_p^p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{r}}_m)\right\} \tag{2.24}$$

conditional on the data. While for functions that are Hadamard differentiable this approach yields a consistent bootstrap (e.g. Gill et al. (1989); Van der Vaart and Wellner (1996)), it has been pointed out by Dümbgen (1993) and more recently by Fang and Santos (2014) that this is in general not true for functions that are only directionally Hadamard differentiable. In particular the plug-in approach fails for the Wasserstein distance.

For the Wasserstein distance there are two alternatives. First, Dümbgen (1993) already pointed out that re-sampling fewer than n (or m, respectively) observations yield a consistent bootstrap. Second, Fang and Santos (2014) propose to plug-in $\rho_{n,m}\left\{(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - (\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\}$ into the derivative of the function.

Recall from Section 2.3 that

$$\phi_p : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}, \quad \phi_p(\boldsymbol{h}_1, \boldsymbol{h}_2) = \max_{\boldsymbol{u} \in \Phi_p^*} \langle \boldsymbol{u}, \boldsymbol{h}_2 - \boldsymbol{h}_1 \rangle \tag{2.25}$$

is the directional Hadamard derivative of $(\boldsymbol{r}, \boldsymbol{s}) \mapsto W_p^p(\boldsymbol{r}, \boldsymbol{s})$ at $\boldsymbol{r} = \boldsymbol{s}$. With this notation, the following Theorem summarizes the implications of the results of Dümbgen (1993) and Fang and Santos (2014) for the Wasserstein distance.

Theorem 9 (Prop. 2 of Dümbgen (1993) and Thms. 3.2 and 3.3 of Fang and Santos (2014))**.** *Under the assumptions of Theorem 1 let $\hat{\boldsymbol{r}}_n^*$ and $\hat{\boldsymbol{s}}_m^*$ be consistent bootstrap versions of $\hat{\boldsymbol{r}}_n$ and $\hat{\boldsymbol{s}}_m$, that is,* (2.23) *converges to zero in probability. Then,*

1. *the plug-in bootstrap* (2.24) *is not consistent, that is,*

$$\sup_{f \in \mathrm{BL}_1(\mathbb{R})} E\left[f(\rho_{n,m}\left\{W_p^p(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - W_p^p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\}) | X_1, \dots X_n, Y_1, \dots, Y_m\right] - E[f(\rho_{n,m} W_p^p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m))]$$

   does not converge to zero in probability.

2. *Under the null hypothesis $\boldsymbol{r} = \boldsymbol{s}$, the derivative plug-in*

$$\phi_p(\rho_{n,m}\left\{(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - (\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\}) \tag{2.26}$$

   is consistent, that is

$$\sup_{f \in \mathrm{BL}_1(\mathbb{R})} E\left[f(\phi_p(\rho_{n,m}\left\{(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - (\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\}))| X_1, \dots, X_n, Y_1, \dots, Y_m\right] - E\left[f\left(\rho_{n,m} W_p^p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right)\right]$$

   converges to zero in probability.

2.8 An alternative representation of the limiting distribution

We give a second representation of the limiting distribution under the alternative $\boldsymbol{r} \neq \boldsymbol{s}$. The random part of the limiting distribution (2.6) is the linear program

$$\max_{(\boldsymbol{u},\boldsymbol{v})\in\Phi_p^*(\boldsymbol{r},\boldsymbol{s})} \sqrt{\lambda}\langle \boldsymbol{G}, \boldsymbol{u}\rangle + \sqrt{1-\lambda}\langle \boldsymbol{H}, \boldsymbol{v}\rangle.$$

With the representation (2.1) of $\Phi_p^*(\boldsymbol{r}, \boldsymbol{s})$ we obtain the dual linear program

$$\begin{aligned} \min \quad & zW_p^p(\boldsymbol{r}, \boldsymbol{s}) + \sum_{x,x'\in\mathcal{X}} w_{x,x'} d^p(x, x') \\ \text{s.t.} \quad & \boldsymbol{w} \geq 0, z \in \mathbb{R} \\ & \sum_{x'\in\mathcal{X}} w_{x,x'} + zr_x = G_x \\ & \sum_{x\in\mathcal{X}} w_{x,x'} + zs_x = H_x \end{aligned}$$

Note that the constraints can only be satisfied if both $\sqrt{\lambda}\boldsymbol{G} - z\boldsymbol{r}$ and $\sqrt{1-\lambda}\boldsymbol{H} - z\boldsymbol{s}$ have only non-negative entries and $z \leq 0$. In this case the second term in the objective function is clearly minimized by $-z\boldsymbol{w}^*$, with $\boldsymbol{w}^*$ an optimal transport plan between these two measures $\boldsymbol{r} - \sqrt{\lambda}\boldsymbol{G}/z$ and $\boldsymbol{s} - \sqrt{1-\lambda}\boldsymbol{H}/z$ and the second term of the objective function is equal to $-zW_p^p(\boldsymbol{r} - \sqrt{\lambda}\boldsymbol{G}/z, \boldsymbol{s} - \sqrt{1-\lambda}\boldsymbol{H}/z)$.

To write this more compactly let us slightly extend our notation. For $\boldsymbol{r}, \boldsymbol{s} \in \mathbb{R}^{\mathcal{X}}$ with $\sum_x r_x = \sum_x s_x = 1$ let

$$\tilde{W}_p^p(\boldsymbol{r}, \boldsymbol{s}) = \begin{cases} W_p^p(\boldsymbol{r}, \boldsymbol{s}) & \text{if } \boldsymbol{r}, \boldsymbol{s} \geq 0; \\ \infty & \text{else.} \end{cases}$$

With this we can thus write the random variable in the limiting distribution

(2.6) as the one-dimensional non-linear optimization problem

(2.27)
$$\frac{1}{p} W_p^{1-p}(\boldsymbol{r}, \boldsymbol{s}) \min_{z \geq 0} z \left\{ \tilde{W}_p^p(\boldsymbol{r} + \sqrt{\lambda}\boldsymbol{G}/z, \boldsymbol{s} + \sqrt{1-\lambda}\boldsymbol{H}/z) - W_p^p(\boldsymbol{r}, \boldsymbol{s}) \right\}.$$

2.9 Simulations and applications

The following numerical experiments were performed using R (R Core Team, 2016). All computations of Wasserstein distances and optimal transport plans as well as their visualizations were performed with the R-package `transport` (Schuhmacher et al., 2014; Gottschlich and Schuhmacher, 2014). The code used for the computation of the limiting distributions is available as an R-package `otinference` (Sommerfeld, 2017).

2.9.1 Speed of convergence

We investigate the speed of convergence to the limiting distribution in Theorem 1 in the one-sample case under the null hypothesis. To this end, we consider as ground space $\mathcal{X}$ a regular two-dimensional $L \times L$ grid with the euclidean distance as the metric d and $L = 3, 5, 10$. We generate five random measures $\boldsymbol{r}$ on $\mathcal{X}$ as realizations of a Dirichlet random variable with concentration parameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha) \in \mathbb{R}^{L \times L}$ for $\alpha = 1, 5, 10$. Note, that $\alpha = 1$ corresponds to a uniform distribution on the probability simplex. For each measure, we generate $20,000$ realizations of $n^{1/2p} W_p(\hat{\boldsymbol{r}}_n, \boldsymbol{r})$ with $n\hat{\boldsymbol{r}}_n \sim \text{Multinom}(\boldsymbol{r})$ for $n = 10, 1000, 1000$ and of the theoretical limiting distribution given in Theorem 1. The Kolmogorov-Smirnov distance (that is, the maximum absolute difference between their cdfs) between these two samples (averaged over the five measures) is shown in Figure 2.1. The experiment shows that the limiting distribution is a good approximation of the finite sample version even for small sample sizes. For the considered parameters the size of the ground space $N = L^2$ seems to slow the convergence only marginally. Similarly, the underlying measure seems to have no sizeable effect on the convergence speed as the dependence on the concentration parameter α demonstrates.

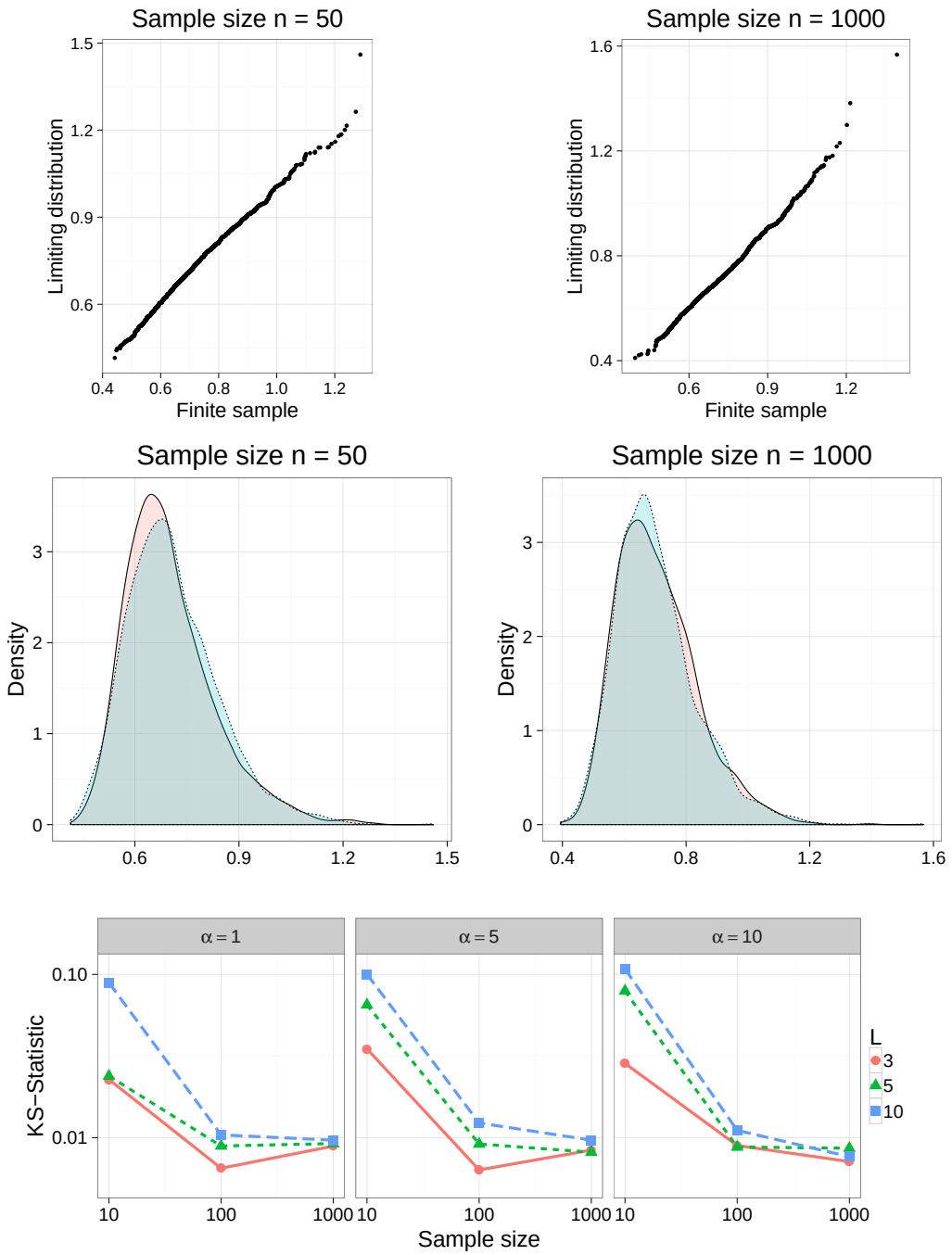

Figure 2.1: Comparison of the finite sample distribution and the theoretical limiting distribution on a regular grid of length L for different sample sizes. The two top rows show Q-Q-plots and kernel density estimates (bandwidth: Silverman's rule of thumb (Silverman, 1986), solid line: finite sample, dotted line: limiting distribution) for $L = 10$. Last row shows the KS statistic between the two distributions as a function of the sample size for different L and for different concentration parameters α.

2.9.2 Testing the null: real and synthetic fingerprints

The generation and recognition of synthetic fingerprints is a topic of great interest in forensic science and current state-of-the-art methods (Cappelli et al., 2000) produce synthetic fingerprints that even human experts fail to recognize as such (Maltoni et al., 2009, p. 292ff). Recently, Gottschlich and Huckemann (2014) presented a method using the Wasserstein distance that is able to distinguish synthetic from real fingerprints with high accuracy. Their method is probabilistic in nature, since it is based on a hypothesized unknown distribution of certain features of the fingerprint. We use our distributional limits to assess the statistical significance of the differences.

Minutiae histograms The basis for the comparison of fingerprints are so called minutiae which are key qualities in biometric identification based on fingerprints (Jain, 2007). They are certain characteristic features such as bifurcations of the line patterns of the fingerprint. Each of the minutiae have a location in the fingerprint and a direction such that it can be characterized by two real numbers and an angle. Figure 2.3 shows a real and a synthetic fingerprint with their minutiae.

The recognition method of Gottschlich and Huckemann (2014) considers pairs of minutiae and records their distance and the difference between their angles. Based on these two values each minutiae pair is put in one of 100 bins arranged in a regular grid (10 directional by 10 distance bins) to obtain a so called minutiae histogram (MH). Based on the bin-wise mean of MHs for several fingerprints to construct a typical MH, they found that the proximity in Wasserstein distance to these references is a good classifier for distinguishing real and synthetic fingerprints.

In order to assess the statistical significance of the difference in minutiae pair distributions, we consider fingerprints from the databases 1 and 4 of the Fingerprint Verification Competition of 2002 (Maio et al., 2002), containing 110 real and synthetic fingerprints, respectively. From each database the minutiae were obtained by automatic procedure using a commercial off-the-shelf program. For each fingerprint we chose disjoint minutiae pairs at

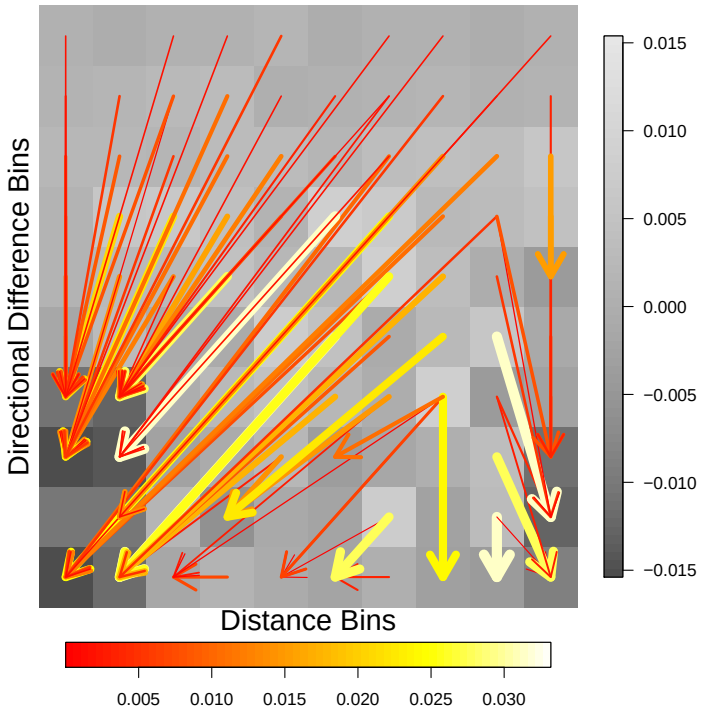

Figure 2.2: The optimal transport plan between the MHs of real and fake fingerprints. The grey values indicate the magnitude of the difference of the two MHs. The arrows show the transport. The amount of mass transported is encoded in the color and thickness of the arrows.

random to avoid the issue of pairs being dependent yielding a total of 1917 and 1437 minutiae pairs from real and synthetic fingerprints, respectively.

While two-sample tests for univariate data are abundant and well studied there are no multivariate methods that could be considered standard in this setting. Therefore, we report on the findings of several tests from the literature for comparison with the Wasserstein based method from (2.5). We tested the null hypothesis of the underlying distributions being equal for the un-centered, the centered and the centered and scaled (to variance 1) data to assess effects beyond first moments using the following methods: 1) comparing the empirical Wasserstein distance W_1 after binning on a regular 10×10 grid with the limiting distribution from Theorem 1; 2) a permutation test; 3) the crossmatch test proposed by (Rosenbaum, 2005) and 4) the kernel based test (Anderson et al., 1994) implemented in the R package ks.

Table 2.1: Results of different two-sample tests for difference in the distribution of MHs of real and fake fingerprints.

	Wasserstein	Crossmatch	Permutation	KDE
Raw	0.00E+00	2.99E-01	1.00E-03	1.12E-08
Centered	4.00E-04	4.48E-05	1.00E-03	2.60E-21
Centered & Scaled	2.54E-02	1.01E-02	1.71E-01	1.79E-14

Table 2.1 shows the resulting empirical distributions on a 10×10 grid and the p-values for the different tests. The differences are highly significant according to all tests, except the permutation test for the centered and scaled data. In this particular example at least, the Wasserstein based test seems to be able to pick up differences in distributions (in the first moment and beyond) at least as good as current state-of-the-art methods.

In addition to testing, the Wasserstein method provides us with an optimal transport plan, transforming one measure into the other. For the minutiae histograms under consideration this is illustrated in Figure 2.2. This transport plan gives information beyond a simple test for equality as it highlights structural changes in the distribution. In this specific application it reveals how in the minutiae histogram of synthetic fingerprints compared to the one of real fingerprints mass has been shifted from large and intermediate directional differences to smaller ones. In particular to small and large distances, and only to a lesser extent to intermediate distances. In conclusion one may say that synthetic fingerprints show smaller differences in the directions of minutiae and stronger clustering of minutiae distances around small and large values. Insight of this sort may lead to improved generation or detection of synthetic fingerprints.

2.9.3 Asymptotic under the alternative: metagenomics

Metagenomics studies microbial communities by analyzing genetic material in an environmental sample such as a stool sample of a human. High-throughput sequencing techniques no longer require cultivated cloned microbial cultures to perform sequencing. Instead, a sample with potentially many different species can be analyzed directly and the abundance of each

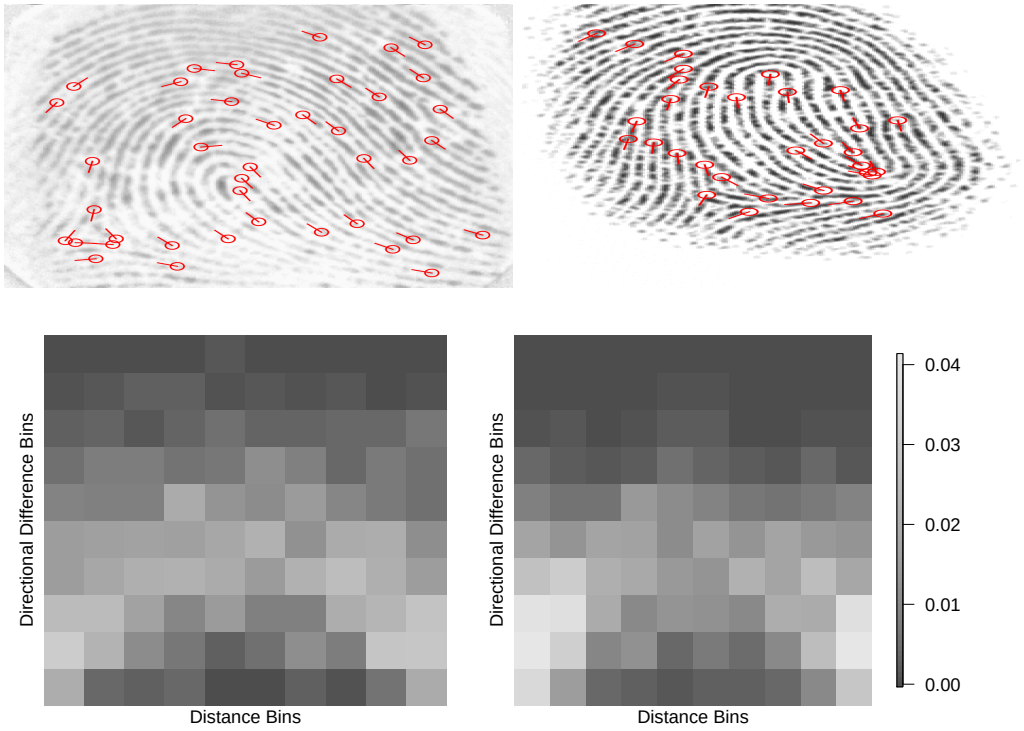

Figure 2.3: Top row: Minutiae of a real (left) and a synthetic (right) fingerprint. Bottom row: Minutiae histograms of real and synthetic fingerprints.

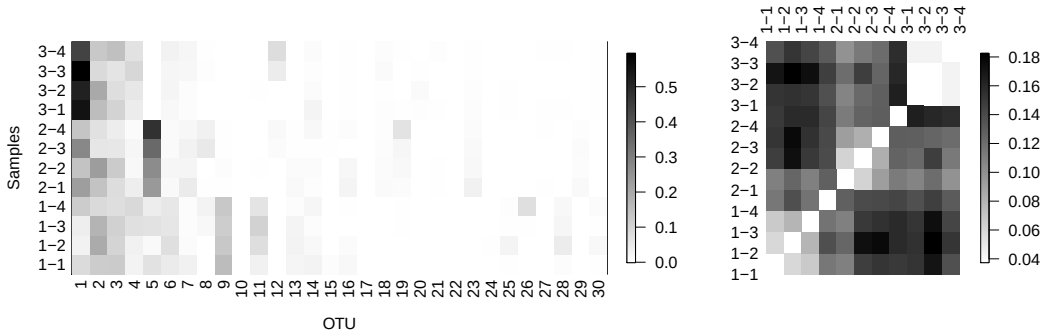

Figure 2.4: Relative abundances of the 30 first OTUs in the 12 samples (left) and Wasserstein distances of the microbial communities (right). Here, ij is the j-th sample of the i-th person.

species in the sample can be recovered. The applications of this technique are countless and constantly growing. In particular, the composition of microbial communities in the human gut has been associated with obesity, inflammatory bowel disease and others (Turnbaugh et al., 2007).

The analysis of a sample with high-throughput sequencing techniques yields several thousands to many hundreds of thousands sequences. After elaborate pre-processing, these sequences are aligned to a reference database and clustered in *operational taxonomic units* (OTUs). These OTUs can be thought of (albeit omitting some biological detail) as the different species present in the sample. For each OTU this analysis yields the number of sequences associated with it, that is how often this particular OTU was detected in the sample. Further, comparing the genetic sequences associated with an OTU yields a biologically meaningful measure of similarity between OTUs - and hence a distance. A metagenomic sample can therefore be regarded as a sample in a discrete metric space with OTUs being the points of the space. Comparing such samples representing microbial communities is of great interest (Kuczynski et al., 2010). The Wasserstein distance has been recognized to provide valuable insight and to facilitate tests for equality of two communities (Evans and Matsen, 2012). This previous application however, relies on a phylogenetic tree that is build on the OTUs and the distance is then measured in the tree. This additional pre-processing step involves many parameter choices and is unnecessary with our method.

A further drawback of the method of Evans and Matsen (2012) is that it only allows for testing the null hypothesis of two communities being equal. In practice, one frequently finds that natural variation is so high that even two samples from the same source taken at different times will be recognized as different. This raises the question whether variation within samples from the same source is smaller than the difference to samples of another source. Statistically speaking we are looking for confidence sets for differences which are assumed to be different from zero. This requires asymptotics under the alternative $\boldsymbol{r} \neq \boldsymbol{s}$, which is provided by Theorem 1.

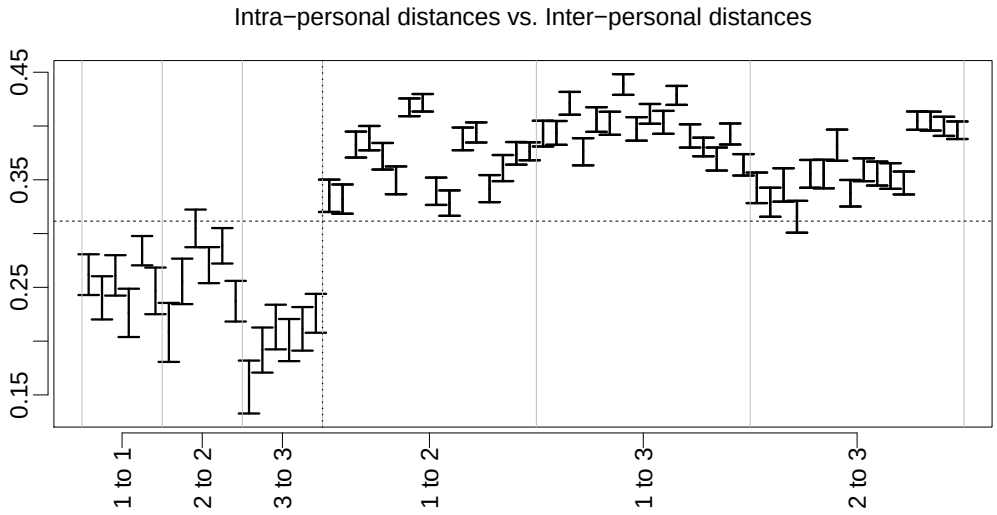

Figure 2.5: Display of 95% confidence intervals of Wasserstein distances of microbial communities. The horizontal axis shows which person pair the distances belong to (separated by gray vertical lines). The dotted vertical line separates intra- (left) from inter- (right) -personal distances.

Data analysis We consider part of the data of Costello et al. (2009). Four stool samples were taken from each of three persons at different times. We used the preparation of this data by P. Schloss available at `https://www.mothur.org/w/images/d/d8/CostelloData.zip`. The reads were pre-processed with the program `mothur` (Schloss et al., 2009) using the procedure outlined in Schloss et al. (2011) and Schloss (2015). The relative abundances of the 30 most frequent OTUs and the Wasserstein-2 distances of the microbial communities are shown in Figure 2.4. In this and all other figures we use $i - j$ to denote sample j of person i. Note that it is typical for this data that most of the mass is concentrated on a few OTUs.

The Wasserstein-2 distances for all 66 pairs and their 99% confidence intervals were computed using the asymptotic distribution in Theorem 1. The results are shown in Figure 2.5. The entire analysis took less than a minute on a standard laptop. The confidence intervals show that intra-personal distances are in fact significantly smaller than inter-personal distances.

2.10 Discussion

We discuss limitations, possible extensions of the presented work and promising directions for future research.

Beyond finite spaces I: rates in the finite and the continuous setting ($d = 1$) The scaling rate in Theorem 1 depends solely on p and is completely independent of the underlying space $\mathcal{X}$. This contrasts known bounds on the rate of convergence in the continuous case (see references in the Introduction), which exhibit a strong dependence on the dimension of the space and the moments of the distribution.

Under the null hypothesis (that is, the two underlying population measures are equal) and when $\mathcal{X} = \mathbb{R}$ and $p = 2$, the scaling rate for a continuous distribution is known to be $n^{1/2}$, at least under additional tail conditions (see e.g. Del Barrio et al. (2005)). This means that in this case the scaling rate for a discrete distribution is slower (namely $n^{1/4}$). Under the alternative (different population measures) the scaling rate is $n^{1/2}$ and coincide in the discrete and the continuous case (see Munk and Czado (1998)).

Beyond finite spaces II: higher dimensions ($d \geq 2$) For a continuous measure μ the Wasserstein distance is the solution of an infinite-dimensional optimization problem. Although differentiability results also exist for such problems (e.g. Shapiro (1992)), there are strong indications that the argument presented here cannot carry over to the this case for $d \geq 2$. This is most easily seen from the classical results of Ajtai et al. (1984). We consider the uniform distribution on the unit square. For two samples of size n independently drawn from this distribution, Ajtai et al. (1984) showed that there exist constants C_1, C_2 such that the 1-Wasserstein distance $\hat{W}_1^{(n)}$ between them satisfies

$$C_1 n^{-1/2} (\log n)^{1/2} \leq \hat{W}_1^{(n)} \leq C_2 n^{-1/2} (\log n)^{1/2}$$

with probability $1 - o(1)$. Hence, for $c_n \hat{W}_1^{(n)}$ to have a non-degenerate limit, we need $c_n = \sqrt{n/\log n}$. However, a common property of all delta-methods

is that they preserve the rate of convergence, which is not satisfied here.

Transport distances on trees Complementing our Theorem 5 a further result on transport distances on trees was proven by Evans and Matsen (2012) in the context of phylogenetic trees for the comparison of metagenomic samples (see also our application in Section 2.9). They point out that the Wasserstein-1 distance on trees is equal to the so-called *weighted unifrac distance* which is very popular in genetics. Inspired by this distance they give a formal generalization mimicking a cost exponent $p > 1$ and consider its asymptotic behavior. However, as they remark, these generalized expressions are no longer related (beyond a formal resemblance) to Wasserstein distances with cost exponent $p > 1$. Comparing the performance of their ad-hoc metric and the true Wasserstein distance on trees that is under consideration here is an interesting topic for further research.

Bootstrap We showed that while the naive n-out-of-n bootstrap fails for the Wasserstein distance (Section 2.7), the m-out-of-n bootstrap is consistent. An interesting and challenging question is how m should be chosen.

Wasserstein barycenters Barycenters in the Wasserstein space (Agueh and Carlier, 2011) have recently received much attention (Cuturi and Doucet, 2014; Del Barrio et al., 2015). We expect that the techniques developed here can be of use in providing a rigorous statistical theory (e.g. distributional limits). The same applies to geodesic principal component analysis in the Wasserstein space (Bigot et al., 2013; Seguy and Cuturi, 2015).

Alternative cost matrices and transport distances Theorem 1 holds in very large generality for arbitrary cost matrices, including in particular the case of a cost matrix derived from a metric but using a cost exponent $p < 1$.

Beyond this obvious modification it seems worthwhile to extend the methodology of directional differentiability in conjunction with a delta-method to other functionals related to optimal transport, e.g. entropically regularized

(Cuturi, 2013) or sliced Wasserstein distances (Bonneel et al., 2015). This would require a careful investigation of the analytical properties of these quantities similar to classical results for the Wasserstein distance.

2.11 Proofs

2.11.1 Proof of Theorem 1

a) With the notation introduced in Theorem 1, $n\hat{\boldsymbol{r}}_n$ is a sample of size n from a multinomial distribution with probabilities $\boldsymbol{r}$. Therefore, $\sqrt{n}(\hat{\boldsymbol{r}}_n - \boldsymbol{r}) \Rightarrow \boldsymbol{G}$ as $n \to \infty$ (Wasserman, 2011, Thm. 14.6). The Hadamard derivative of the map $(\boldsymbol{r}, \boldsymbol{s}) \mapsto W_p^p(\boldsymbol{r}, \boldsymbol{s})$ as given in Theorem 4 can now be used in the delta-method from Theorem 3. Together with the representation (2.10) of the set of dual solutions $\Phi_p^*(\boldsymbol{r}, \boldsymbol{s})$, this yields

$$\sqrt{n} W_p^p(\hat{\boldsymbol{r}}_n, \boldsymbol{r}) \Rightarrow \max_{(\boldsymbol{u},\boldsymbol{v}) \in \Phi_p^*(\boldsymbol{r},\boldsymbol{r})} -\langle \boldsymbol{u}, \boldsymbol{G} \rangle \overset{D}{\sim} \max_{\boldsymbol{u} \in \Phi_p^*} \langle \boldsymbol{u}, \boldsymbol{G} \rangle.$$

Here and in the following $Z_1 \overset{D}{\sim} Z_2$ means the distributional equality of the random variables Z_1 and Z_2. Applying to this the Continuous Mapping Theorem with the map $t \mapsto t^{1/p}$ gives the assertion.

b) Consider the map $(\boldsymbol{r}, \boldsymbol{s}) \mapsto W_p(\boldsymbol{r}, \boldsymbol{s}) = (W_p^p(\boldsymbol{r}, \boldsymbol{s}))^{1/p}$. By Theorem 4 and the chain rule for Hadamard directional derivatives (Shapiro, 1990, Prop. 3.6), the Hadamard derivative of this map at $(\boldsymbol{r}, \boldsymbol{s})$ is given by

$$(\boldsymbol{h}_1, \boldsymbol{h}_2) \mapsto p^{-1} W_p^{1-p}(\boldsymbol{r}, \boldsymbol{s}) \left\{ \max_{(\boldsymbol{u},\boldsymbol{v}) \in \Phi_p^*(\boldsymbol{r},\boldsymbol{s})} -(\langle \boldsymbol{u}, \boldsymbol{h}_1 \rangle + \langle \boldsymbol{v}, \boldsymbol{h}_2 \rangle) \right\}. \tag{2.28}$$

An application of the delta-method of Theorem 3 concludes this part.

c) and d). Note that under the assumptions of the Theorem

$$\sqrt{\frac{nm}{n+m}} \left((\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) - (\boldsymbol{r}, \boldsymbol{s}) \right) \Rightarrow (\sqrt{\lambda}\boldsymbol{G}, \sqrt{1-\lambda}\boldsymbol{H}). \tag{2.29}$$

Part d) follows with the delta-method from (2.28) and (2.29).

For part c) we use, as we did for a), the derivative given in Theorem 4 and the Continuous Mapping Theorem. The limit distribution is

$$\left\{ \max_{(\boldsymbol{u},\boldsymbol{v})\in\Phi_p^*(\boldsymbol{r},\boldsymbol{s})} (\sqrt{\lambda}\langle \boldsymbol{G},\boldsymbol{u}\rangle + \sqrt{1-\lambda}\langle \boldsymbol{H},\boldsymbol{v}\rangle) \right\}^{1/p}.$$

Note that if $\boldsymbol{r} = \boldsymbol{s}$ we have $(\boldsymbol{u},\boldsymbol{v}) \in \Phi_p^*(\boldsymbol{r},\boldsymbol{s})$ if and only if $\boldsymbol{u} \in \Phi_p^*$ and $\boldsymbol{v} = -\boldsymbol{u}$, by (2.10) and (2.1). Hence, with $\boldsymbol{G} \overset{D}{\sim} \boldsymbol{H}$ we conclude

$$\begin{aligned}
\max_{(\boldsymbol{u},\boldsymbol{v})\in\Phi_p^*(\boldsymbol{r},\boldsymbol{s})} (\sqrt{\lambda}\langle \boldsymbol{G},\boldsymbol{u}\rangle + \sqrt{1-\lambda}\langle \boldsymbol{H},\boldsymbol{v}\rangle) &\overset{D}{\sim} \max_{\boldsymbol{u}\in\Phi_p^*} (\sqrt{\lambda}\langle \boldsymbol{G},\boldsymbol{u}\rangle - \sqrt{1-\lambda}\langle \boldsymbol{H},\boldsymbol{u}\rangle) \\
&\overset{D}{\sim} \max_{\boldsymbol{u}\in\Phi_p^*} \sqrt{\lambda + (1-\lambda)}\langle \boldsymbol{G},\boldsymbol{u}\rangle \\
&= \max_{\boldsymbol{u}\in\Phi_p^*} \langle \boldsymbol{G},\boldsymbol{u}\rangle.
\end{aligned}$$

2.11.2 Proof of Theorem 4

By (Gal et al., 1997, Ch. 3, Thm. 3.1) the function $(\boldsymbol{r},\boldsymbol{s}) \mapsto W_p^p(\boldsymbol{r},\boldsymbol{s})$ is directionally differentiable with derivative (2.9) in the sense of Gâteaux, that is, the limit (2.7) exists for a fixed $\boldsymbol{h}$ and not a sequence $\boldsymbol{h}_n \to \boldsymbol{h}$ (see e.g. Shapiro (1990)). To see that this is also a directional derivative in the Hadamard sense (2.7) it suffices (Shapiro, 1990, Prop. 3.5) to show that $(\boldsymbol{r},\boldsymbol{s}) \mapsto W_p^p(\boldsymbol{r},\boldsymbol{s})$ is locally Lipschitz. That is, we need to show that for $\boldsymbol{r},\boldsymbol{r}',\boldsymbol{s},\boldsymbol{s}' \in \mathcal{P}_{\mathcal{X}}$

$$|W_p^p(\boldsymbol{r},\boldsymbol{s}) - W_p^p(\boldsymbol{r}',\boldsymbol{s}')| \leq C||(\boldsymbol{r},\boldsymbol{s}) - (\boldsymbol{r}',\boldsymbol{s}')||,$$

for some constant $C > 0$ and some (and hence all) norm $||\cdot||$ on $\mathbb{R}^N \times \mathbb{R}^N$. Exploiting symmetry, it suffices to show that

$$W_p^p(\boldsymbol{r},\boldsymbol{s}) - W_p^p(\boldsymbol{r},\boldsymbol{s}') \leq C||\boldsymbol{s} - \boldsymbol{s}'||$$

for some constant $C > 0$ and some norm $||\cdot||$. To this end, we employ an argument similar to that used to prove the triangle inequality for the Wasserstein distance (see e.g. (Villani, 2008, p. 94)). Indeed, by the gluing

Lemma (Villani, 2008, Ch. 1) there exist random variables X_1, X_2, X_3 with marginal distributions $\boldsymbol{r}, \boldsymbol{s}$ and $\boldsymbol{s}'$, respectively, such that $E[d^p(X_1, X_3)] = W_p^p(\boldsymbol{r}, \boldsymbol{s}')$ and $E[d(X_2, X_3)] = W_1(\boldsymbol{s}, \boldsymbol{s}')$. Then, since (X_1, X_2) has marginals $\boldsymbol{r}$ and $\boldsymbol{s}$, we have

$$\begin{aligned} W_p^p(\boldsymbol{r}, \boldsymbol{s}) - W_p^p(\boldsymbol{r}, \boldsymbol{s}') &\leq E\left[d^p(X_1, X_2) - d^p(X_1, X_3)\right] \\ &\leq p \operatorname{diam}(\mathcal{X})^{p-1} E\left[|d(X_1, X_2) - d(X_1, X_3)|\right] \\ &\leq p \operatorname{diam}(\mathcal{X})^{p-1} E\left[d(X_2, X_3)\right] = p \operatorname{diam}(\mathcal{X})^{p-1} W_1(\boldsymbol{s}, \boldsymbol{s}') \\ &\leq p \operatorname{diam}(\mathcal{X})^{p} ||\boldsymbol{s} - \boldsymbol{s}'||_1, \end{aligned}$$

where the last inequality follows from (Villani, 2008, Thm. 6.15). This completes the proof.

2.11.3 Proof of Theorem 5

Simplify the set of dual solutions Φ_p^* As a first step, we rewrite the set of dual solutions Φ_p^* given in (2.1) in our tree notation as

$$\Phi_p^* = \left\{\boldsymbol{u} \in \mathbb{R}^{\mathcal{X}} : u_x - u_{x'} \leq d_{\mathcal{T}}(x, x')^p, \quad x, x' \in \mathcal{X}\right\}. \tag{2.30}$$

The key observation is that in the condition $u_x - u_{x'} \leq d_{\mathcal{T}}(x, x')^p$ we do not need to consider all pairs of vertices $x, x' \in \mathcal{X}$, but only those which are joined by an edge. To see this, assume that only the latter condition holds. Let $x, x' \in \mathcal{X}$ arbitrary and $x = x_1, \ldots, x_l = x'$ the sequence of vertices defining the unique path joining x and x', such that $(x_j, x_{j+1}) \in E$ for $j = 1, \ldots, n-1$. Then

$$u_x - u_{x'} = \sum_{j=1}^{n-1} (u_{x_j} - u_{x_{j+1}}) \leq \sum_{j=1}^{n-1} d_{\mathcal{T}}(x_j, x_{j+1})^p \leq d_{\mathcal{T}}(x, x')^p,$$

such that the condition is satisfied for all $x, x' \in \mathcal{X}$. Noting that if two vertices are joined by an edge than one has to be the parent of the other, we

can write the set of dual solutions as

$$\Phi_p^* = \left\{ \boldsymbol{u} \in \mathbb{R}^{\mathcal{X}} : |u_x - u_{\mathrm{par}(x)}| \leq d_{\mathcal{T}}(x, \mathrm{par}(x))^p, \quad x \in \mathcal{X} \right\}. \tag{2.31}$$

Rewrite the target function We define linear operators $S_{\mathcal{T}}, D_{\mathcal{T}} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ by

$$(D_{\mathcal{T}} v)_x = \begin{cases} v_x - v_{\mathrm{par}(x)} & x \neq \mathrm{root}(\mathcal{T}) \\ v_{\mathrm{root}(\mathcal{T})} & x = \mathrm{root}(\mathcal{T}). \end{cases}, \quad (S_{\mathcal{T}} u)_x = \sum_{x' \in \mathrm{children}(x)} u_{x'}.$$

Lemma 1. *For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{\mathcal{X}}$ we have $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \langle S_{\mathcal{T}} \boldsymbol{u}, D_{\mathcal{T}} \boldsymbol{v} \rangle$.*

Proof. We compute

$$\begin{aligned} \langle S_{\mathcal{T}} \boldsymbol{u}, D_{\mathcal{T}} \boldsymbol{v} \rangle &= \sum_{x \in \mathcal{X}} (S_{\mathcal{T}} \boldsymbol{u})_x (D_{\mathcal{T}} \boldsymbol{v})_x \\ &= \sum_{x \in \mathcal{X} \setminus \{\mathrm{root}(\mathcal{T})\}} \sum_{x' \in \mathrm{children}(x)} (v_x - v_{\mathrm{par}(x)}) u_{x'} \\ &\quad + \sum_{x' \in \mathrm{children}(\mathrm{root}(\mathcal{T}))} v_{\mathrm{root}(\mathcal{T})} u_{x'} \\ &= \sum_{x \in \mathcal{X}} \sum_{x' \in \mathrm{children}(x)} v_x u_{x'} \\ &\quad - \sum_{x \in \mathcal{X} \setminus \{\mathrm{root}(\mathcal{T})\}} \sum_{x' \in \mathrm{children}(x)} v_{\mathrm{par}(x)} u_{x'} \\ &= \sum_{x \in \mathcal{X}} u_x v_x, \end{aligned}$$

which proves the Lemma. To see how the last line follows let $\mathrm{children}^1(x)$ be the set of immediate predecessors of x, that is children of x that are connected to x by an edge. Then we can write the second term in the second to last line above as

$$\begin{aligned} \sum_{x \in \mathcal{X} \setminus \{\mathrm{root}(\mathcal{T})\}} \sum_{x' \in \mathrm{children}(x)} v_{\mathrm{par}(x)} u_{x'} &= \sum_{y \in \mathcal{X}} \sum_{x \in \mathrm{children}^1(y)} \sum_{x' \in \mathrm{children}(x)} v_y u_{x'} \\ &= \sum_{y \in \mathcal{X}} \sum_{x' \in \mathrm{children}(y) \setminus \{y\}} v_y u_{x'} \end{aligned}$$

and the claim follows. □

If $\boldsymbol{u} \in \Phi_p^*$, as given in (2.31), we have for $x \neq \text{root}(\mathcal{T})$ that

$$|(D_{\mathcal{T}}\boldsymbol{u})_x| = |u_x - u_{\text{par}(x)}| \leq d_{\mathcal{T}}(x, \text{par}(x))^p.$$

With these two observations and Lemma 1, we get for $\boldsymbol{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ and $\boldsymbol{u} \in \Phi_p^*$ that

$$\langle \boldsymbol{G}, \boldsymbol{u} \rangle = \langle S_{\mathcal{T}}\boldsymbol{G}, D_{\mathcal{T}}\boldsymbol{u} \rangle \leq \sum_{\text{root}(\mathcal{T}) \neq x \in \mathcal{X}} |(S_{\mathcal{T}}\boldsymbol{G})_x| d_{\mathcal{T}}(x, \text{par}(x))^p. \tag{2.32}$$

Therefore, $\max_{\boldsymbol{u} \in \Phi_p^*} \langle \boldsymbol{G}, \boldsymbol{u} \rangle$ is bounded by $\sum_{\text{root}(\mathcal{T}) \neq x \in \mathcal{X}} |(S_{\mathcal{T}}\boldsymbol{G})_x| d_{\mathcal{T}}(x, \text{par}(x))^p$. Since $D_{\mathcal{T}}$ is an isomorphism, we can define a vector $\boldsymbol{v} \in \mathbb{R}^{\mathcal{X}}$ by

$$(D_{\mathcal{T}}\boldsymbol{v})_x = \text{sgn}\left((S_{\mathcal{T}}\boldsymbol{G})_x\right) d_{\mathcal{T}}(x, \text{par}(x))^p.$$

From (2.31) we see that $\boldsymbol{v} \in \Phi_p^*$ and Lemma 1 shows that $\langle \boldsymbol{G}, \boldsymbol{v} \rangle$ attains the upper bound in (2.32). This concludes the proof.

2.11.4 Proof of Corollary 1

In order to use Theorem 5 we define the tree $\mathcal{T}$ with vertices $\{x_1, \ldots, x_N\}$, edges $E = \{(x_j, x_{j+1}), j = 1, \ldots, N-1\}$ and $\text{root}(\mathcal{T}) = x_N$. Then, if $\boldsymbol{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$, we have that $\{(S_{\mathcal{T}}\boldsymbol{G})_j\}_{j=1,\ldots,N}$ is a Gaussian vector such that for $i \leq j$

$$\begin{aligned}\text{cov}((S_{\mathcal{T}}\boldsymbol{G})_i, (S_{\mathcal{T}}\boldsymbol{G})_j) &= \sum_{\substack{k \leq i \\ l \leq j}} E\left[G_k G_l\right] = \sum_{k \leq i} r_k(1 - r_k) - \sum_{\substack{k \leq i \\ l \leq j \\ k \neq l}} r_k r_l \\ &= \bar{r}_i - \sum_{\substack{k \leq i \\ l \leq i}} r_k r_l - \sum_{\substack{k \leq i \\ i < l \leq j}} r_k r_l = \bar{r}_i - \bar{r}_i^2 - \bar{r}_i(\bar{r}_j - \bar{r}_i)) = \bar{r}_i - \bar{r}_i \bar{r}_j.\end{aligned}$$

Therefore, we have that for a standard Brownian bridge B

$$S_{\mathcal{T}}\boldsymbol{G} \sim (B(\bar{r}_1), \ldots, B(\bar{r}_N)).$$

Together with $d(x_j, \mathrm{par}(x_j)) = (x_{j+1} - x_j)^2$, and (2.12) this proves the Corollary.

Chapter 3

Strategies for inference in large-scale problems

This chapter proposes a strategy to apply the distributional limits presented in the previous chapter for two-sample testing in the case of very large problems, that is, when the number of support points of the measures makes exact computation of the involved quanities computationally infeasible.

The first section shows how thresholding the ground distance yields a lower bound for the Wasserstein distance, while the second section gives a stochastic upper bound for the limiting distribution by using the explicit expression for tree metrics. These results are combined to yield a conservative but fast two-sample test which is applied to microscopy data in the third section.

3.1 Thresholded Wasserstein distance

As outlined in the introduction, a lower bound on $W_p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)$ is enough to test the null hypothesis $\boldsymbol{r} = \boldsymbol{s}$ with pre-specified significance level. To this end, we use an idea of Pele and Werman (2009) who showed that one can obtain such a lower bound by thresholding the ground distance and that this reduces computation time and memory requirements by one polynomial order.

Thresholded ground distance For a thresholding parameter $t \geq 0$ define the thresholded metric

$$d_t(x, x') = \min \{d(x, x'), t\} . \tag{3.1}$$

Then, d_t is again a metric. Let $W^{(t)}(\boldsymbol{r}, \boldsymbol{s})$ be the Wasserstein distance with respect to d_t. Since $d_t(x, x') \leq d(x, x')$ for all $x, x' \in \mathcal{X}$ we have that $W_p^{(t)}(\boldsymbol{r}, \boldsymbol{s}) \leq W_p(\boldsymbol{r}, \boldsymbol{s})$ for all $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_\mathcal{X}$ and all $t \geq 0$.

Computing the thresholded Wasserstein distance The thresholded distance $W_p^{(t)}$ is often much faster to compute than the exact Wasserstein distance. The reason for this is that many of N^2 distances between points in $\mathcal{X}$ have length t. Since the transport problem (1.1) can be written as a network-flow problem (Luenberger and Ye, 2008), we can leverage this fact to redirect all edges with length t through a virtual node and thus dramatically reduce the number of edges. The resulting network-flow problem can be tackled with existing efficient solvers (see e.g. (Bertsekas, 1992); in practice, we achieved the best results with the network solver of the CPLEX (`www.ibm.com/software/commerce/optimization/cplex-optimizer/`)).

For details we refer to the original source Pele and Werman (2009). Among other things, they show that if each point in $\mathcal{X}$ has $\mathcal{O}(1)$ neighbors with distance at most t, the thresholded distance can be computed in $\mathcal{O}(N^2 \log N)$ time with $\mathcal{O}(N)$ memory requirement. This is a considerable reduction compared to the exact distance, which requires $\mathcal{O}(N^3 \log N)$ time and $\mathcal{O}(N^2)$ memory.

In practice this difference proves to be very meaningful as we demonstrate in Section 3.3 where we use the thresholded Wasserstein distance for inference on a large grid.

We remark at this point that it is entirely possible to use d_t as ground distance on $\mathcal{X}$ and Theorem 1 will give the exact limiting distribution also in this case. Since this entails changing the given structure on $\mathcal{X}$ we do not pursue this approach any further in this work.

3.2 Stochastically bounding the limiting distribution

In order to use the distributional limits from Section 2.1 when N is large, we need to compute the limiting distribution. When N is large, however, the limiting distribution in Theorem 1 is a linear program with essentially the same number of constraints and variables as the dual of the Wasserstein problem. Therefore, computing the limiting distribution is essentially as hard as computing the Wasserstein distance itself. This renders a naive Monte-Carlo approach to obtain quantiles infeasible. But we can use the explicit formula for the case of tree metrics to stochastically bound the limiting distribution.

This is based on the following simple observation: Let $\mathcal{T}$ a spanning tree of $\mathcal{X}$ and $d_{\mathcal{T}}$ the tree metric generated by $\mathcal{T}$ and the weights $(x, x') \mapsto d(x, x')$ as described in Section 2.4. Then for any $x, x' \in \mathcal{X}$ we have $d(x, x') \leq d_{\mathcal{T}}(x, x')$. Let $\Phi^*_{p,\mathcal{T}}$ denote the set defined in (2.1) with the metric $d_{\mathcal{T}}$ instead of d. Then $\Phi^*_p \subset \Phi^*_{p,\mathcal{T}}$ and hence

$$\max_{\boldsymbol{u} \in \Phi^*_p} \langle \boldsymbol{v}, \boldsymbol{u} \rangle \leq \max_{\boldsymbol{u} \in \Phi^*_{p,\mathcal{T}}} \langle \boldsymbol{v}, \boldsymbol{u} \rangle$$

for all $\boldsymbol{v} \in \mathbb{R}^N$. In view of formula (2.12) define

$$Z_{\mathcal{T},p}(\boldsymbol{u}) = \left\{ \sum_{x \in \mathcal{X}} |(S_{\mathcal{T}} \boldsymbol{u})_x| d_{\mathcal{T}}(x, \mathrm{par}(x))^p \right\}^{\frac{1}{p}} \tag{3.2}$$

for $\boldsymbol{u} \in \mathbb{R}^N$. It follows that

$$\max_{\boldsymbol{u} \in \Phi^*_p} \langle \boldsymbol{v}, \boldsymbol{u} \rangle \leq Z_{\mathcal{T},p}(\boldsymbol{v}).$$

for all $\boldsymbol{v} \in \mathbb{R}^N$ and this proves the following main result of this section.

Theorem 10. *Let $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_{\mathcal{X}}$ and $\hat{\boldsymbol{r}}_n$ $\hat{\boldsymbol{s}}_m$ be generated by i.i.d. $X_1, \ldots, X_n \sim \boldsymbol{r}$ and $Y_1, \ldots, Y_m \sim \boldsymbol{s}$, respectively. Let further $\mathcal{T}$ be a spanning tree of $\mathcal{X}$. Then, under the null hypothesis $\boldsymbol{r} = \boldsymbol{s}$ we have as n and m approach infinity*

such that $n \wedge m \to \infty$ *and* $n/(n+m) \to \lambda \in [0,1]$ *that*

$$\limsup_{n,m\to\infty} P\left[\left(\frac{nm}{n+m}\right)^{1/2p} W_p^{(t)}(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m) \geq z\right] \leq P\left[Z_{\mathcal{T},p}(\boldsymbol{G}) \geq z\right], \tag{3.3}$$

where $\boldsymbol{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ *with* $\Sigma(\boldsymbol{r})$ *as defined in* (2.2).

In (3.3) the important parameter is the threshold t. While the stochastic bound of the limiting distribution $Z_{\mathcal{T},p}$ is very fast to compute, the thresholded Wasserstein distance $W_p^{(t)}(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)$ is a computational bottleneck. A large threshold t will result in a better approximation of the true Wasserstein distance and hence a higher power of the test but also requires a longer computation time.

Regular Grids

When $\mathcal{X}$ is a regular grid a spanning tree can be constructed from a dyadic partition. Let D be a positive integer, L a power of two and $\mathcal{X}$ the regular grid of L^D points in the unit hypercube $[0,1]^D$. For $0 \leq l \leq l_{\max}$ with

$$l_{\max} = \log_2 L$$

let P_l be the natural partition of $\mathcal{X}$ into 2^{Dl} squares of each $L^D/2^{Dl}$ points. We define $\mathcal{X}'$ by adding to $\mathcal{X}$ all center-points of sets in P_l for $0 \leq l < l_{\max}$. We identify center points of $P_{l_{\max}}$ with the points in $\mathcal{X}$. A tree with vertices $\mathcal{X}'$ can now be build using the inclusion relation of the sets $\{P_l\}_{0\leq l\leq l_{\max}}$ as ancestry relation. More precisely, the leaves of the tree are the points of $\mathcal{X}$ and the parent of the center point of $F \in P_l$ is the center point of the unique set in P_{l-1} that contains F.

If we use the Euclidean metric to define the distance between neighboring vertices we get

$$d_{\mathcal{T}}(x, \mathrm{par}(x)) = \frac{\sqrt{D}2^{-l}}{2},$$

if $x \in P_l$.

A measure $\boldsymbol{r}$ on $\mathcal{X}$ naturally extends to a measure on $\mathcal{X}'$ if we give zero

mass to all inner vertices. We also denote this measure by $\boldsymbol{r}$. Then, if $x \in \mathcal{X}'$ is the center point of the set $F \in P_l$ for some $0 \leq l \leq l_{\max}$, we have that $(S_{\mathcal{T}}\boldsymbol{r})_x = S_F\boldsymbol{r}$ where $S_F\boldsymbol{r} = \sum_{x \in F} r_x$. Therefore, we have in (3.3)

$$Z_{\mathcal{T},p}(\boldsymbol{u}) = \left\{ \sum_{l=0}^{l_{\max}} D^{p/2} 2^{-p(l+1)} \sum_{F \in P_l} |S_F \boldsymbol{u}| \right\}^{1/p}.$$

This expression can be evaluated efficiently and used with Theorem 10 to obtain a two-sample test.

3.3 Application: single-marker switching microscopy

Single Marker Switching (SMS) Microscopy (Betzig et al., 2006; Rust et al., 2006; Egner et al., 2007; Heilemann et al., 2008; Fölling et al., 2008) is a living cell fluorescence microscopy technique in which fluorescent markers which are tagged to a protein structure in the probe are stochastically switched from a no-signal giving (off) state into a signal-giving (on) state. A marker in the on state emits a bunch of photons some of which are detected on a detector before it is either switched off or bleached. From the photons registered on the detector, the position of the marker (and hence of the protein) can be determined. The final image is assembled from all observed individual positions recorded in a sequence of time intervals (frames) in a position histogram, typically a pixel grid.

SMS microscopy is based the principle that at any given time only a very small number of markers are in the on state. As the probability of switching from the off to the on state is small for each individual marker and they remain in the on state only for a very short time (1-100ms). This allows SMS microscopy to resolve features below the diffraction barrier that limits conventional far-field microscopy (see Hell (2007) for a survey) because with overwhelming probability at most one marker within a diffraction limited

spot is in the on state. At the same time this property requires much long acquisition times (1min-1h) to guarantee sufficient sampling of the probe. As a consequence, if the probe moves during the acquisition, the final image will be blurred.

Correcting for this drift and thus improving image quality is an area of active research (Geisler et al., 2012; Deschout et al., 2014; Hartmann et al., 2014; Aspelmeier et al., 2015). In order to investigate the validity of such a drift correction method we introduce a test of the Wasserstein distance between the image obtained from the fist half of the recording time and the second half. This test is based on the distributional upper bound of the limiting distribution which was developed in Section 3.2 in combination with a lower bound of the Wasserstein distance (Pele and Werman, 2009). In fact, there is no standard method for problems of this kind and we argue that the (thresholded) Wasserstein distance is particular useful in such a situation as the specimen moves between the frames without loss of mass, hence the drift induces a transport structure between successive frames. In the following we compare the distribution from the first half of frames with the distribution from the second half scaled with the sample sizes (as in (2.12)). We reject the hypothesis that the distributions from the first and the second half are the same, if our test statistic is larger than the $1-\alpha$ quantile of the distributional bound of the limiting distribution in (3.3). If we have statistical evidence that the tresholded Wasserstein distance is not zero, we can also conclude that there is a significant difference in the Wasserstein distance.

Statistical Model It is common to assume the bursts of photons registered on the detector as independent realizations of a random variable with a density that is proportional to the density of markers in the probe (Aspelmeier et al., 2015). As it is expected that the probe drifts during the acquisition this density will vary over time. In particular, the locations registered at the beginning of the observation will follow a different distribution than those observed at the end.

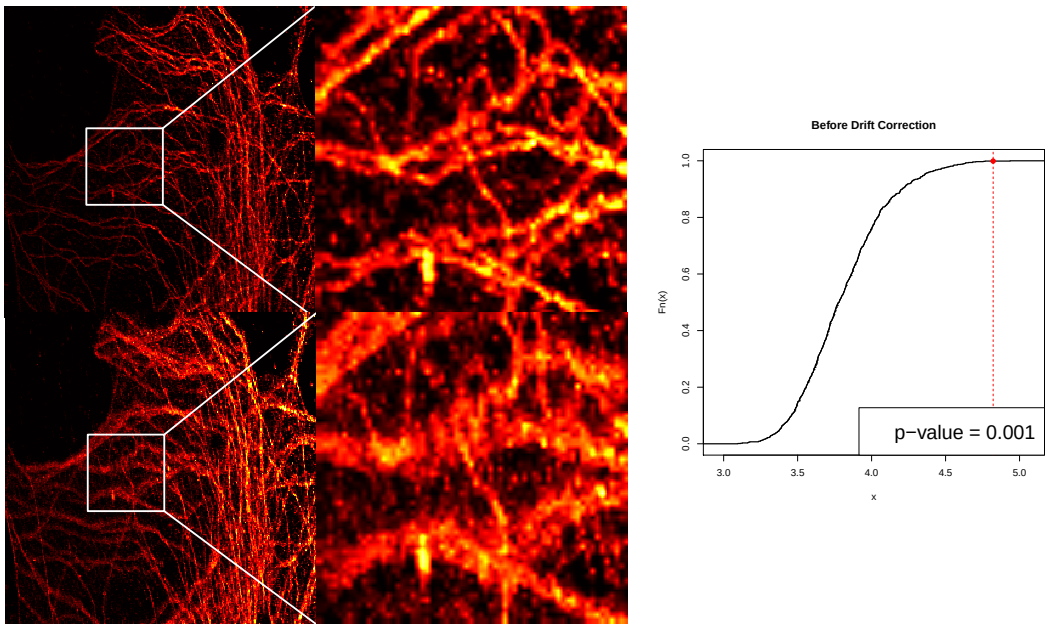

Figure 3.1: Left: Aggregated samples of the first (first row) and the last (second row) 50% of the observation time as heat maps of relative frequency without correction for the drift of the probe. Magnifications of a small area are shown to highlight the blurring of the picture. Right: Empirical distribution function of a sample from the upper bound (tree approximation) of the limiting distribution. The red dot (line) indicates the scaled thresholded Wasserstein distance for $t = 6/256$.

Data and Results We consider an SMS image of a tubulin structure presented in Hartmann et al. (2014) to assess their drift correction method. This image is recorded in 40.000 single frames over a total recording time of 10 minutes (i.e., 15 ms per frame). We compare the aggregated sample collected during the first 50% ($\hat{=}$ 20.000 frames) of the total observation time with the aggregated sample obtained in the last 50% on a 256×256 grid for both the original uncorrected values and for the values where the drift correction of Hartmann et al. (2014) was applied. Heat maps of these four samples are shown in the left hand side of Figure 3.1 (no correction) and Figure 3.2 (corrected), respectively.

The question we will address is: "To what extend has the drift being properly removed by the drift correction?" From the application of the thresholded

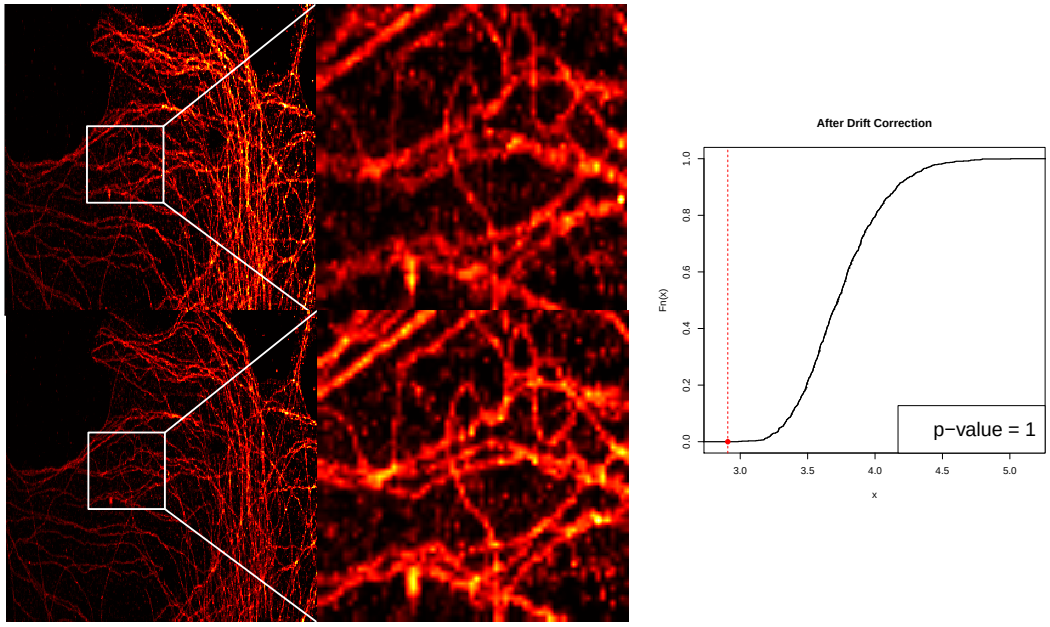

Figure 3.2: Left: Aggregated samples of the first (first row) and the last (second row) 50% of the observation time as heat maps of relative frequency with correction for the drift of the probe. Magnifications of a small area are shown to highlight the drift correction of the picture. Right: Empirical distribution function of a sample from the upper bound (tree approximation) of the limiting distribution. The red dot (line) indicates the scaled thresholded Wasserstein distance after drift correction for $t = 6/256$. The difference between the fist and the second 50% is no longer significant.

Wasserstein distance for different thresholds we expect to obtain detailed understanding of which scales the drift has been removed. As Hartmann et al. (2014) have corrected with a global drift function one might expect that on small spatial scales not all effects have been removed.

We compute the thresholded Wasserstein distance $W_1^{(t)}$ between the two pairs of samples as described in Section 3.1 with different thresholds $t \in \{2, 3, \ldots, 14\}/256$. We compare these values with a sample from the stochastic upper bound for the limiting distribution on regular grids obtained as described in Section 3.2. This allows us to obtain a test for the null hypothesis 'no difference' from Theorem 10. To visualize the outcomes of theses tests for different thresholds t we have plotted the corresponding p-values in Figure 3.3. The red line indicates the magnitude of the drift over the total recording time. As the magnitude is approximately $6/256$, we plot in the right hand side of Figure 3.1 and Figure 3.2 the empirical distribution functions of the upper bound (3.3) and indicate the value of the test-statistic for $t = 6/256$ with a red dot for the data before the correction and after the correction, respectively.

As shown in Figure 3.3 the differences caused by the drift of the probe are recognized as highly statistically significant ($p \leq 0.05$) for thresholds larger than $t = 4/256$. After the drift correction method is applied, the difference is no longer significant for thresholds smaller than $t = 14/256$. The estimated shift during the first and the second 50% of the observations is three pixels in x-direction and one pixel in y-direction. That shows that the significant difference that is detected when comparing the images without drift correction for $t \in \{5, 6, 7, 8, 9, 10\}/256$ is caused in fact by the drift. The fact that there is still a significant difference for large thresholds ($t \geq 14$) in the corrected pictures suggests further intrinsic and local inhomogeneous motion of the specimen or non-polynomial drift that is not captured by the drift model and bleaching effects of fluorescent markers.

In summary, this example demonstrates that our strategy of combining a lower bound for the Wasserstein distance with a stochastic bound of the limiting distribution is capable of detecting subtle differences in a large N setting.

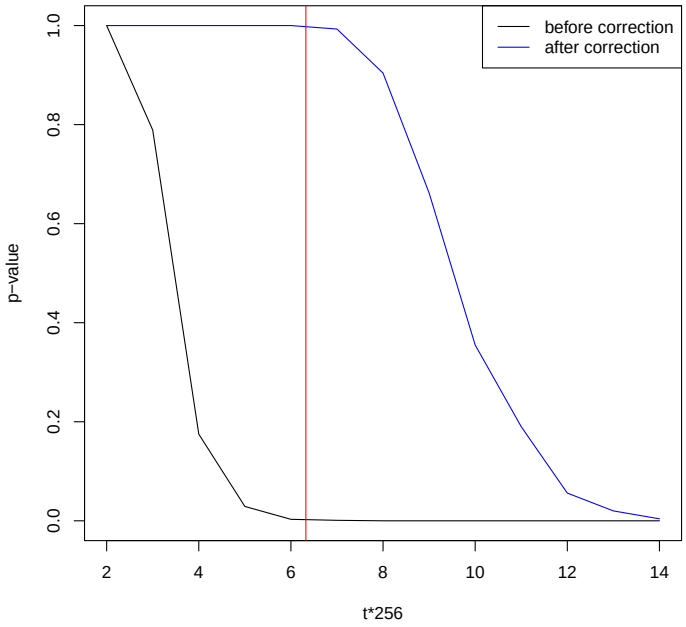

Figure 3.3: P-values for the null hypothesis 'no difference' for different thresholds t before and after the drift correction. The red line indicates the magnitude of the total dirft.

Chapter 4

Probabilistic approximation via exact solvers

This chapter proposes a scheme which utilizes an arbitrary exact solver for the Wasserstein (or any other transport) distance in order to obtain a fast probabilistic approximation.

The first section presents the algorithm. The second section gives theoretical results on the approximation quality, in particular, assessing the dependence of the quality on the size of the underlying space. The third section contains numerical experiments to demonstrate the practical performance of the algorithm.

The chapter is concluded with a discussion section and a section containing the proofs of the presented results.

4.1 Problem and algorithm

Although our meta-algorithm is applicable to any optimal transport distance between probability measures, the theory concerns the *Wasserstein distance*. The idea of the proposed algorithm is to replace a probability measure $\boldsymbol{r} \in \mathcal{P}(\mathcal{X})$ with the empirical measure $\hat{\boldsymbol{r}}_S$ based on i.i.d. picks $X_1, \ldots, X_S \sim \boldsymbol{r}$ for some natural number S. Likewise, replace $\boldsymbol{s}$ with $\hat{\boldsymbol{s}}_S$. Then, use $W_p(\hat{\boldsymbol{r}}_S, \hat{\boldsymbol{s}}_S)$ as a random approximation of $W_p(\boldsymbol{r}, \boldsymbol{s})$.

Algorithm 1 Statistical approximation of $W_p(\boldsymbol{r}, \boldsymbol{s})$

1: **Input:** Probability measures $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_{\mathcal{X}}$, sample size S and number of repetitions B
2: **for** $i = 1 \dots B$ **do**
3: Sample i.i.d. $X_1, \dots, X_S \sim \boldsymbol{r}$ and $Y_1, \dots, Y_S \sim \boldsymbol{s}$
4: $\hat{r}_{S,x} \leftarrow \#\{k : X_k = x\}/S$ **for all** $x \in \mathcal{X}$
5: $\hat{s}_{S,x} \leftarrow \#\{k : Y_k = x\}/S$ **for all** $x \in \mathcal{X}$
6: Compute $\hat{W}^{(i)} \leftarrow W_p(\hat{\boldsymbol{r}}_S, \hat{\boldsymbol{s}}_S)$
7: **end for**
8: **Return:** $\hat{W}_p^{(S)}(\boldsymbol{r}, \boldsymbol{s}) \leftarrow B^{-1} \sum_{i=1}^{B} \hat{W}^{(i)}$

In each of the B iterations in Algorithm 1, the Wasserstein distance between two sets of S point masses has to be computed. For the exact Wasserstein distance, two measures on N points need to be compared. If we take the super-cubic runtime of the auction algorithm as a basis, Algorithm 1 has runtime

$$\mathcal{O}(BS^3 \log S)$$

compared to $\mathcal{O}(N^3 \log N)$ for the exact distance. This means a dramatic reduction of computation time if S is small compared to N.

The application of Algorithm 1 to other optimal transport distances is straightforward. One can simply replace $W_p(\hat{\boldsymbol{r}}_S, \hat{\boldsymbol{s}}_S)$ with the desired distance, e.g. the Sinkhorn distance ((Cuturi, 2013), see also our numerical experiments below).

4.2 Theoretical results

In this chapter, we give general non-asymptotic guarantees for the quality of the approximation $\hat{W}_p^{(S)}(\boldsymbol{r}, \boldsymbol{s})$. To this end, we first give non-asymptotic bounds for the expected L_1-error made by the approximation. That is, we look for bounds of the form

$$E\left[\left|\hat{W}_p^{(S)}(\boldsymbol{r}, \boldsymbol{s}) - W_p(\boldsymbol{r}, \boldsymbol{s})\right|\right] \leq g(S, \mathcal{X}, p), \tag{4.1}$$

for some function g. We are particularly interested in the dependence of the bound on the size N of the space $\mathcal{X}$ and on the sample size S as this determines how the number of sampling points S (and hence the computational burden of Algorithm 1 must be increased for increasing problem size N in order to retain (on average) a certain approximation quality.

In a second step, in Subsection 4.2.2 we use bounds of the form (4.1) to obtain deviation inequalities for $\hat{W}^{(S)}(\boldsymbol{r}, \boldsymbol{s})$ via concentration of measure techniques.

Comparison with results for general measures The question of the convergence of empirical measures to the true measure in expected Wasserstein distance has been considered in detail by Boissard and Gouic (2014) and Fournier and Guillin (2014). The case of the underlying measures being different (that is, the convergence of $EW_p(\hat{\boldsymbol{r}}_S, \hat{\boldsymbol{s}}_S)$ to $W_p(\boldsymbol{r}, \boldsymbol{s})$ when $\boldsymbol{r} \neq \boldsymbol{s}$) has not been considered to the best of our knowledge. Theorem 11 is very similar to the main result of Boissard and Gouic (2014). However, we give a result here, which is explicitly tailored to finite spaces and makes explicit the dependence of the constants on the size N of the underlying space $\mathcal{X}$. In fact, when we consider finite spaces $\mathcal{X}$ which are subsets of $\mathbb{R}^D$ later in Theorem 13, we will see that in contrast to the results of Boissard and Gouic (2014), the rate of convergence (in S) does not change when the dimension gets large, but rather the dependence of the constants on N changes. This is a valuable insight as our main concern here is how the subsample size S (driving the computational cost) must be chosen when N grows in order to retain a certain approximation quality.

4.2.1 Expected absolute error

For $\delta > 0$ the *covering number* $\mathcal{N}(\mathcal{X}, \delta)$ of $\mathcal{X}$ is defined as the minimal number of closed balls with radius δ and center in $\mathcal{X}$ that is needed to cover all of $\mathcal{X}$. Note that in contrast to continuous spaces, $\mathcal{N}(\mathcal{X}, \delta)$ is bounded by N for all $\delta > 0$. With this, we have the following

Theorem 11. *Let $\hat{\boldsymbol{r}}_S$ be the empirical measure obtained from i.i.d. samples $X_1, \dots, X_S \sim \boldsymbol{r}$, then*

$$E\left[W_p^p(\hat{\boldsymbol{r}}_S, \boldsymbol{r})\right] \leq \mathcal{E}_q / \sqrt{S} \tag{4.2}$$

for every $2 \leq q \in \mathbb{N}$ and

$$\begin{aligned} &\mathcal{E}_q := \mathcal{E}_q(\mathcal{X}, p) \\ &\quad := 2^{p-1} q^{2p} (\operatorname{diam}(\mathcal{X}))^p \left(q^{-(l_{\max}+1)p} \sqrt{N} + \sum_{l=0}^{l_{\max}} q^{-lp} \sqrt{\mathcal{N}(\mathcal{X}, q^{-l} \operatorname{diam}(\mathcal{X}))} \right) \end{aligned} \tag{4.3}$$

with $l_{\max} \in \mathbb{N}$ a parameter that can be chosen freely to minimize the upper bound.

Based on Theorem 11, we can formulate a bound for the mean approximation of Algorithm 1.

Theorem 12. *If $\boldsymbol{r} \neq \boldsymbol{s}$ and $\hat{W}_p^{(S)}(\boldsymbol{r}, \boldsymbol{s})$ is obtained from Algorithm 1 then for every natural $q \geq 2$*

$$E\left[\left|\hat{W}_p^{(S)}(\boldsymbol{r}, \boldsymbol{s}) - W_p(\boldsymbol{r}, \boldsymbol{s})\right|\right] \leq 2\mathcal{E}_q^{1/p} S^{-1/(2p)}. \tag{4.4}$$

Note that the upper bound in Theorem 12 behaves as $\mathcal{O}(1/S^{1/(2p)})$ for large S, which does not reflect the $\sqrt{n}$ scaling rate under the alternative in the distributional limits of Theorem 1. This issue is discussed in more detail in Section 4.4.

Regular Grids While the constant $\mathcal{E}_q$ in Theorem 11 may be difficult to compute or estimate in general, we can give explicit bounds in the case when $\mathcal{X}$ is a finite set of points in Euclidean space. They exhibit the dependence of the approximation error on the size of the space N.

In particular, it comprises the case when the measures represent images (two- or more dimensional).

Theorem 13. *Let $\mathcal{X} \subset [0, L]^D$ a subset of Euclidean space with $L > 0$ and let the metric d on $\mathcal{X}$ be the usual Euclidean metric. Then,*

$$\mathcal{E}_q \leq 2^p q^{2p+2} (\operatorname{diam}(\mathcal{X}))^p \left(\frac{4D}{\operatorname{diam}(\mathcal{X})} \right)^{D/2} C_{D,p}(N)$$

where

$$C_{D,p}(N) = \begin{cases} 1 & \text{if } D/2 - p < 0, \\ 1 + \frac{1}{p} \log_q N & \text{if } D/2 - p = 0, \\ 1 + N^{\frac{1}{2}(1 - \frac{2p}{D})} & \text{if } D/2 - p > 0. \end{cases}$$

This result gives control over the error made by the approximation $\hat{W}_p^{(S)}(\boldsymbol{r}, \boldsymbol{s})$ of $W_p(\boldsymbol{r}, \boldsymbol{s})$. Of particular interest is the behavior of this error for high resolution images, that is $N \to \infty$. We distinguish three cases. In the *low-dimensional case* $p' = D/2 - p < 0$, we have $C_{D,p}(N) = \mathcal{O}(1)$. Hence, in this case, the approximation error is $\mathcal{O}(S^{-\frac{1}{2p}})$ independent of the size of the image. In the *critical case* $p' = 0$ the approximation error is no longer independent of N but is of order $\mathcal{O}\left(\log(N) S^{-\frac{1}{2p}}\right)$. Finally, in the *high-dimensional case* the dependence on N becomes stronger with an approximation error of order

$$\mathcal{O}\left(\left(\frac{N^{(1 - \frac{2p}{D})}}{S} \right)^{\frac{1}{2p}} \right).$$

In all cases one can choose $S = o(N)$ while still guaranteeing vanishing approximation error for $N \to \infty$. In practice, this means that for large images, S can typically be chosen (much) smaller than N to obtain a good approximation of the Wasserstein distance.

4.2.2 Concentration bounds

Based on the bounds for the expected approximation error we now give non-asymptotic guarantees for the approximation error in the form of deviation bounds using standard concentration of measure techniques.

Theorem 14. *If $\hat{W}_p^{(S)}(\boldsymbol{r}, \boldsymbol{s})$ is obtained from Algorithm 1, then for every $z \geq 0$*

$$P\left[|\hat{W}_p^{(S)}(\boldsymbol{r}, \boldsymbol{s}) - W_p(\boldsymbol{r}, \boldsymbol{s})| \geq z + \frac{2\mathcal{E}_q^{1/p}}{S^{1/2p}}\right] \leq 2\exp\left(-\frac{SBz^{2p}}{8\ \mathrm{diam}(\mathcal{X})^{2p}}\right). \tag{4.5}$$

Note that while the mean approximation quality $2\mathcal{E}_q^{1/p}/S^{1/(2p)}$ only depends on the subsample size S, the stochastic variability (see the right hand side term in (4.5)) depends on the product SB. This means that the repetition number B cannot decrease the expected error but it can decrease the probability of large deviations from it.

4.3 Simulations

In this section we report the performance of Algorithm 1 in numerical experiments.

4.3.1 Setup

We compute optimal transport distances *exactly*, that is, solving the full problem, as well as *approximately* with Algorithm 1 for all possible combinations of the following parameters:

- three different solvers computing the Wasserstein distance. These are 1) CPLEX [1] using the network solver; 2) the transportation simplex and the 3) shortlist method, both are described in Gottschlich and Schuhmacher (2014) and implemented in the R-package `transport` (Schuhmacher et al., 2014), which we use.

  Additionally, we compute the Sinkhorn distance (Cuturi, 2013), that is an entropically regularized optimal transport distance. For this we use the implementation in the R-package `barycenter` (Klatt, 2016) of the algorithm presented in Cuturi (2013).

[1] `www.ibm.com/software/commerce/optimization/cplex-optimizer/`

- nine pairs of 2-D images (using the normalized grayscale values as the probability mass). The images are taken from the benchmark database DOTmark (Schrieber et al., 2016) and consist of one pair of images from each of the three classes "White Noise", "Cauchy Density" and "Classic Images", where each pair is considered in the resolutions 32×32, 64×64 and 128×128.

- the ground distance (that is, d in our notation) is always the euclidean distance between pixels in the image, where the pixels are assumed to be equally spaced in the unit square $[0, 1]^2$. For the cost exponent we take $p \in \{1, 2, 3\}$.

For the approximate computation with Algorithm 1 we use all combinations of the following parameters:

- the subsample size S runs through the values $\{100, 500, 1000, 2000, 4000\}$.

- the number of repetitions B runs through the values $\{1, 2, 5\}$.

Each approximate combination was repeated 5 times for every possible combination of the above parameters. All calculations were run on one core of a Linux server (AMD Opteron Processor 6140 from 2011 with 2.6 GHz) and the result as well as the computation time were recorded.

4.3.2 Results

Overall performance In order to assess the performance of the algorithm and relate the approximation quality to the reduction in computation time, we report the mean absolute error made by the Algorithm 1 and the ratio of the runtime of the approximation and the runtime for the exact computation on the same instance.

Figure 4.1 plots the relative errors against the relative runtimes, averaging over all different solvers, image classes and choices of parameters S and B. Figure 4.2 shows the same results but separated for each parameter pair S,B, in order to assess their influence.

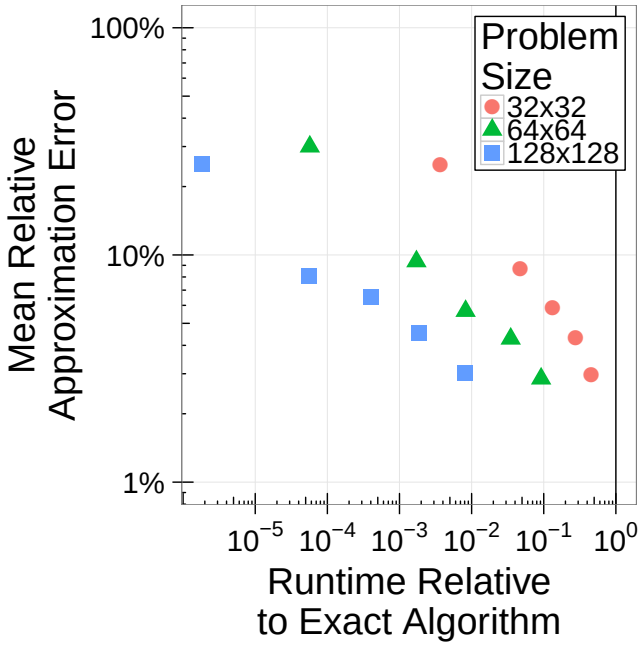

Figure 4.1: Relative error and relative runtime compared to the exact computation of the proposed scheme. Optimal transport distances and its approximations were computed between images of different sizes (32×32, 64×64, 128×128). Each point represents a specific parameter choice in the scheme and is a mean over different problem instances, solvers and cost exponents. For the relative runtimes the geometric mean is reported. For details on the parameters see Figure 4.2.

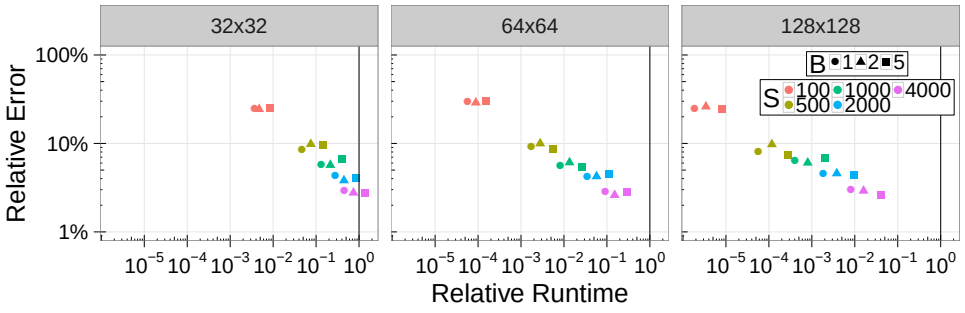

Figure 4.2: Relative errors vs. relative runtimes relative to the exact computation for different parameters S and B and different problem sizes. Both axes are on log-scale.

- The main driving factor of the approximation quality and the reduction in runtime is the subsample size S. Relatively low subsample sizes S yield good approximations and (depending on the resolution) considerable reductions in computation time. For example, $S = 4000$ on a 128×128 image yields (on average) an approximation error of 3% while reducing the computation time by a factor of 100.

- The repetition number B has hardly any effect on the approximation quality, while increasing the computation time of the algorithm linearly.

- The resolution has little effect on the approximation quality, as suggested by the theoretical bounds in Section 4.1. However, it greatly influences the relative runtime, as the runtime of the exact algorithms scales with the resolution while the runtime of Algorithm 1 only scales in S and B.

Figure 4.3 shows a scatter plot of the relative error of the approximation as S varies. Each point in the scatter plot corresponds to a different set of parameters or a different trial. The experiments are distinguished by the image class and the target quantity (Wasserstein or Sinkhorn distance), respectively.

- The approximation error appears to decay polynomially in S in all cases.

- The class of images considered has a considerable influence on the approximation quality. Specifically, the Algorithm 1 performs best for images generated from a Cauchy density, somewhat worse but still comparable for classic images and much worse for white noise images. This could lead to the interpretation that the proposed approximation performs better, the more structure the images have.

- The algorithm performs equally well for the Wasserstein and the Sinkhorn distance, with the latter showing a marginally but consistently better approximation error.

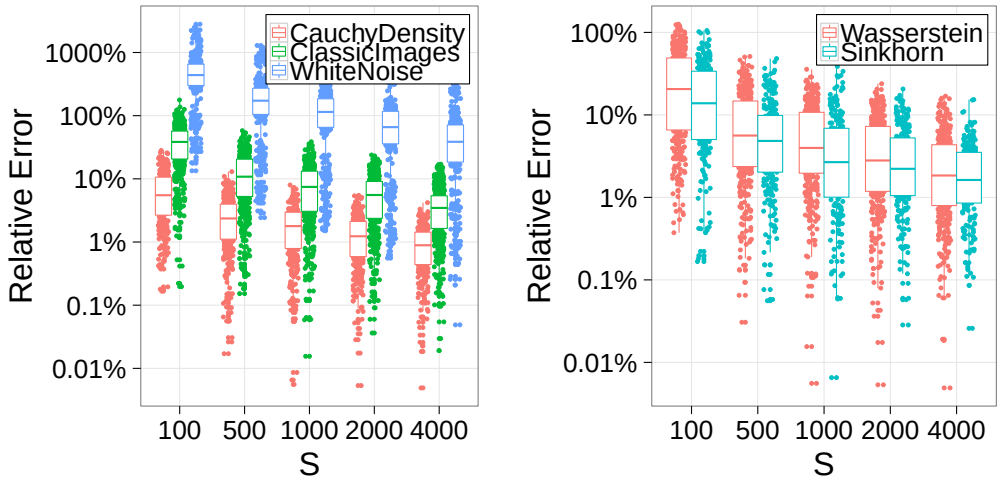

Figure 4.3: A comparison of the relative errors for different image classes (left) and between the approximations of the Wasserstein and Sinkhorn distances (right).

Figure 4.4 shows the signed relative error of the approximation relative to S. Its distribution is strongly skewed for smaller values of S while this skew vanishes almost completely for $S \geq 2000$. This means that the approximation generated by Algorithm 1 will often overestimate the true transportation distance when S is small.

4.4 Discussion

As our simulations demonstrate, subsampling is a simple, yet powerful tool to obtain good approximations to Wasserstein distances with only a small fraction of the runtime and memory required for exact computation. It is especially remarkable that for a fixed amount of subsampled points, and therefore a fixed amount of time and memory, the relative error is independent of the resolution of the images. Based on these results, we expect the subsampling algorithm to return similarly precise results with even higher resolutions of the images it is applied to, while the effort to obtain them stays the same.

The numerical results (Figure 4.2) show a inverse polynomial decrease of

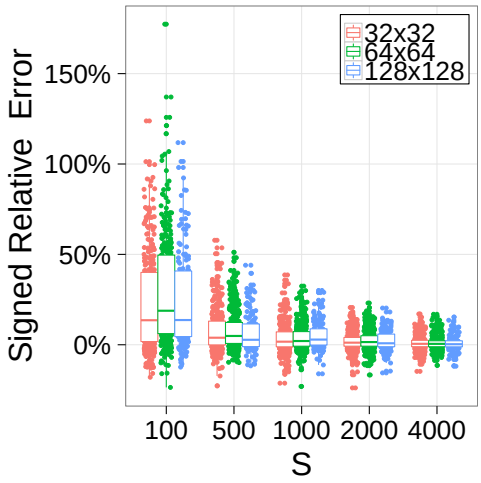

Figure 4.4: The signed relative approximation error $\left(\hat{W}_p^{(S)}(\boldsymbol{r},\boldsymbol{s}) - W_p(\boldsymbol{r},\boldsymbol{s})\right)/W_p(\boldsymbol{r},\boldsymbol{s})$ showing that the approximation overestimates the exact distance for small S but the bias vanishes for larger S.

the approximation error with S, in accordance with the theoretical results. As we see little dependence on the cost exponent p we suspect that the rate $O(S^{-1/2p})$ might be improved upon. In fact, recent work on asymptotics of empirical Wasserstein distances would suggest an $\mathcal{O}(S^{-1/2})$ rate (Sommerfeld and Munk, 2016).

When applying the algorithm, it is important to note that the quality of the returned values depend on the structure of the data. In very irregular instances it might be necessary to increase the sample size in order to obtain similarly precise results, while in regular structures a small sample size might suffice.

Our scheme allows the parameters to be easily tuned towards faster runtimes or more precise results, as desired. Increases and decreases of the sample size S are recommended to influence the performance in either direction, while the parameter B should only be increased, if a particularly low variability of the estimate is required or if the repetitions can be computed in parallel. Otherwise, the higher runtime should be spent with a higher sample size (compare Figure 4.2).

The scheme presented here can readily be applied to other optimal transport distances, as long as an exact solver is available, as we demonstrated with the Sinkhorn distance (Cuturi, 2013). Empirically, we can report good performance in this case, suggesting that entropically regularized distances might be even more amenable to subsampling approximation than the Wasserstein distance itself. Extending the theoretical results to this case would require an analysis of the mean speed of convergence of empirical Sinkhorn distances.

All in all, subsampling proves to be a very powerful and versatile tool that can be used with virtually any optimal transport solver as back-end and has both theoretical approximation error guarantees, and a convincing performance in practice.

4.5 Proofs

4.5.1 Proof of Theorem 11

Proof strategy The method used in this proof has been employed before to bound the mean rate of convergence of the empirical Wasserstein distance (in Boissard and Gouic (2014); Fournier and Guillin (2014)). In essence, it constructs on the space $\mathcal{X}$ a tree and bounds the Wasserstein distance with some transport metric in the tree, which can either be computed explicitly or bounded easily.

More precisely, in our case of finite spaces, let $\mathcal{T}$ be a spanning tree of $\mathcal{X}$ and $d_{\mathcal{T}}$ the metric on $\mathcal{X}$ defined by path length in the tree. Clearly, the tree metric $d_{\mathcal{T}}$ dominates the original metric on $\mathcal{X}$ and hence $W_p(\boldsymbol{r}, \boldsymbol{s}) \leq W_p^{\mathcal{T}}(\boldsymbol{r}, \boldsymbol{s})$ for all $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}(\mathcal{X})$, where $W_p^{\mathcal{T}}$ denotes the Wasserstein distance evaluated with respect to the tree metric. The goal is now to bound $E\left[(W_p^{\mathcal{T}}(\hat{\boldsymbol{r}}_S, \boldsymbol{r}))^p\right]$.

Building the tree We build a q-ary tree on $\mathcal{X}$. In the following we set $l_{\max} = \lceil \log_q N \rceil$. For $l \in \{0, \dots, l_{\max}\}$ we let $Q_l \subset \mathcal{X}$ be the center points of a $q^{-l} \operatorname{diam}(\mathcal{X})$ covering of $\mathcal{X}$, that is

$$\bigcup_{x \in Q_l} B(x, q^{-l} \operatorname{diam}(\mathcal{X})) = \mathcal{X}, \text{ and } |Q_l| = \mathcal{N}(\mathcal{X}, q^{-l} \operatorname{diam}(\mathcal{X})),$$

where $B(x,\epsilon) = \{x' \in \mathcal{X} : d(x,x') \leq \epsilon\}$. Additionally set $Q_{l_{\max}+1} = \mathcal{X}$.

Now define $\tilde{Q}_l = Q_l \times \{l\}$ and we will build a tree structure on $\cup_{l=0}^{l_{\max}+1} \tilde{Q}_l$.

Since we must have $|\tilde{Q}_0| = 1$ we can take this element as the root. Assume now that the tree already contains all elements of $\cup_{j=0}^{l} \tilde{Q}_j$ Then, we add to the tree all elements of $\tilde{Q}_{l+1}$ by choosing for $(x, l+1) \in \tilde{Q}_{l+1}$ (exactly one) parent element $(x', l) \in \tilde{Q}_l$ such that $d(x,x') \leq q^{-l} \operatorname{diam}(\mathcal{X})$. This is possible, since Q_l is a $q^{-l} \operatorname{diam}(\mathcal{X})$ covering of $\mathcal{X}$. We set the length of the connecting edge to $q^{-l} \operatorname{diam}(\mathcal{X})$.

In this fashion we obtain a spanning tree $\mathcal{T}$ of $\cup_{l=0}^{l_{\max}+1} \tilde{Q}_l$ and a partition $\{\tilde{Q}_l\}_{l=0,\ldots,l_{\max}+1}$. About this tree we know that

- it is in fact a tree. First, it is connected, because the construction starts with one connected component and in every subsequent step all additional vertices are connected to it. Second, it contains no cycles. To see this let $((x_1,l_1),\ldots,(x_K,l_K))$ a cycle in $\mathcal{T}$. Without loss of generality we may assume $l_1 = \min\{l_1,\ldots,l_K\}$. Then, (x_1,l_1) must have at least two edges connecting it to vertices in a $\tilde{Q}_l$ with $l \geq l_1$ which is impossible by construction.
- $|\tilde{Q}_l| = \mathcal{N}(\mathcal{X}, q^{-l} \operatorname{diam}(\mathcal{X}))$ for $0 \leq l \leq l_{\max}$.
- $d(x, \operatorname{par}(x)) = q^{-l+1} \operatorname{diam}(\mathcal{X})$ whenever $x \in \tilde{Q}_l$.
- $d(x,x') \leq d_{\mathcal{T}}((x, l_{\max}+1), (x', l_{\max}+1))$.

Since the leafs of $\mathcal{T}$ can be identified with $\mathcal{X}$ a measure $\boldsymbol{r} \in \mathcal{P}(\mathcal{X})$ canonically defines a probability measure $\boldsymbol{r}^{\mathcal{T}} \in \mathcal{P}(\mathcal{T})$ for which $r^{\mathcal{T}}_{(x,l_{\max}+1)} = r_x$ and $r^{\mathcal{T}}_{(x,l)} = 0$ for $l \leq l_{\max}$. In slight abuse of notation we will denote the measure $\boldsymbol{r}^{\mathcal{T}}$ simply by $\boldsymbol{r}$. With this notation, we have $W_p(\boldsymbol{r},\boldsymbol{s}) \leq W_p^{\mathcal{T}}(\boldsymbol{r},\boldsymbol{s})$ for all $\boldsymbol{r},\boldsymbol{s} \in \mathcal{P}(\mathcal{X})$.

Wasserstein distance on trees Note also that $\mathcal{T}$ is *ultra-metric* that is, all its leaves are at the same distance from the root. For trees of this type, we can define a height function $h : \mathcal{X} \to [0,\infty)$ such that $h(x) = 0$ if $x \in \mathcal{X}$ is a leaf and $h(\operatorname{par}(x)) - h(x) = d_{\mathcal{T}}(x, \operatorname{par}(x))$ for all $x \in \mathcal{X} \setminus \operatorname{root}(\mathcal{X})$.

There is an explicit formula for the Wasserstein distance on ultra-metric trees (Kloeckner, 2013). Indeed, if $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}(\mathcal{X})$ then

$$(W_p^{\mathcal{T}}(\boldsymbol{r}, \boldsymbol{s}))^p = 2^{p-1} \sum_{x \in \mathcal{X}} \left(h(\mathrm{par}(x))^p - h(x)^p\right) \left|(S_{\mathcal{T}}\boldsymbol{r})_x - (S_{\mathcal{T}}\boldsymbol{s})_x\right|. \tag{4.6}$$

with the operator $S_{\mathcal{T}}$ as defined in (2.11). For the tree $\mathcal{T}$ constructed above and $x \in \tilde{Q}_l$ with $l = 0, \ldots, l_{\max}$ we have

$$h(x) = \sum_{j=l}^{l_{\max}} q^{-j} \operatorname{diam}(\mathcal{X}).$$

and therefore

$$\operatorname{diam}(\mathcal{X}) q^{-l} \leq h(x) \leq 2 \operatorname{diam}(\mathcal{X}) q^{-l}.$$

This yields

$$(h(\mathrm{par}(x))^p - (h(x))^p) \leq (\operatorname{diam}(\mathcal{X}))^p q^{-(l-2)p}.$$

The formula (4.6) thus yields

$$\begin{aligned} &E\left[W_p^p(\hat{\boldsymbol{r}}_S, \boldsymbol{r})\right] \\ &\leq 2^{p-1} q^{2p} (\operatorname{diam}(\mathcal{X}))^p \sum_{l=0}^{l_{\max}+1} q^{-lp} \sum_{x \in \tilde{Q}_l} E|(S_{\mathcal{T}}\hat{\boldsymbol{r}}_S)_x - (S_{\mathcal{T}}\boldsymbol{r})_x|. \end{aligned}$$

Since $(S_{\mathcal{T}}\hat{\boldsymbol{r}}_S)_x$ is the mean of S i.i.d. Bernoulli variables with expectation $(S_{\mathcal{T}}\boldsymbol{r})_x$ we have

$$\begin{aligned} \sum_{x \in \tilde{Q}_l} E|(S_{\mathcal{T}}\hat{\boldsymbol{r}}_S)_x - (S_{\mathcal{T}}\boldsymbol{r})_x| &\leq \sum_{x \in \tilde{Q}_l} \sqrt{\frac{(S_{\mathcal{T}}\boldsymbol{r})_x (1 - (S_{\mathcal{T}}\boldsymbol{r})_x)}{S}} \\ &\leq \frac{1}{\sqrt{S}} \left(\sum_{x \in \tilde{Q}_l} (S_{\mathcal{T}}\boldsymbol{r})_x\right)^{1/2} \left(\sum_{x \in \tilde{Q}_l} (1 - (S_{\mathcal{T}}\boldsymbol{r})_x)\right)^{1/2} \\ &\leq \sqrt{|\tilde{Q}_l|/S}, \end{aligned}$$

using Hölder's inequality and the fact that $\sum_{x\in\tilde{Q}_l}(S_{\mathcal{T}}\boldsymbol{r})_x = 1$ for all $l = 0,\dots,l_{\max}+1$.

This finally yields

$$
\begin{aligned}
&E\left[W_p^p(\hat{\boldsymbol{r}}_S,\boldsymbol{r})\right]\\
&\leq 2^{p-1}q^{2p}(\operatorname{diam}(\mathcal{X}))^p\left(q^{-(l_{\max}+1)p}\sqrt{N}+\sum_{l=0}^{l_{\max}}q^{-lp}\sqrt{\mathcal{N}(\mathcal{X},q^{-l}\operatorname{diam}(\mathcal{X}))}\right)/\sqrt{S}\\
&\leq \mathcal{E}_q(\mathcal{X},p)/\sqrt{S},
\end{aligned}
$$

using in the last inequality that $l_{\max} = \lceil \log_q N\rceil$.

4.5.2 Proof of Theorem 12

The statement of the theorem is an immediate consequence of the reverse triangle inequality for the Wasserstein distance, Jensen's inequality and Theorem 11,

$$
\begin{aligned}
E\left[\left|\hat{W}_p^{(S)}(\boldsymbol{r},\boldsymbol{s})-W_p(\boldsymbol{r},\boldsymbol{s})\right|\right] &\leq E\left[W_p(\hat{\boldsymbol{r}}_S,\boldsymbol{r})+W_p(\hat{\boldsymbol{s}}_S,\boldsymbol{s})\right]\\
&\leq E\left[W_p^p(\hat{\boldsymbol{r}}_S,\boldsymbol{r})\right]^{1/p}+E\left[W_p^p(\hat{\boldsymbol{s}}_S,\boldsymbol{s})\right]^{1/p}\\
&\leq 2\mathcal{E}_q^{1/p}/S^{1/(2p)}.
\end{aligned}
$$

4.5.3 Proof of Theorem 13

We want to use (4.3). First, note that (Shalev-Shwartz and Ben-David, 2014, Example 27.1)

$$\mathcal{N}([0,L]^D,\epsilon)\leq 2^D D^D\epsilon^{-D}.$$

Therefore,

$$\mathcal{N}(\mathcal{X},q^{-l}\operatorname{diam}(\mathcal{X}))\leq\mathcal{N}([0,L]^D,q^{-l}\operatorname{diam}(\mathcal{X})/2)\leq\left(\frac{4D}{\operatorname{diam}(\mathcal{X})}\right)^D q^{lD}.$$

This yields

$$\begin{aligned}
&\sum_{l=0}^{l_{\max}} q^{-lp}\sqrt{\mathcal{N}(\mathcal{X}, q^{-l}\operatorname{diam}(\mathcal{X}))} \\
&\leq \left(\frac{4D}{\operatorname{diam}(\mathcal{X})}\right)^{D/2} \sum_{l=0}^{l_{\max}} q^{l(D/2-p)} \\
&\leq q\left(\frac{4D}{\operatorname{diam}(\mathcal{X})}\right)^{D/2} \times \begin{cases} 1 + q^{l_{\max}(D/2-p)} & \text{if } D/2-p \neq 0, \\ l_{\max} & \text{if } D/2-p = 0. \end{cases}
\end{aligned}$$

Setting $l_{\max} = \lceil \beta \log_q N \rceil$ for some $\beta > 0$ to be specified yields (using (4.3))

$$\begin{aligned}
\mathcal{E}_q \leq 2^{p-1} q^{2p+2} (\operatorname{diam}(\mathcal{X}))^p \left(\frac{4D}{\operatorname{diam}(\mathcal{X})}\right)^{D/2} \\
\times \begin{cases} 1 + N^{-\beta p+1/2} + N^{\beta(D/2-p)} & \text{if } D/2-p \neq 0 \\ 1 + N^{-\beta p+1/2} + \beta \log_q N & \text{if } D/2-p = 0. \end{cases}
\end{aligned}$$

If $D/2 - p < 0$ we can choose β large enough such that $1 + N^{-\beta p+1/2} + N^{\beta(D/2-p)} \leq 2$. If $D/2 - p > 0$ we choose $\beta = 1/D$ such that $1 + N^{-\beta p+1/2} + N^{\beta(D/2-p)} \leq 1 + N^{\frac{1}{2}(1-2p/D)}$. Finally, for $D/2 - p = 0$ we set $\beta = 1/(2p)$ such that $1 + N^{-\beta p+1/2} + \beta \log_q N \leq 2 + \frac{1}{p} \log_q N$. This concludes the proof.

4.5.4 Proof of Theorem 14

We introduce some additional notation. For $(x, y), (x', y') \in \mathcal{X}^2$ we set

$$d_{\mathcal{X}^2}((x, y), (x', y')) = \{d^p(x, x') + d^p(y, y')\}^{1/p}$$

We further define the function $Z : (\mathcal{X}^2)^{SB} \to \mathbb{R}$ via

$$\begin{aligned}
&((x_{11}, y_{11}), \ldots, (x_{SB}, y_{SB})) \\
&\mapsto \frac{1}{B}\sum_{i=1}^{B}\left[W_p\left(\frac{1}{S}\sum_{j=1}^{S}\delta_{x_{ji}}, \frac{1}{S}\sum_{j=1}^{S}\delta_{y_{ji}}\right) - W_p(\boldsymbol{r}, \boldsymbol{s})\right].
\end{aligned}$$

Since $W_p^p(\cdot,\cdot)$ is jointly convex (Villani, 2008, Thm.4.8), we have

$$W_p\left(\frac{1}{S}\sum_{j=1}^{S}\delta_{x_j},\frac{1}{S}\sum_{j=1}^{S}\delta_{y_j}\right) \le \left\{\frac{1}{S}\sum_{j=1}^{S}W_p^p(\delta_{x_j},\delta_{y_j})\right\}^{1/p}$$
$$= S^{-1/p}\left\{\sum_{j=1}^{S}d^p(x_j,y_j)\right\}^{1/p}.$$

Our first goal is to show that Z is Lipschitz continuous. To this end, let $((x_{11},y_{11}),\dots,(x_{SB},y_{SB}))$ and $((x'_{11},y'_{11}),\dots,(x'_{SB},y'_{SB}))$ arbitrary elements of $(\mathcal{X}^2)^{SB}$. Then, using the reverse triangle inequality and the relations above

$$\begin{aligned}
&|Z((x_{11},y_{11}),\dots,(x_{SB},y_{SB})) - Z((x'_{11},y'_{11}),\dots,(x'_{SB},y'_{SB}))|\\
&\le \frac{1}{B}\sum_{i=1}^{B}\left|W_p\left(\frac{1}{S}\sum_{j=1}^{S}\delta_{x_{ji}},\frac{1}{S}\sum_{j=1}^{S}\delta_{y_{ji}}\right) - W_p\left(\frac{1}{S}\sum_{j=1}^{S}\delta_{x'_{ji}},\frac{1}{S}\sum_{j=1}^{S}\delta_{y'_{ji}}\right)\right|\\
&\le \frac{1}{B}\sum_{i=1}^{B}\left[W_p\left(\frac{1}{S}\sum_{j=1}^{S}\delta_{x_{ji}},\frac{1}{S}\sum_{j=1}^{S}\delta_{x'_{ji}}\right) + W_p\left(\frac{1}{S}\sum_{j=1}^{S}\delta_{y_{ji}},\frac{1}{S}\sum_{j=1}^{S}\delta_{y'_{ji}}\right)\right]\\
&\le \frac{S^{-1/p}}{B}\sum_{i=1}^{B}\left[\left\{\sum_{j=1}^{S}d^p(x_{ji},x'_{ji})\right\}^{1/p} + \left\{\sum_{j=1}^{S}d^p(y_{ji},y'_{ji})\right\}^{1/p}\right]\\
&\le \frac{S^{-1/p}}{B}(2B)^{\frac{p-1}{p}}\left\{\sum_{i,j}d^p_{\mathcal{X}^2}((x_{ji},y_{ji}),(x'_{ji},y'_{ji}))\right\}^{1/p}
\end{aligned}$$

Hence, $Z/2$ is Lipschitz continuous with constant $(SB)^{-1/p}$ relative to the p-metric generated by $d_{\mathcal{X}^2}$ on $(\mathcal{X}^2)^{SB}$.

For $\tilde{\boldsymbol{r}} \in \mathcal{P}(\mathcal{X}^2)$ let $H(\cdot \mid \tilde{\boldsymbol{r}})$ denote the relative entropy with respect to $\tilde{\boldsymbol{r}}$. Since $\mathcal{X}^2$ has $d_{\mathcal{X}^2}$-diameter $2^{1/p}\operatorname{diam}(\mathcal{X})$, we have by (Bolley and Villani, 2005, Particular case 5) that for every $\tilde{\boldsymbol{s}}$

$$W_p(\tilde{\boldsymbol{r}},\tilde{\boldsymbol{s}}) \le \left(8\operatorname{diam}(\mathcal{X})^{2p}H(\tilde{\boldsymbol{r}} \mid \tilde{\boldsymbol{s}})\right)^{1/2p}. \tag{4.7}$$

If $X_{11},\dots,X_{SB} \sim \boldsymbol{r}$ and $Y_{11},\dots,Y_{SB} \sim \boldsymbol{s}$ are all independent, we have

$$Z((X_{11},Y_{11}),\dots,(X_{SB},Y_{SB})) \sim \hat{W}_p^{(S)}(\boldsymbol{r},\boldsymbol{s}) - W_p(\boldsymbol{r},\boldsymbol{s}).$$

The Lipschitz continuity of Z and the transportation inequality (4.7) yields a concentration result for this random variable. In fact, by (Gozlan and Léonard, 2007, Lemma 6) we have

$$P\left[\hat{W}_p^{(S)}(\boldsymbol{r},\boldsymbol{s}) - W_p(\boldsymbol{r},\boldsymbol{s}) \geq E\left[\hat{W}_p^{(S)}(\boldsymbol{r},\boldsymbol{s}) - W_p(\boldsymbol{r},\boldsymbol{s})\right] + z\right] \leq \exp\left(\frac{-SBz^{2p}}{8\ \mathrm{diam}(\mathcal{X})^{2p}}\right).$$

for all $z \geq 0$. Note that $-Z$ is Lipschitz continuous as well and hence, by the union bound,

$$P\left[\left|\hat{W}_p^{(S)}(\boldsymbol{r},\boldsymbol{s}) - W_p(\boldsymbol{r},\boldsymbol{s})\right| \geq E\left[\left|\hat{W}_p^{(S)}(\boldsymbol{r},\boldsymbol{s}) - W_p(\boldsymbol{r},\boldsymbol{s})\right|\right] + z\right] \leq 2\exp\left(\frac{-SBz^{2p}}{8\ \mathrm{diam}(\mathcal{X})^{2p}}\right).$$

Now, with the reverse triangle inequality, Jensen's inequality and Theorem 11,

$$\begin{aligned} E\left[\left|\hat{W}_p^{(S)}(\boldsymbol{r},\boldsymbol{s}) - W_p(\boldsymbol{r},\boldsymbol{s})\right|\right] &\leq E\left[W_p(\hat{\boldsymbol{r}}_S,\boldsymbol{r}) + W_p(\hat{\boldsymbol{s}}_S,\boldsymbol{s})\right] \\ &\quad E\left[W_p^p(\hat{\boldsymbol{r}}_S,\boldsymbol{r})\right]^{1/p} + \left[W_p^p(\hat{\boldsymbol{s}}_S,\boldsymbol{s})\right]^{1/p} \\ &\leq 2\mathcal{E}_q^{1/p}/S^{1/(2p)}. \end{aligned}$$

Together with the last concentration inequality above, this concludes the proof of Theorem 14.

Bibliography

Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Agulló-Antolín, M., Cuesta-Albertos, J. A., Lescornel, H., and Loubes, J.-M. (2015). A parametric registration model for warped distributions with Wasserstein's distance. *Journal of Multivariate Analysis*, 135:117–130.

Ajtai, M., Komlós, J., and Tusnády, G. (1984). On optimal matchings. *Combinatorica*, 4(4):259–264.

Ambrosio, L. (2003). Lecture Notes on Optimal Transport Problems. In *Mathematical Aspects of Evolving Interfaces*, pages 1–52. Springer.

Anderson, N. H., Hall, P., and Titterington, D. M. (1994). Two-Sample Test Statistics for Measuring Discrepancies Between Two Multivariate Probability Density Functions Using Kernel-Based Density Estimates. *Journal of Multivariate Analysis*, 50(1):41–54.

Aspelmeier, T., Egner, A., and Munk, A. (2015). Modern statistical challenges in high-resolution fluorescence microscopy. *Annual Review of Statistics and Its Application*, 2(1):163–202.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bertsekas, D. P. (1992). Auction algorithms for network flow problems: A tutorial introduction. *Computational Optimization and Applications*, 1(1):7–66.

Betzig, E., Patterson, G. H., Sougrat, R., Lindwasser, O. W., Olenych, S., Bonifacino, J. S., Davidson, M. W., Lippincott-Schwartz, J., and Hess, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793):1642–1645.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217.

Bigot, J., Gouet, R., Klein, T., and López, A. (2013). Geodesic PCA in the Wasserstein space. *arXiv:1307.7721*.

Bobkov, S. and Ledoux, M. (2014). One-dimensional empirical measures, order statistics and Kantorovich transport distances. *preprint*.

Boissard, E. and Gouic, T. L. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539–563.

Boissard, E., Gouic, T. L., and Loubes, J.-M. (2015). Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759.

Bolley, F. and Villani, C. (2005). Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. In *Annales de La Faculté Des Sciences de Toulouse: Mathématiques*, volume 14, pages 331–352.

Bonnans, J. F. and Shapiro, A. (2013). *Perturbation Analysis of Optimization Problems*. Springer.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45.

Cappelli, R., Erol, A., Maio, D., and Maltoni, D. (2000). Synthetic fingerprint-image generation. In *Proceedings of The15th International Conference on Pattern Recognition*, volume 3, pages 471–474.

Chandrasekaran, R., Kabadi, S. N., and Murthy, K. G. (1982). Some NP-complete problems in linear programming. *Operations Research Letters*, 1(3):101–104.

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science (New York, N.Y.)*, 326(5960):1694–1697.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300.

Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters. In *Proceedings of The 31st International Conference on Machine Learning*, pages 685–693.

Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., and Rodríguez-Rodríguez, J. M. (1999). Tests of goodness of fit based on the L2-Wasserstein distance. *The Annals of Statistics*, 27(4):1230–1239.

Del Barrio, E., Giné, E., and Utzet, F. (2005). Asymptotics for L2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1):131–189.

Del Barrio, E., Lescornel, H., and Loubes, J.-M. (2015). A statistical analysis of a deformation model with Wasserstein barycenters: Estimation procedure and goodness of fit test. *arXiv:1508.06465*.

Del Barrio, E. and Loubes, J.-M. (2017). Central Limit Theorems for empirical transportation cost in general dimension. *arXiv preprint arXiv:1705.01299*.

Deschout, H., Zanacchi, F. C., Mlodzianoski, M., Diaspro, A., Bewersdorf, J., Hess, S. T., and Braeckmans, K. (2014). Precisely and accurately localizing single emitters in fluorescence microscopy. *Nature methods*, 11(3):253–266.

Dobrushin, R. (1970). Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3):458–486.

Donoho, D. L. and Liu, R. C. (1988). Pathologies of some Minimum Distance Estimators. *The Annals of Statistics*, 16(2):587–608.

Dorea, C. C. Y. and Ferreira, D. B. (2012). Conditions for equivalence between Mallows distance and convergence to stable laws. *Acta Mathematica Hungarica*, 134(1-2):1–11.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95(1):125–140.

Dümbgen, L., Samworth, R., and Schuhmacher, D. (2011). Approximation by log-concave distributions, with applications to regression. *The Annals of Statistics*, 39(2):702–730.

Egner, A., Geisler, C., von Middendorff, C., Bock, H., Wenzel, D., Medda, R., Andresen, M., Stiel, A. C., Jakobs, S., Eggeling, C., Schönle, A., and Hell, S. W. (2007). Fluorescence nanoscopy in whole cells by asynchronous localization of photoswitching emitters. *Biophysical Journal*, 93(9):3285–3290.

Erbar, M. and Maas, J. (2012). Ricci Curvature of Finite Markov Chains via Convexity of the Entropy. *Archive for Rational Mechanics and Analysis*, 206(3):997–1038.

Evans, S. N. and Matsen, F. A. (2012). The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592.

Fang, Z. and Santos, A. (2014). Inference on directionally differentiable functions. *arXiv:1404.3763*.

Fölling, J., Bossi, M., Bock, H., Medda, R., Wurm, C. A., Hein, B., Jakobs, S., Eggeling, C., and Hell, S. W. (2008). Fluorescence nanoscopy by ground-state depletion and single-molecule return. *Nature Methods*, 5(11):943–945.

Fournier, N. and Guillin, A. (2014). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, pages 1–32.

Freitag, G., Czado, C., and Munk, A. (2007). A nonparametric test for similarity of marginals—With applications to the assessment of population bioequivalence. *Journal of Statistical Planning and Inference*, 137(3):697–711.

Freitag, G. and Munk, A. (2005). On Hadamard differentiability in k-sample semiparametric models—with applications to the assessment of structural relationships. *Journal of Multivariate Analysis*, 94(1):123–158.

Gal, T., Greenberg, H. J., and Hillier, F. S., editors (1997). *Advances in Sensitivity Analysis and Parametric Programming*, volume 6 of *International Series in Operations Research & Management Science*. Springer.

Gangbo, W. and McCann, R. J. (2000). Shape recognition via Wasserstein distance. *Quarterly of Applied Mathematics*, LVIII(4):705–737.

Geisler, C., Hotz, T., Schönle, A., Hell, S. W., Munk, A., and Egner, A. (2012). Drift estimation for single marker switching based imaging schemes. *Optics express*, 20(7):7274–7289.

Gelbrich, M. (1990). On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203.

Gill, R. D., Wellner, J. A., and Præstgaard, J. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1) [with discussion and reply]. *Scandinavian Journal of Statistics*, 16(2):97–128.

Gottschlich, C. and Huckemann, S. (2014). Separating the real from the synthetic: Minutiae histograms as fingerprints of fingerprints. *IET Biometrics*, 3(4):291–301.

Gottschlich, C. and Schuhmacher, D. (2014). The Shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems. *PLoS ONE*, 9(10):e110214.

Gozlan, N. and Léonard, C. (2007). A large deviation approach to some transportation cost inequalities. *Probability Theory and Related Fields*, 139(1-2):235–283.

Gozlan, N., Roberto, C., Samson, P.-M., and Tetali, P. (2013). Displacement convexity of entropy and related inequalities on graphs. *Probability Theory and Related Fields*, 160(1-2):47–94.

Gray, R. M. (1988). *Probability, Random Processes, and Ergodic Properties*. Springer.

Halder, A. and Bhattacharya, R. (2011). Model validation: A probabilistic formulation. In *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, pages 1692–1697.

Hartmann, A., Huckemann, S., Dannemann, J., Laitenberger, O., Geisler, C., Egner, A., and Munk, A. (2014). Drift estimation in sparse sequential dynamic imaging: With application to nanoscale fluorescence microscopy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, to appear.

Heilemann, M., van de Linde, S., Schüttpelz, M., Kasper, R., Seefeldt, B., Mukherjee, A., Tinnefeld, P., and Sauer, M. (2008). Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. *Angewandte Chemie (International Ed. in English)*, 47(33):6172–6176.

Hell, S. W. (2007). Far-field optical nanoscopy. *Science*, 316(5828):1153–1158.

Horowitz, J. and Karandikar, R. L. (1994). Mean rates of convergence of empirical measures in the Wasserstein metric. *Journal of Computational and Applied Mathematics*, 55(3):261–273.

Hung, M. S., Rom, W. O., and Waren, A. D. (1986). Degeneracy in transportation problems. *Discrete Applied Mathematics*, 13(2):223–237.

Jain, A. K. (2007). Technology: Biometric recognition. *Nature*, 449(7158):38–40.

Johnson, O. and Samworth, R. (2005). Central limit theorem and convergence to stable laws in Mallows distance. *Bernoulli*, 11(5):829–845.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17.

Kantorovich, L. V. and Rubinstein, G. S. (1958). On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59.

Klatt, M. (2016). Barycenter: Wasserstein Barycenter.

Klee, V. and Witzgall, C. (1968). Facets and vertices of transportation polytopes. In George Bernard Dantzig and Veinott, A. F., editors, *Mathematics of the Decision Sciences*, volume II of *Lectures in Applied Mathematics*, pages 257–282. American Mathematical Soc., Providence, RI.

Kloeckner, B. R. (2013). A geometric study of Wasserstein spaces: Ultrametrics. *Mathematika*, pages 1–17.

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature Methods*, 7(10):813–819.

Ling, H. and Okada, K. (2007). An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):840–853.

Luenberger, D. G. and Ye, Y. (2008). *Linear and Nonlinear Programming*. Springer.

Maio, D., Maltoni, D., Cappelli, R., Wayman, J. L., and Jain, A. K. (2002). FVC2002: Second fingerprint verification competition. In *Proceedings of the 16th International Conference on Pattern Recognition*, volume 3, pages 811–814. IEEE.

Mallows, C. L. (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 43(2):508–515.

Maltoni, D., Maio, D., Jain, A. K., and Prabhakar, S. (2009). *Handbook of Fingerprint Recognition*. Springer.

Mason, D. M. (2016). A Weighted Approximation Approach to the Study of the Empirical Wasserstein Distance. In *High Dimensional Probability VII*, pages 137–154. Birkhäuser, Cham.

Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):223–241.

Ni, K., Bresson, X., Chan, T., and Esedoglu, S. (2009). Local histogram based segmentation using the Wasserstein distance. *International Journal of Computer Vision*, 84(1):97–111.

Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E. B., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S., Moore, W., Moss, R. B., Herzenberg, L. A., and Walther, G. (2016). Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations. *PLOS ONE*, 11(3):e0151859.

Otto, F. (2001). The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174.

Oudre, L., Jakubowicz, J., Bianchi, P., and Simon, C. (2012). Classification of periodic activities using the Wasserstein distance. *IEEE Transactions on Biomedical Engineering*, 59(6):1610–1619.

Pele, O. and Werman, M. (2009). Fast and robust earth mover's distances. In *IEEE 12th International Conference on Computer Vision*, pages 460–467.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rachev, S. T. (1985). The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4):647–676.

Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory*. Springer.

Rippl, T., Munk, A., and Sturm, A. (2015). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis, to appear*.

Robinson, S. M. (1977). A Characterization of Stability in Linear Programming. *Operations Research*, 25(3):435–447.

Rockafellar, R. T. (1984). Directional differentiability of the optimal value function in a nonlinear programming problem. In *Sensitivity, Stability and Parametric Analysis*, number 21 in Mathematical Programming Studies, pages 213–226. Springer.

Römisch, W. (2004). Delta Method, Infinite Dimensional. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc.

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Rust, M. J., Bates, M., and Zhuang, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods*, 3(10):793–796.

Ruttenberg, B. E., Luna, G., Lewis, G. P., Fisher, S. K., and Singh, A. K. (2013). Quantifying spatial relationships from whole retinal images. *Bioinformatics*, 29(7):940–946.

Samworth, R. and Johnson, O. (2004). Convergence of the empirical process in Mallows distance, with an application to bootstrap performance. *arXiv:math/0406603*.

Samworth, R. and Johnson, O. (2005). The empirical process in Mallows distance, with application to goodness-of-fit tests. *arXiv:math/0504424*.

Schloss, P. D. (2015). Schloss lab 454 standard operating procedure - http://www.mothur.org/wiki/454_SOP - 2015-07-01 17:53:34. `http://www.mothur.org/wiki/454_SOP`.

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE*, 6(12):e27310.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541.

Schmitzer, B. (2016). A Sparse Multi-Scale Algorithm for Dense Optimal Transport. *Journal of Mathematical Imaging and Vision*.

Schrieber, J., Schuhmacher, D., and Gottschlich, C. (2016). DOTmark - A Benchmark for Discrete Optimal Transport. *arXiv:1610.03368 [cs, math]*.

Schuhmacher, D., Gottschlich, C., and Baehre, B. (2014). R-package transport: Optimal transport in various forms - https://cran.r-project.org/package=transport.

Seguy, V. and Cuturi, M. (2015). An algorithmic approach to compute principal geodesics in the Wasserstein space. *arXiv:1506.07944*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of optimization theory and applications*, 66(3):477–487.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186.

Shapiro, A. (1992). Perturbation analysis of optimization problems in Banach spaces. *Numerical Functional Analysis and Optimization*, 13(1-2):97–116.

Shirdhonkar, S. and Jacobs, D. W. (2008). Approximate earth mover's distance in linear time. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley series in probability and mathematical statistics. Wiley, New York.

Sierksma, G. (2001). *Linear and Integer Programming: Theory and Practice, Second Edition*. CRC Press.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press.

Sommerfeld, M. (2017). Otinference: Inference for Optimal Transport - https://cran.r-project.org/package=otinference.

Sommerfeld, M. and Munk, A. (2016). Inference for Empirical Wasserstein Distances on Finite Spaces. *arXiv:1610.03287 [stat]*.

Sommerfeld, M. and Munk, A. (2017). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages n/a–n/a.

Srivastava, S., Li, C., and Dunson, D. B. (2015). Scalable Bayes via barycenter in Wasserstein space. *arXiv:1508.05880*.

Talagrand, M. (1992). Matching random samples in many dimensions. *The Annals of Applied Probability*, pages 846–856.

Talagrand, M. (1994). The transportation cost from the uniform measure to the empirical measure in dimension$\geq$ 3. *The Annals of Probability*, pages 919–959.

Tameling, C., Sommerfeld, M., and Munk, A. (2017). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *arXiv:1707.00973 [math, stat]*.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature*, 449(7164):804–810.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. Springer.

Vasershtein, L. N. (1969). Markov processes over denumerable products of spaces describing large system of automata. *Problemy Peredači Informacii*, 5(3):64–72.

Villani, C. (2003). *Topics in Optimal Transportation*. Number 58. American Mathematical Soc.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer.

Wasserman, L. (2011). *All of Statistics*. Springer Science & Business Media.

List of Figures

Max Sommerfeld

Friedrich-Ebert-Str. 50 d – +49 176 26 31 70 88 –
max.sommerfeld@mathematik.uni-goettingen.de

Positionen

Wissenschaftlicher Mitarbeiter

Felix-Bernstein Institut für Mathematische Statistik, Göttingen

Seit 1. Juni 2014

Visiting Graduate Fellow

Statistical and Applied Mathematichal Sciences Institute,

Raleigh, North Carolina, USA

September 2013 – Juni 2014

Ausbildung

MSc Mathematik

Universität Göttingen

September 2013

BSc Mathematik

Universität Hannover

September 2011

Abitur

Gymnasium Mellendorf, 30900 Wedemark

Juni 2007

Publikationen

- Huckemann, Stephan, Kwang-Rae Kim, Axel Munk, Florian Rehfeldt, Max Sommerfeld, Joachim Weickert, and Carina Wollnik. "The Circular SiZer, Inferred Persistence of Shape Parameters and Application to Early Stem Cell Differentiation." *Bernoulli* 22, no. 4 (November 2016): 2113–42. doi:10.3150/15-BEJ722.
- Sommerfeld, Max, and Axel Munk. "Inference for Empirical Wasserstein Distances on Finite Spaces." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017.doi:10.1111/rssb.12236.
- Sommerfeld, Max, Stephan Sain, and Armin Schwartzman. "Confidence Regions for Spatial Excursion Sets from Repeated Random Field Observations, with an Application to Climate." *Journal of the American Statistical Association*, (2017). doi:10.1080/01621459.2017.1341838.

Reviewer für wissenschaftliche Journale

Annals of Statistics, Bernoulli, Annals of Applied Statistics, Electronic Journal of Statistics