Empirical Optimal Transport on Discrete Spaces: Limit Theorems, Distributional Bounds and Applications

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen

Doktorgrades

"Doctor rerum naturalium"

der Georg-August-Universität zu Göttingen

im Promotionsprogramm

"PhD School of Mathematical Sciences (SMS)"

der Georg-August University School of Science (GAUSS)

vorgelegt von

Carla Tameling

aus Münster

Göttingen, 2018

Betreuungsausschuss:

Prof. Dr. Axel Munk
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Anja Sturm
Institut für Mathematische Stochastik, Universität Göttingen

Mitglieder der Prüfungskommission:

Referent:
Prof. Dr. Axel Munk
Institut für Mathematische Stochastik, Universität Göttingen

Korreferentin:
Prof. Dr. Anja Sturm
Institut für Mathematische Stochastik, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Jörg Brüdern
Mathematisches Institut, Universität Göttingen

Jun.-Prof. Dr. Daniel Rudolf
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Anita Schöbel
Institut für Numerische und Angewandte Mathematik, Universität Göttingen

Dr. Yoav Zemel
Institut für Mathematische Stochastik, Universität Göttingen

Tag der mündlichen Prüfung: 11.12.2018

Acknowledgments

Preface

The beginning of *optimal transport* dates back more than two centuries. Since then it has played its way into different mathematical disciplines. Among others it is a well-established tool in probability theory to study for example limit laws, derive concentration inequalities or for point process approximations. During the last two decades distances based on optimal transport became also well-known in statistical theory and found applications in a broad range of fields. Examples include machine learning, risk measures in finance, classification and goodness-of-fit testing.

From the statistical point of view the interesting questions are the rate of convergence, concentration results and distributional limits for the empirical optimal transport distance, i.e., based on the empirical measure generated from a sample. Distributional limits are an essential tool in statistics for hypothesis testing and to derive confidence bands. The theory of distributional limits for the empirical optimal transport distance was restricted for a long time only to the one-dimensional case. These results were extended to different special cases in higher dimensions during the last five years, but are still incomplete. In this thesis we enhance the distributional limit results for the empirical optimal transport distance on *countable* spaces.

Optimal transport is suitable to measure spatial distances between structures recorded as images as it finds the optimal matching between these structures. Moreover, the optimal transport plan (the optimal solution of optimal transport) allows to deduce how far different parts of the structures are apart. To quantify spatial proximity of structures recorded as images is especially important in biology, e.g. to study protein distributions. We take advantage of the potential of optimal transport to measure spatial distances to derive a new method based on optimal transport to analyze spatial proximity of proteins in super-resolution microscopy images.

This thesis is organized as follows. In Chapter 1 we give an introduction to optimal transport based on the historical development of this field of research. Furthermore, we

state the main results from this thesis and compare them to existing literature. Chapter 2 deals with distributional limits for the empirical optimal transport distance on countable metric spaces. As a special case we consider ground spaces that are trees (see Section 2.2). The results for the trees are used to derive a distributional upper bound for the limiting distribution on general countable metric spaces. We conclude this chapter with the numerical evaluation of this upper bound.

In Chapter 3 we derive a new method for colocalization analysis of images generated by coordinate-targeted super-resolution microscopy methods based on optimal transport. This new method - optimal transport colocalization (OTC) - is evaluated on different real data sets to deduce different properties such as robustness against background and independence of resolution.

This thesis is concluded with Chapter 4 - a discussion of the presented results and an outlook to open research questions deduced from the results in this thesis.

Previous publications and joint work   Large parts of this thesis have already been published in Tameling et al. (2017), Tameling and Munk (2018) and Tameling et al. (2018). The preprint Tameling et al. (2017) considers the theory for the limit laws for the empirical Wasserstein distance for measures supported on countable spaces (Chapter 2).

The explicit limit distribution in Chapter 2.2 is joint work with Max Sommerfeld. The author of this dissertation and Max Sommerfeld contributed equally to the derivation of these results. The work on the distributional upper bound for the limiting distribution was done by Max Sommerfeld. The numerical evaluation of this distributional upper bound given in Chapter 2.2 was already published in Tameling and Munk (2018).

Most parts of Chapter 3 were published in the preprint Tameling et al. (2018). All STED images in this chapter were generated by the Jakob's Lab (Till Stephan, Stefan Stoldt) from the Max Planck Institute for Biophysical Chemistry, Göttingen.

Contents

List of symbols

$\mathbb{N}$ — Set of positive integers

$\mathbb{Z}$ — Set of integers

$\mathbb{R}$ — Set of real numbers

$\mathbb{R}_+$ — Set of non-negative real numbers

ℓ^1 — Space of absolutely summable sequences

ℓ^∞ — Space of bounded sequences

$\ell^1_{d_{x_0}^p}$ — Weighted ℓ^1-space

$\ell^\infty_{d_{x_0}^{-p}}$ — Weighted ℓ^∞-space

$\mathcal{P}(X)$ — Set of probability measures on X

$\mathcal{P}_p(X)$ — Set of probability measures on X with finite p-th moment

$|\cdot|$ — Absolute value

$\|\cdot\|$ — Euclidean distance

$\|\cdot\|_{\ell^1(d_{x_0}^p)}$ — Weighted ℓ^1-norm

$\|\cdot\|_{\ell^\infty(d_{x_0}^{-p})}$ — Weighted ℓ^∞-norm

$\xrightarrow{\mathscr{D}}$ — Convergence in distribution

$T \sharp \mu$ — Push-forward measure

$\mu \otimes \nu$ — Product measure of μ and ν

$\operatorname{supp}(\mu)$ — Support of the measure μ

$\operatorname{Var}(X)$ — Variance of X

$\operatorname{Cov}(X, Y)$ — Covariance of X and Y

$\mathcal{N}(0, \Sigma)$ — Gaussian process with mean zero and covariance structure Σ

$\langle \cdot, \cdot \rangle$ — Dual pairing

$\mathcal{S}(r, s)$ — Set of primal optimal solutions in dependence of the marginals (r, s)

$\mathcal{S}^*(r, s)$ — Set of dual optimal solutions in dependence of the marginals (r, s)

B^C — Complement of the set B

CHAPTER 1

Introduction

The theory of optimal transport dates back to the early work of the French mathematician Monge from the 18th century (Monge, 1781). Monge was concerned with the problem of building fortresses. Therefore, he considered the question of how to move a certain amount of material that is extracted from the earth or a mine to the building site of the fortress in the most efficient way (see Figure 1.1). Here, efficiency means the least possible transportation cost, which he assumed to be given by the product of the mass and the distance. To formalize this problem in mathematical terms, we can understand the ground level of the hole and the building site of the fortress as spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Furthermore, we model the material as probability measures on $\mathcal{X}$ and $\mathcal{Y}$ as obviously the volume of the hole and the fortress have to be the same and describe the cost of transporting material from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$ by a measurable non-negative function $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. The question of finding an optimal (minimal cost) assignment reads

$$\min_{T} \int_{\mathcal{X}} c(x, T(x)) \, d\mu(x), \tag{1.1}$$
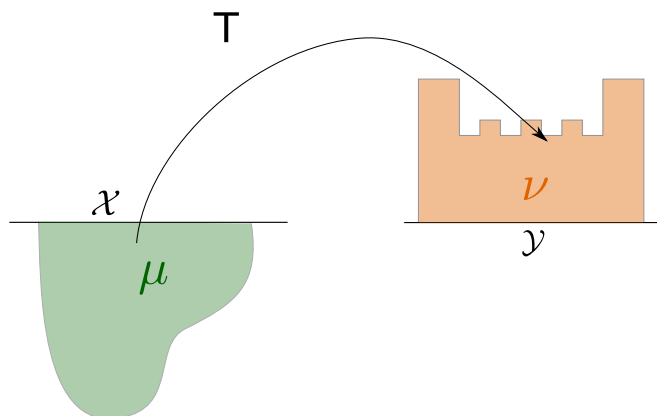
Figure 1.1: **Monge's problem** of transporting extracted material to the building site of the fortress.

where we minimize over all measurable maps $T \colon \mathcal{X} \to \mathcal{Y}$ such that $T \sharp \mu = \nu$. Here, $T \sharp \mu$ denotes the push-forward measure, i.e., for a measurable set $A \subset \mathcal{Y}$ the push-forward measure is given by $T \sharp \mu(A) = \mu(T^{-1}(A))$. Monge originally formulated this problem for $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^D$ and $c(x, y) = \|x - y\|$, where $\| \cdot \|$ denotes the Euclidean distance. The drawback of this formulation of the optimal transport problem is that mass splitting is not allowed. This means that for each $x \in \mathcal{X}$ we need to find a unique destination $y = T(x)$. Therefore, one needs to impose fairly strong regularity conditions on μ and ν to make this problem solvable. The fact that it took more than two centuries until Sudakov (1979) claimed that he found a proof for the existence of an optimal map to Monge's original problem (the correct version of this proof can be found in Ambrosio (2003)) and that Evans and Gangbo (1999) gave a rigorous proof, shows how difficult this problem is.

A relaxed version of Monge's optimal transport problem was introduced by Kantorovich (Kantorovich, 1948, 1958). For probability measures μ supported on the space $\mathcal{X}$ and ν on $\mathcal{Y}$ the Kantorovich optimal transport problem is given as

$$\min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\pi(x, y), \tag{1.2}$$

where

$$\Pi(\mu, \nu) := \{ \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \colon \pi(A \times \mathcal{Y}) = \mu(A), \quad \pi(\mathcal{X} \times B) = \nu(B)$$
$$\text{for all } A \subset \mathcal{X}, B \subset \mathcal{Y} \text{ measurable} \} \tag{1.3}$$

is the set of *couplings* of μ and ν, i.e., the set of probability measures on the product space $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν, respectively. In contrast to Monge's problem (1.1), the feasible set $\Pi(\mu, \nu)$ of this formulation of the optimal transport problem is never empty as the product measure $\mu \otimes \nu$ is a feasible coupling, which may has non-finite transportation cost. The Kantorovich formulation can be seen as a relaxation of Monge's problem, as it allows mass splitting, i.e., the mass located at $x \in \mathcal{X}$ can be distributed to several $y \in \mathcal{Y}$. If T is an optimal solution of Monge's problem (1.1), then $\pi = (Id \times T) \sharp \mu$ in a feasible coupling, i.e., an element of $\Pi(\mu, \nu)$. This transport plan is the optimal solution of (1.2) if the cost function c is continuous and μ has no atoms, i.e., there is no $x \in \mathcal{X}$ with $\mu(\{x\}) > 0$ (Ambrosio, 2003, Thm. 2.1). In the rest of this thesis we only consider the formulation of Kantorovich (1.2) and will refer to this problem as the *optimal transport problem*. Kantorovich also introduced the dual of this problem

$$\sup_{(\phi, \psi) \in \Phi_c} \int_{\mathcal{Y}} \psi \, d\nu - \int_{\mathcal{X}} \phi \, d\mu, \tag{1.4}$$

where

$$\Phi_c := \Big\{(\phi, \psi) \in L^1(d\mu) \times L^1(d\nu) \colon \psi(y) - \phi(x) \leq c(x, y)$$

$$\text{for } \mu\text{-almost all } x \in \mathcal{X} \text{ and } \nu\text{-almost all } y \in \mathcal{Y}\Big\} \quad (1.5)$$

and proved that strong duality holds (the original proof can be found in Kantorovich (1958), for an accessible proof see Villani (2003, Thm. 1.3)).

The heuristic in terms of the building fortresses problem between the primal optimal transport problem (1.2) and its dual (1.4) is as follows. In the primal setting the goal is to find the optimal transference plan to achieve the minimal possible total transport cost for transporting the material from the mine to the building site of the fortress. On the contrary, in the dual setting one can think of an external company that is hired to take care of the transportation. So for this external company $\phi(x)$ is the price for which they can buy the material from the mine and $\psi(y)$ is the price for which they can sell the material to the building site of the fortress. Their profit is given by $\psi(y) - \phi(x)$. To make sure that they are competitive their profit should be less or equal than the transport cost $c(x, y)$. Otherwise, there would be no need to hire them. To sum up, the dual problem asks for maximizing the profit of the external company with the constraint that they have to be competitive.

1.1 The discrete case

If we restrict $\mathcal{X}$ and $\mathcal{Y}$ to be finite spaces, i.e., $\mathcal{X} = \{x_1, \ldots, x_N\}$ and $\mathcal{Y} = \{y_1, \ldots, y_M\}$, the optimal transport problem can be written as a linear program

$$\min_{\boldsymbol{w}} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} c(x, y) w_{x,y}$$

$$\text{s. t.} \sum_{y \in \mathcal{Y}} w_{x,y} = r_x,$$

$$\sum_{x \in \mathcal{X}} w_{x,y} = s_y, \quad (1.6)$$

$$\boldsymbol{w} \geq 0,$$

where $\boldsymbol{w} \geq 0$ has to be understood component-wise. Here, $\boldsymbol{w}, \boldsymbol{r}, \boldsymbol{s}$ are elements in $\mathbb{R}^{N \times M}, \mathbb{R}^N$ and $\mathbb{R}^M$, respectively. Furthermore, c is a $N \times M$ matrix, such that $c(x, y)$ describes the cost to transport one unit from x to y.

In case of finite spaces, the probability measures can be represented by vectors that are

non-negative and sum up to one. More precisely, we denote by

$$\mathcal{P}(X) = \left\{ r = (r_x)_{x \in X} : r_x \geq 0 \; \forall x \in X, \sum_{x \in X} r_x = 1 \right\} \tag{1.7}$$

the set of probability measures on X. A vector $r \in \mathcal{P}(X)$ represents the probability measure $\sum_{x \in X} r_x \delta_x$ with δ_x being the dirac measure at x.

For the linear program in (1.6) r and s do not need to be probability vectors. There exists a solution for this problem as long as the supply equals the demand, i.e., $\sum_{x \in X} r_x = \sum_{y \in \mathcal{Y}} s_y$. If we are concerned with the optimal matching of commodities that are countable it is reasonable to use the number of items for the right hand site of the constraints and even restrict w to be in $\mathbb{N}^{N \times M}$. However, there are also commodities which can be better modeled as probability measures. For example, water or sand or any other good that can be split into infinitesimal small portions.

Parallel to Kantorovich, also Koopmans (Koopmans, 1949) and Hitchcock (Hitchcock, 1941) worked on the optimal transport problem on finite spaces primarily motivated by economical research questions. The Nobel Prize for economics which was awarded to Kantorovich and Koopmans in 1975 demonstrates the outstanding importance of the theory related to optimal transport.

The above introduced setting can be generalized to countable spaces X and $\mathcal{Y}$. Here, the probability measures can still be described by (1.7), they are sequences that are non-negative and sum up to one.

1.2 Distance based on optimal transport

Based on optimal transport a distance on the space of probability measures was developed. This distance is known as Wasserstein distance (Vasershtein, 1969), Earth Mover's distance (Rubner et al., 2000), Kantorovich-Rubinshtein distance (Kantorovich and Rubinshtein, 1958) or Mallows distance (Mallows, 1972).

Definition 1.1 (Wasserstein distance)**.** *Let (X, d) be a Polish metric space and μ and ν probability measures on X. For $p \geq 1$ the p-th Wasserstein distance is given by*

$$W_p(\mu, \nu) = \left\{ \min_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d^p(x, y) \, d\pi(x, y) \right\}^{1/p}, \tag{1.8}$$

where $\Pi(\mu, \nu)$ is the set of couplings defined in (1.3)*.*

To see that this is actually a distance we refer to Villani (2008, p. 94).

In the case of a countable space $\mathcal{X} = \{x_1, x_2, \ldots\}$ the Wasserstein distance is defined via a linear program given in (1.6) with distance as cost function. To be more precise, $W_p^p(r, s)$ is the optimal value of

$$
\begin{aligned}
\min_{w} \quad & \sum_{x,x' \in \mathcal{X}} d^p(x, x') w_{x,x} \\
\text{s. t.} \quad & \sum_{x' \in \mathcal{X}} w_{x,x'} = r_x, \\
& \sum_{x \in \mathcal{X}} w_{x,x'} = s_{x'}, \\
& w \geq 0,
\end{aligned}
\tag{1.9}
$$

where r and s are probability measures on $\mathcal{X}$, i.e., elements in $\mathcal{P}(\mathcal{X})$ defined in (1.7). If we do not impose any condition on r and s the p-th Wasserstein distance $W_p(r, s)$ may be infinite. We can guarantee that $W_p(r, s)$ is finite if the p-th moments of r and s exist, i.e., $r, s \in \mathcal{P}_p(\mathcal{X})$ with

$$
\mathcal{P}_p(\mathcal{X}) := \left\{ r \in \mathcal{P}(\mathcal{X}) : \sum_{x \in \mathcal{X}} d^p(x_0, x) r_x < \infty \right\}.
\tag{1.10}
$$

Here, x_0 is some fixed but arbitrary element in $\mathcal{X}$. Note, that the space is independent of the choice of x_0 as for some other element $\tilde{x}_0$ in $\mathcal{X}$ it holds

$$
\begin{aligned}
\sum_{x \in \mathcal{X}} d^p(\tilde{x}_0, x) r_x &\leq \sum_{x \in \mathcal{X}} 2^{p-1} \left(d^p(\tilde{x}_0, x_0) + d^p(x_0, x) \right) r_x \\
&= 2^{p-1} \left(d^p(\tilde{x}_0, x_0) + \sum_{x \in \mathcal{X}} d^p(x_0, x) r_x \right) < \infty.
\end{aligned}
$$

The Wasserstein distance metrizes weak convergence in $\mathcal{P}_p$ and implies convergence of the moments of order p. During the last two decades this distance has become a standard tool in probability, e.g. to study limit laws (Johnson and Samworth (2005); Rachev and Rüschendorf (1994); Shorack and Wellner (1986)), to derive bounds for Monte Carlo computation schemes such as MCMC (Eberle (2014); Rudolf and Schweizer (2015)), for point process approximations (Barbour and Brown, 1992; Schuhmacher, 2009), bootstrap convergence (Bickel and Freedman, 1981) or to quantify measures of risk (Rachev et al., 2011).

Besides of its theoretical importance, the Wasserstein distance is used in many applications as an empirical measure to compare complex objects, e.g. in image retrieval (Rubner et al., 2000), deformation analysis (Panaretos and Zemel, 2016), meta genomics (Evans and Matsen, 2012), computer vision (Ni et al., 2009), goodness-of-fit testing

(Munk and Czado, 1998; del Barrio et al., 2000), machine learning (Rolet et al., 2016; Arjovsky et al., 2017) and two-sample testing (Ramdas et al., 2017). There are several reasons why practitioners choose the Wasserstein distance. Among others, advantages of the Wasserstein distance are that it incorporates the distance from the underlying space, the intuitive interpretation as amount of work and that it performs exceptionally well at capturing human perception of similarity.

In such applications we usually have only access to a finite sample of the underlying measure, i.e., we have data $X_1, \ldots, X_n \sim r$. The empirical measure associated with this data is given by

$$\hat{\mathbf{r}}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Similar, $\hat{\mathbf{s}}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{Y_i}$ is based on the sample $Y_1, \ldots, Y_m \sim s$. The most obvious estimate of the Wasserstein distance is the plug-in estimate, i.e., the empirical Wasserstein distance $W_p(\hat{\mathbf{r}}_n, s)$ in the one sample case and $W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m)$ in the two sample case. The natural question that arises is the behavior of the empirical Wasserstein distance ($W_p(\hat{\mathbf{r}}_n, s)$ or $W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m)$) compared to its population version. It is a well known result (see e.g. Villani (2008, Cor. 6.11)) that $W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \to W_p(r, s)$ almost surely if the p-th moments of r and s exist. One can now ask for the rate of convergence or for distributional limits regarding the empirical Wasserstein distance to understand the convergence behavior in more detail. Distributional limits are essential for statistical applications, e.g. in hypothesis testing or in derivation of confidence statements.

1.3 Application of optimal transport - colocalization

In biology, spatial proximity, or colocalization, is an important feature to understand interactions between proteins. The investigation of these protein synergies is a valuable tool as many cellular processes depend on protein networks. Usually, the distribution of proteins is visualized by fluorescence microscopy. The analysis of the images generated by fluorescence microscopy give some indication of the spatial proximity of the proteins under investigation. Commonly the resolution of these images was diffraction limited. This diffraction limit was overcome during the last two decades by super-resolution microscopy (nanoscopy). All nanoscopy techniques are based on an 'on' (emitting photons) and 'off' (dark state) switching of certain fluorophores attached to the proteins under investigation (Sahl et al., 2017; Huang et al., 2009). These nanoscopy methods can be clustered into two major groups - the coordinate-stochastic methods (Betzig et al., 2006; Hess et al., 2006; Rust et al., 2006) and the coordinate-targeted methods (Sahl et al., 2017; Klar et al., 2000; Hofmann et al., 2005). In coordinate-stochastic

nanoscopy the super-resolution is achieved via a sparse subsets of fluorophores which are in the 'on' state. It is assumed that only one fluorophore in a neighborhood below the diffraction limit is in the 'on' state. Hence, it is possible to estimate the locations of the fluorophores from the raw data. Therefore, the data provided by coordinate-stochastic nanoscopy are lists containing fluorophore locations.

Contrary to the coordinate-stochastic methods, coordinated-targeted methods as STED (STimulated Emission Depletion) or RESOLFT (REversible Saturable/Switchable Optical Linear (Fluorescence) Transitions) are based on scanning over the sample while the off-switching in desired spatial coordinates is accomplished with targeted reversible light. Due to the scanning the raw data generated by coordinate-targeted nanoscopy are pixel images that represent the intensities of fluorescence distributions.

The colocalization analysis of such super-resolution data comes with new challenges as the overlap between two different channels (images of protein distributions) is drastically reduced due to the absence of large blurring in super-resolution data. Therefore, we propose a new method based on optimal transport to evaluate the colocalization in STED images. Optimal transport is especially well-suited for this task as it finds the optimal matching between the protein distributions. This matching serves as an indicator for possible interactions between the investigated proteins.

1.4 Literature review and connections to existing work

In this section an overview of literature regarding the relationship between the empirical Wasserstein distance and the population Wasserstein distance is given. Furthermore, existing literature regarding colocalization analysis is presented.

Rate of convergence of empirical Wasserstein distance The beginning of research related to the rate of convergence of the empirical Wasserstein distance started already more than 40 years ago. Ajtai et al. (1984) investigated the rate of convergence of the empirical Wasserstein distance for the uniform measure on the unit square, Talagrand (1992, 1994) extended this to higher dimensions. The two-sample case with equality of the underlying measures for general measures on $[0, 1]^D$ with $D \geq 3$ was derived by Dobrić and Yukich (1995). Horowitz and Karandikar (1994) then provided non-asymptotic bounds for the average speed of convergence for the empirical 2-Wasserstein distance. There are several refinements of these results, e.g. Boissard and Gouic (2014), Fournier and Guillin (2015) and Weed and Bach (2017).

Distributional limits for the Wasserstein distance   As a natural extension of general convergence results, there is a long standing interest in distributional limits for the empirical Wasserstein distance, in particular motivated from statistical applications. Most of this work is restricted to the univariate case $\mathcal{X} \subset \mathbb{R}$. Munk and Czado (1998) derived central limit theorems for a trimmed Wasserstein distance on the real line when $r \neq s$ whereas del Barrio et al. (1999a,b) consider the empirical Wasserstein distance when r belongs to a parametric family of distributions for the assessment of goodness of fit, e.g. for a Gaussian location scale family. In a similar fashion del Barrio et al. (2005) provided asymptotics for a weighted version of the empirical 2-Wasserstein distance in one dimension and Freitag and Munk (2005) derived limit laws for semiparametric models, still restricted to the univariate case. There are also several results for dependent data in one dimension, e.g. Dede (2009), Dedecker and Merlevede (2015). For a recent survey we refer to Bobkov and Ledoux (2016) and Mason (2016) and references therein. A major reason of the limitation to dimension $D = 1$ is that only for $\mathcal{X} \subset \mathbb{R}$ (or more generally a totally ordered space) the coupling which solves (1.9) is known explicitly and can be expressed in terms of the quantile functions F^{-1} and G^{-1} of r and s, respectively, as $\pi = (F^{-1} \times G^{-1})\#\mathcal{L}$, where $\mathcal{L}$ is the Lebesgue measure on $[0, 1]$ (see Mallows (1972)). All the above mentioned work relies essentially on this fact. For higher dimensions only in specific settings such a coupling can be computed explicitly and then be used to derive limit laws (Rippl et al., 2016). Already for $D = 2$ Ajtai et al. (1984) indicate that the scaling rate for the limiting distribution of $W_1(\hat{r}_n, r)$ when r is the uniform measure on $\mathcal{X} = [0, 1]^2$ (if the limiting distribution exists) must be of complicated nature as it is bounded from above and below by a rate of order $\sqrt{n \log(n)}$. Recently, del Barrio and Loubes (2017) gave distributional limits for the quadratic empirical Wasserstein distance in general dimension with a scaling rate $\sqrt{n}$. This yields a (non-degenerate) normal limit in the case $r \neq s$, i.e., when the data generating measure is different from the measure to be compared with (extending Munk and Czado (1998) to $D > 1$). Their result centers the empirical Wasserstein distance with an expected empirical Wasserstein distance (whose value is typically unknown) instead of the true Wasserstein distance and requires r and s to have a positive Lebesgue density on the interior of their convex support. Their proof uses the uniqueness and stability of the optimal transportation potential (i.e., the minimizer of the dual transportation problem (1.4), see Villani (2003) for further information) and the Efron-Stein variance inequality. However, in the case $r = s$, their distributional limit degenerates to a point mass at 0, underlining the fundamental difficulty of this problem again. An alternative approach has been advocated recently in Sommerfeld and Munk (2018) who restrict to finite spaces $\mathcal{X} = \{x_1, \ldots, x_N\}$. They derive limit laws for the empirical

Wasserstein distance for $r = s$ (and $r \neq s$). Under equality of measures they get for the p-th empirical Wasserstein distance the scaling rate $n^{1/2p}$ and in the case of different underlying measures a $\sqrt{n}$ rate.

Colocalization Colocalization analysis of conventional microscopy images is a widely used tool to investigate protein interactions (Landmann and Marbet, 2004; Humpert et al., 2015; Li et al., 2004; Herce et al., 2013; Demandolx and Davoust, 1997; Adler and Parmryd, 2010; Agnati et al., 2005; Bolte and Cordelières, 2006; Eggert et al., 2014; Moser et al., 2017; Neumann et al., 2010; Osterwald et al., 2012; Worz et al., 2010; Xu et al., 2016; Zinchuk and Grossenbacher-Zinchuk, 2009; Zinchuk and Grossenbacher-Zinchuk, 2014; Zinchuk and Zinchuk, 2008). Common methods are visual inspection of the overlay of two colored channels and methods based on pixel intensity correlation (*pixel-based* methods). These pixel-based methods rely on a pixel to pixel comparison between the two images. The most widely used methods based on the correlation principle are Manders' colocalization coefficient, Pearson's correlation and a thresholded version of Pearson's correlation (Costes et al., 2004; Dunn et al., 2011; Manders et al., 1992, 1993; Barlow et al., 2010). All theses coefficients are a measure for the average degree of colocalization. Therefore, detecting correlated regions and evaluating the colocalization in different spatial areas which is accompanied by the evaluation of colocalization on different spatial scales is another relevant topic in this research area (Wang et al., 2016, 2017, 2018).

For data sets recorded with coordinate-stochastic nanoscopy, it is quite common to use colocalization methods based on concepts from spatial statistics (Coltharp et al., 2014; Georgieva et al., 2016; Lagache et al., 2015; Lehmann et al., 2011; Malkusch et al., 2012), including k-nearest neighbors and Ripley's K. These methods are known as *object-based* methods as they operate on the coordinates of the fluorophores.

1.5 Main results

In this section we give an overview of the main results of this thesis regarding the two topics distributional limits for the empirical Wasserstein distance and optimal transport based methods for colocalization analysis.

1.5.1 Distributional limits

The main contribution is a distributional limit result for the empirical Wasserstein distance $W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m)$ on countable spaces $\mathcal{X} = \{x_1, x_2, \ldots\}$ (see Chapter 2). For all results we assume that the empirical measures are formed based on independent and identically

distributed (i.i.d.) data. We further assume that r and s have finite p-th moments, i.e., both measures are elements of $\mathcal{P}_p(\mathcal{X})$ in (1.10). In case that the diameter of the countable metric space $(\mathcal{X}, d)$ is infinite, we need the assumption that r and s fulfill the following condition (respective formula for s)

$$\sum_{x \in \mathcal{X}} d^p(x_0, x) \sqrt{r_x} < \infty, \tag{1.11}$$

where x_0 is some fixed but arbitrary element in $\mathcal{X}$. This condition is extensively discussed in Chapter 2.1.2. Under equality of measures, i.e., $r = s$, we prove that

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2p}} W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \xrightarrow{\mathscr{D}} \left\{\max_{\lambda \in \mathcal{S}^*(r)} \langle \mathbf{G}, \lambda \rangle\right\}^{\frac{1}{p}}.$$

Here, $\xrightarrow{\mathscr{D}}$ indicates weak convergence, $\mathbf{G}$ is a centered Gaussian which is the limit of $\sqrt{n}(\hat{\mathbf{r}}_n - r)$ and $\mathcal{S}^*(r)$ denotes a convex set that is related to the set of dual solutions of the Wasserstein distance (see (2.4)). These results can be used for statistical testing and confidence statements under the null hypothesis of equality of measures.

We also given an analogous result for the case that the underlying measures are different, i.e., $r \neq s$. Further, we consider the one-sample case and derive analogous results in both cases (under equality of measures and different underlying measures).

The distributional limits are derived as consequence of the directional Hadamard differentiability (see 2.3.2) tangentially to the set of measures with finite p-th moment (1.10) of the optimal value of the linear program given in (1.9) in conjunction with a generalized delta method for non-linear derivatives.

All derived limit distributions are given implicitly via a maximization problem. In the case of the underlying ground space $\mathcal{X}$ being a tree and under the assumption of equality of measures we can explicitly calculate this maximum (see Chapter 2.2). This explicit result can be used to upper bound the limiting distribution on general spaces (see Sommerfeld (2017)). The accuracy of this upper bound is numerically investigated.

1.5.2 Colocalization

We derive a new method based on optimal transport - Optimal Transport Colocalization (OTC) - to analyze colocalization in coordinate-targeted super-resolution microscopy. OTC is a pixel-based method and hence, can be directly applied on raw STED data, as the raw data are pixel-images. The major benefit of OTC over the conventional pixel-based coefficients is the capability of capturing colocalization on different scales

simultaneously. In contrast to the widely used colocalization coefficients, OTC represents a curve that measures colocalization in dependency of a spatial parameter. Additionally, we propose methods to analyze the OTC curves from a statistical point of view.

CHAPTER 2

Distributional limits

In this chapter we introduce important notation and derive distributional limit results for the empirical Wasserstein distance on countable spaces. Further, we give explicit limit results in the case that the underlying countable space is given by a weighted tree and use this to derive an upper bound for the limiting distribution on general spaces. The accuracy of this upper bound is investigated in simulations.

2.1 Wasserstein distance on countable metric spaces

Let throughout the following $\mathcal{X} = \{x_1, x_2, \ldots\}$ be a countable metric space equipped with a metric $d \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. The probability measures on $\mathcal{X}$ are infinite dimensional vectors (sequences) r in $\mathcal{P}(\mathcal{X})$ given in (1.7).

We want to emphasize that we consider the discrete topology on $\mathcal{X}$ and do not embed $\mathcal{X}$ for example in $\mathbb{R}^D$. This implies that the support of any probability measure $r \in \mathcal{P}(\mathcal{X})$ is the union of points $x \in \mathcal{X}$ such that $r_x > 0$.

The p-th Wasserstein distance ($p \geq 1$) is given by the p-th root of the optimal value of the linear program in (1.9). As mentioned in the introduction this distance is finite for all measures with finite p-th moment, more precisely for all $r, s \in \mathcal{P}_p(\mathcal{X})$ in (1.10). We need to introduce the weighted ℓ^1-space $\ell^1_{d^p_{x_0}}(\mathcal{X})$ which is defined via the weighted ℓ^1-norm

$$\|r\|_{\ell^1(d^p_{x_0})} = \sum_{x \in \mathcal{X}} d^p(x, x_0) |r_x| + |r_{x_0}| \tag{2.1}$$

with the same $x_0 \in \mathcal{X}$ as in the definition of $\mathcal{P}_p(\mathcal{X})$. The necessity arises due to the fact that the set of probability measures with finite p-th moment is a closed subset of $\ell^1_{d^p_{x_0}}(\mathcal{X})$ and hence complete itself. This will play a crucial role in the proof of the directional Hadamard differentiability (see Section 2.3.2). The weighted ℓ^1-norm (2.1)

can be extended in the following way to sequences on $\mathcal{X} \times \mathcal{X}$ and hence to $\mathcal{P}_p(\mathcal{X} \times \mathcal{X})$

$$\|w\|_{\ell^1(d_{x_0}^p)} = \sum_{x,x' \in \mathcal{X}} d^p(x_0, x) |w_{x,x'}| + |w_{x_0,x'}| + \sum_{x,x' \in \mathcal{X}} d^p(x_0, x') |w_{x,x'}| + |w_{x,x_0}|.$$

2.1.1 Main results

Define the empirical measure generated by i.i.d. random variables $X_1, \ldots, X_n$ from the measure r as

$$\hat{\mathbf{r}}_n = (\hat{r}_{n,x})_{x \in \mathcal{X}}, \text{ where } \hat{r}_{n,x} = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{X_k = x\}}, \tag{2.2}$$

and $\hat{\mathbf{s}}_m$ is defined in the same way by $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} s$. We assume that the collections of random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent. Furthermore, let

$$\ell^1(\mathcal{X}) = \left\{ (a_x)_{x \in \mathcal{X}} \in \mathbb{R}^{\mathcal{X}} : \sum_{x \in \mathcal{X}} |a_x| < \infty \right\}$$

and its dual

$$\ell^\infty(\mathcal{X}) = \left\{ (a_x)_{x \in \mathcal{X}} \in \mathbb{R}^{\mathcal{X}} : \sup_{x \in \mathcal{X}} |a_x| < \infty \right\}.$$

We also require the dual norm of $\| \cdot \|_{\ell^1(d_{x_0}^p)}$, a weighted version of the ℓ^∞-norm to characterize the set of dual solutions:

$$\|a\|_{\ell^\infty(d_{x_0}^{-p})} = \max\left(|a_{x_0}|, \sup_{x \neq x_0 \in \mathcal{X}} |d^{-p}(x, x_0) a_x| \right),$$

for $p \geq 1$. The space $\ell^\infty_{d_{x_0}^{-p}}(\mathcal{X})$ contains all elements which have a finite $\| \cdot \|_{\ell^\infty(d_{x_0}^{-p})}$-norm. This is the dual of the weighted ℓ^1-space $\ell^1_{d_{x_0}^p}(\mathcal{X})$ (see Section 2.3.1 for further explanations).

For $r, s \in \mathcal{P}_p(\mathcal{X})$ we define the following convex sets

$$S^*(r, s) = \Big\{ (\lambda, \mu) \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) \times \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) : \langle r, \lambda \rangle + \langle s, \mu \rangle = W_p^p(r, s)$$
$$\lambda_x + \mu_{x'} \leq d^p(x, x') \quad \forall x, x' \in \mathcal{X} \Big\} \tag{2.3}$$

and

$$S^*(r) = \Big\{ \lambda \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) : \lambda_x - \lambda_{x'} \leq d^p(x, x') \quad \forall x, x' \in \text{supp}(r) \Big\}, \tag{2.4}$$

with $\text{supp}(r) = \{x \in \mathcal{X} : r_x > 0\}$. For our limiting distributions we define the following (multinomial) covariance structure

$$\Sigma(r) = \begin{cases} r_x(1 - r_x) & \text{if } x = x', \\ -r_x r_{x'} & \text{if } x \neq x'. \end{cases} \tag{2.5}$$

Theorem 2.1. *Let $(\mathcal{X}, d)$ be a countable metric space and $r, s \in \mathcal{P}_p(\mathcal{X})$ such that (1.11) holds for some $x_0 \in \mathcal{X}$, $p \geq 1$, and $\hat{\mathbf{r}}_n$ be generated by i.i.d. samples $X_1, ..., X_n \sim r$. Furthermore, let $\mathbf{G} \sim \mathcal{N}(0, \Sigma(r))$ be a Gaussian process with Σ as defined in (2.5). Then*

a) For $r = s$ it holds for $n \to \infty$

$$n^{\frac{1}{2p}} W_p(\hat{\mathbf{r}}_n, r) \xrightarrow{\mathcal{D}} \left\{ \max_{\lambda \in \mathcal{S}^*(r)} \langle \mathbf{G}, \lambda \rangle \right\}^{\frac{1}{p}}. \tag{2.6}$$

b) For $r \neq s$ it holds for $n \to \infty$

$$n^{\frac{1}{2}}(W_p(\hat{\mathbf{r}}_n, s) - W_p(r, s)) \xrightarrow{\mathcal{D}} \frac{1}{p} W_p^{1-p}(r, s) \left\{ \max_{(\lambda, \mu) \in \mathcal{S}^*(r, s)} \langle \mathbf{G}, \lambda \rangle \right\}. \tag{2.7}$$

Note, that we obtain different scaling rates under equality of measures $r = s$ (null-hypothesis) and the case $r \neq s$ (alternative), which has important statistical consequences. For $r \neq s$ we are in the regime of the standard C.L.T. rate $\sqrt{n}$, but for $r = s$ we get the rate $n^{\frac{1}{2p}}$, which is strictly slower for $p > 1$.

Remark 2.2. *a) Note, that in Theorem 2.1 for $r \neq s$ the objective function in (2.7) is independent of the second component μ of the feasible set $\mathcal{S}^*(r, s)$. This is due to the fact that in $W_p(\hat{\mathbf{r}}_n, s)$ the second component is not random.*

b) Observe, that the limit in (2.7) is normally distributed if the set $\mathcal{S}^*(r, s)$ is a singleton up to a constant shift. This is the case if the linear program underlying the definition of the Wasserstein distance (1.9) is non-degenerate. In the case of equality of measures the underlying linear program (1.9) is for all r degenerate and hence the set $\mathcal{S}^*(r)$ is never a singleton up to constant shift.

c) We would like to emphasize that the set of dual solutions $\mathcal{S}^*(r)$ is independent of r, if the support of r is full, i.e.,

$$\mathcal{S}^* = \left\{ \lambda \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) : \lambda_x - \lambda_{x'} \leq d^p(x, x') \quad \forall x, x' \in \mathcal{X} \right\}. \tag{2.8}$$

This offers a universal strategy to simulate the limiting distribution on trees independent of r. For more details see Appendix 2.3.3.

Remark 2.3 (Degeneracy of limit law)**.** *In the case that r has full support and $p > 1$ the limit in (2.6) degenerates to a point mass at zero if $\mathcal{S}^*$ contains only constant elements, i.e., $\lambda_x = c$ for all $x \in \mathcal{X}$ and some $c \in \mathbb{R}$. In this case the variance of $\langle \mathbf{G}, \boldsymbol{\lambda} \rangle$ is given by*

$$\sum_{x \in \mathcal{X}} c^2 \operatorname{Var}(G_x) + \sum_{x,x' \in \mathcal{X}, x \neq x'} c^2 \operatorname{Cov}(G_x, G_{x'})$$

$$= \sum_{x \in \mathcal{X}} c^2 r_x (1 - r_x) - \sum_{x,x' \in \mathcal{X}, x \neq x'} c^2 r_x r_{x'}$$

$$= \sum_{x \in \mathcal{X}} c^2 r_x (1 - r_x) - \sum_{x \in \mathcal{X}} c^2 r_x \sum_{x' \in \mathcal{X}, x' \neq x} r_{x'}$$

$$= \sum_{x \in \mathcal{X}} c^2 r_x (1 - r_x) - \sum_{x \in \mathcal{X}} c^2 r_x (1 - r_x) = 0.$$

Hence, as the variance is zero for all elements in $\mathcal{S}^*$ so is the variance of the maximum and this yields that the right hand side in (2.6) is a dirac measure at zero.
The set of dual solutions $\mathcal{S}^*$ contains only constant elements if and only if the space $\mathcal{X}$ has no isolated point. To see this recall the definition of $\mathcal{S}^*$ for full support of r in (2.8)

$$\mathcal{S}^* = \left\{ \boldsymbol{\lambda} \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) : \lambda_x - \lambda_{x'} \leq d^p(x, x') \quad \forall x, x' \in \mathcal{X} \right\}.$$

The condition that defines $\mathcal{S}^*$ is equivalent to

$$|\lambda_x - \lambda_{x'}| \leq d^p(x, x').$$

For fixed $x \in \mathcal{X}$ there exits a sequence $(x'_n)_{n \in \mathbb{N}}$ such that the distance $d(x, x'_n)$ gets arbitrary small if and only if x is not an isolated point. If this holds for all $x \in \mathcal{X}$, i.e., the space $\mathcal{X}$ contains no isolated point, the dual solution $\boldsymbol{\lambda}$ has to be constant.
Now, the question arises if there exists another scaling rate than $n^{\frac{1}{2p}}$ such that the limit is not degenerated. This question can be answered for $\mathcal{X}$ being a subset of the real line $\mathbb{R}$ that has no isolated point as it follows from Theorem 7.11. in Bobkov and Ledoux (2016) that scaling with $\sqrt{n}$ provides then a non-degenerate limit law. On the other hand, as soon as $\mathcal{X} \subset \mathbb{R}$ contains an isolated point our rate coincides with the rate given in Bobkov and Ledoux (2016).

For statistical applications it is also interesting to consider the two sample case, extensions to k-samples, $k \geq 2$ being obvious then.

Theorem 2.4. *Under the same assumptions as in Theorem 2.1 and with $\hat{\mathbf{s}}_m$ generated by $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} \mathbf{s}$, independently of $X_1, \ldots, X_n \overset{i.i.d.}{\sim} r$ and $\mathbf{H} \sim \mathcal{N}(0, \Sigma(\mathbf{s}))$, which is independent of $\mathbf{G}$, and the extra assumption that $\mathbf{s}$ also fulfills (1.11) the following holds.*

a) *Let $\rho_{n,m} = (nm/(n+m))^{1/2}$. For $\mathbf{r} = \mathbf{s}$ and $\min(n, m) \to \infty$ such that $m/(n+m) \to \alpha \in [0, 1]$ we have*

$$\rho_{n,m}^{1/p} W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \xrightarrow{\mathscr{D}} \left\{ \max_{\lambda \in \mathcal{S}^*(\mathbf{r})} \langle \mathbf{G}, \lambda \rangle \right\}^{\frac{1}{p}}. \tag{2.9}$$

b) *For $\mathbf{r} \neq \mathbf{s}$ and $n, m \to \infty$ such that $\min(n, m) \to \infty$ and $m/(n+m) \to \alpha \in [0, 1]$ we have*

$$\rho_{n,m}(W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) - W_p(\mathbf{r}, \mathbf{s})) \xrightarrow{\mathscr{D}}$$
$$\frac{1}{p} W_p^{1-p}(\mathbf{r}, \mathbf{s}) \left\{ \max_{(\lambda, \mu) \in \mathcal{S}^*(\mathbf{r}, \mathbf{s})} \sqrt{\alpha} \langle \mathbf{G}, \lambda \rangle + \sqrt{1-\alpha} \langle \mathbf{H}, \mu \rangle \right\}. \tag{2.10}$$

Remark 2.5. *In the case of dependent data analogous results to Theorem 2.1 and 2.4 will hold, as soon as the weak convergence of the empirical process w.r.t. the $\| \cdot \|_{\ell^1(d_{x_0}^p)}$-norm is valid. All other steps of the proof remain unchanged.*

The rest of this subsection is devoted to the proofs of Theorem 2.1 and Theorem 2.4. To prove these two theorems we use the delta method for mappings that are directionally Hadamard differentiable tangentially to a set (see Definition 2.20).

Theorem 2.6 (Römisch (2004), Theorem 1)**.** *Let $\mathcal{U}$ and $\mathcal{Y}$ be normed spaces and K be a subset of $\mathcal{U}$, $f: K \to \mathcal{Y}$ a mapping and assume that the following two conditions are satisfied:*

i) *The mapping f is Hadamard directionally differentiable at $u \in K$ tangentially to K with derivative $f_u'(\cdot): T_K(u) \to \mathcal{Y}$.*

ii) *For each n, $X_n: \Omega_n \to K$ are maps such that $a_n(X_n - u) \xrightarrow{\mathscr{D}} X$ for some sequence $a_n \to +\infty$ and some random element X.*

Then we have $a_n(f(X_n) - f(u)) \xrightarrow{\mathscr{D}} f_u'(X)$.

Proof of Theorem 2.1 and Theorem 2.4. To use the delta method, we need to verify (1.) directional Hadamard differentiability of $W_p^p(\cdot, \cdot)$ and (2.) weak convergence of $\sqrt{n}(\hat{\mathbf{r}}_n - \mathbf{r})$. We mention that the delta method required here is not standard as the directional Hadamard derivative is not linear (see Römisch (2004), Shapiro (1991) or Dümbgen (1993)).

1. Theorem 2.21 in Section 2.3.2 proves directional Hadamard differentiability of W_p^p tangentially to the set of probability measures with finite p-th moment ($\mathcal{P}_p(\mathcal{X})$) with respect to the $\| \cdot \|_{\ell^1(d_{x_0}^p)}$-norm (2.1).

2. The weak convergence of the empirical process w.r.t. the $\| \cdot \|_{\ell^1(d_{x_0}^p)}$-norm is addressed in the following lemma.

Lemma 2.7. *Let $X_1, \ldots, X_n \sim r$ be i.i.d. taking values in a countable metric space (X, d) and let $\hat{\mathbf{r}}_n$ be the empirical measure as defined in (2.2). Then,*

$$\sqrt{n}(\hat{\mathbf{r}}_n - r) \xrightarrow{\mathscr{D}} \mathbf{G}$$

with respect to the $\| \cdot \|_{\ell^1(d_{x_0}^p)}$-norm if and only if condition (1.11) is fulfilled. Here, $\mathbf{G}$ is a Gaussian process with mean 0 and covariance structure

$$\Sigma(r) = \begin{cases} r_x(1 - r_x) & \text{if } x = x', \\ -r_x r_{x'} & \text{if } x \neq x', \end{cases}$$

as given in (2.5).

Proof of Lemma. The weighted ℓ^1-space $\ell^1_{d_{x_0}^p}$ is according to Proposition 3, Maurey (1972) of cotype 2. Hence, $\sqrt{n}(\hat{\mathbf{r}}_n - r)$ converges weakly w.r.t. the $\ell^1(d_{x_0}^p)$-norm by Corollary 1 in Jain (1976) if and only if the summability condition (1.11) is fulfilled. $\square$

Theorem 2.1 a) is now a straight forward application of the delta method 2.6 and the continuous mapping theorem for $f(x) = x^{1/p}$.

For Theorem 2.1 b) we use again the delta method, but this time in combination with the chain rule for directional Hadamard differentiability (Proposition 3.6 (i), Shapiro (1990)).

The proof of Theorem 2.4 works analogously. Note, that under the assumptions of the theorem it holds in the case of $r = s$

$$\rho_{n,m}((\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) - (r, s)) = \left(\sqrt{\frac{m}{n+m}} \sqrt{n}(\hat{\mathbf{r}}_n - r), \sqrt{\frac{n}{n+m}} \sqrt{m}(\hat{\mathbf{s}}_m - s) \right)$$

$$\xrightarrow{\mathscr{D}} (\sqrt{\alpha}\mathbf{G}, \sqrt{1-\alpha}\mathbf{G}') \quad (2.11)$$

with $\mathbf{G}' \overset{\mathcal{D}}{=} \mathbf{G}$.

The delta method together with the continuous mapping theorem and equation (2.11) gives

$$\rho_{n,m}^{1/p} W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \xrightarrow{\mathscr{D}} \left\{ \max_{(\lambda,\mu)\in\mathcal{S}^*(r,r)} \sqrt{\alpha}\langle \lambda, \mathbf{G} \rangle + \sqrt{1-\alpha}\langle \mu, \mathbf{G}' \rangle \right\}^{1/p} .$$

Nevertheless, for all $x \in X$ where $r_x > 0$ it holds $\lambda_x = -\mu_x$ and for all $x \in X$ where $r_x = 0$ the limit element G_x is degenerate. Hence, the limit distribution above is equivalent in

distribution to

$$\left\{ \max_{\lambda \in \mathcal{S}^*(r,r)} \sqrt{\alpha} \langle \boldsymbol{\lambda}, \mathbf{G} \rangle - \sqrt{1-\alpha} \langle \boldsymbol{\lambda}, \mathbf{G}' \rangle \right\}^{1/p}.$$

The independence of $\mathbf{G}$ and $\mathbf{G}'$ yields that $\sqrt{\alpha} \langle \boldsymbol{\lambda}, \mathbf{G} \rangle - \sqrt{1-\alpha} \langle \boldsymbol{\lambda}, \mathbf{G}' \rangle$ equals $\sqrt{\alpha + (1-\alpha)} \langle \boldsymbol{\lambda}, \mathbf{G} \rangle$ in distribution and hence the limit reduces to

$$\left\{ \max_{\lambda \in \mathcal{S}^*(r)} \langle \boldsymbol{\lambda}, \mathbf{G} \rangle \right\}^{1/p}.$$

$\square$

2.1.2 Examination of the summability condition (1.11)

According to Lemma 2.7 condition (1.11) is necessary and sufficient for the weak convergence with respect to the $\| \cdot \|_{\ell^1(d^p_{x_0})}$-norm defined in (2.1). As this condition is crucial for our main theorem on spaces that have an unbounded diameter and we are not aware of a comprehensive discussion, we will provide one in this section. As this condition is not needed in the case of bounded diameter (see Section 2.1.4) we will assume throughout this section that the diameter of $\mathcal{X}$ is infinite.

The first question to investigate is whether this condition it valid for all x_0 if it is valid for one x_0. Contrary to the independence of the space $\mathcal{P}_p(\mathcal{X})$ of the choice of x_0 the summability condition (1.11) is in general not independent of the choice of x_0. However, in the case that $\mathcal{X}$ has no accumulation point, i.e., is discrete in the topological sense, and the unit balls are totally bounded the condition is independent of the choice of x_0.

Lemma 2.8. *Let $\mathcal{X}$ be a space without any accumulation point with respect to the metric d and assume that the unit ball, $B_1(x) = \{x' \in \mathcal{X} \colon d(x, x') \leq 1\}$, is totally bounded for each $x \in \mathcal{X}$. If condition (1.11) holds for one $x_0 \in \mathcal{X}$ then it holds for all x_0.*

Proof. Let the condition (1.11) be fulfilled for x_0 and let $\tilde{x}_0$ be another element in $\mathcal{X}$. Then, it holds

$$\sum_{x \in \mathcal{X}} d^p(\tilde{x}_0, x) \sqrt{r_x} \leq \sum_{x \in \mathcal{X}} 2^{p-1} (d^p(\tilde{x}_0, x_0) + d^p(x_0, x)) \sqrt{r_x}$$

$$\leq 2^{p-1} \left(\sum_{x \in \mathcal{X}} d^p(\tilde{x}_0, x_0) \sqrt{r_x} + \sum_{x \in \mathcal{X}} d^p(x_0, x) \sqrt{r_x} \right).$$

The second sum is finite due to condition (1.11). The first sum can be handled as follows.

$$\sum_{x \in \mathcal{X}} \sqrt{r_x} \leq \sum_{x \in B_1(x_0)} \sqrt{r_x} + \sum_{x \in B_1^C(x_0)} d^p(x_0, x) \sqrt{r_x},$$

here, the second sum is again finite due to condition (1.11). The unit ball $B_1(x_0)$ is complete and totally bounded and hence, compact. This yields that it can only contain finitely many points and the first sum is finite. From theses observations, the claim follows. $\qquad\square$

In the case that x_0 is not an accumulation point another property holds.

Lemma 2.9. *Let $x_0 \in \mathcal{X}$ an isolated point with respect to the metric d. If condition (1.11) holds for p, then it also holds for all $1 \le p' \le p$.*

Proof. As x_0 is an isolated point there exists $\epsilon > 0$ such that $d(x, x_0) > \epsilon$ for all $x \ne x_0 \in \mathcal{X}$. Then,

$$\sum_{x \in \mathcal{X}} d^p(x_0, x) \sqrt{r_x} = \epsilon^p \sum_{x \in \mathcal{X}} \left(\frac{d(x_0, x)}{\epsilon}\right)^p \sqrt{r_x}$$

$$\ge \epsilon^p \sum_{x \in \mathcal{X}} \left(\frac{d(x_0, x)}{\epsilon}\right)^{p'} \sqrt{r_x}$$

$$= \epsilon^{p/p'} \sum_{x \in \mathcal{X}} d^{p'}(x_0, x) \sqrt{r_x}.$$

$\qquad\square$

Exponential families   As we will see, condition (1.11) is fulfilled for many well known distributions including the Poisson distribution, geometric distribution or negative binomial distribution with the Euclidean distance as the metric d on the ground space $\mathcal{X} = \mathbb{N}$.

Theorem 2.10. *Let $(\mathcal{P}_\eta)_\eta$ be an s-dimensional standard exponential family (SEF) (see Lehmann and Casella (1998), Sec. 1.5) of the form*

$$r_x^\eta = h_x \exp\left(\sum_{i=1}^{s} \eta_i T_x^i - A(\eta)\right). \tag{2.12}$$

The summability condition (1.11) is fulfilled if $(\mathcal{P}_\eta)_\eta$ satisfies

1.) *$h_x \ge 1$ for all $x \in \mathcal{X}$,*

2.) *the natural parameter space $\mathcal{N}$ is closed with respect to multiplication by $\frac{1}{2}$, i.e., $\sum_{x \in \mathcal{X}} r_x^\eta < \infty \Rightarrow \sum_{x \in \mathcal{X}} r_x^{\eta/2} < \infty$,*

3.) *the p-th moment w.r.t. the metric d on $\mathcal{X}$ exists, i.e., $\sum_{x \in \mathcal{X}} d^p(x, x_0) r_x^\eta < \infty$ for some arbitrary, but fixed $x_0 \in \mathcal{X}$.*

Proof. For the SEF in (2.12) condition (1.11) reads

$$
\begin{aligned}
\sum_{x \in X} d^p(x_0, x) &\sqrt{\exp\left(\sum_{i=1}^{s} \eta_i T_x^i - A(\boldsymbol{\eta})\right) h_x} \\
&= \frac{1}{\sqrt{\lambda(\boldsymbol{\eta})}} \sum_{x \in X} d^p(x_0, x) \exp\left(\tfrac{1}{2} \sum_{i=1}^{s} \eta_i T_x^i\right) \sqrt{h_x} \qquad (2.13) \\
&\leq \frac{\lambda(\tfrac{1}{2}\boldsymbol{\eta})}{\sqrt{\lambda(\boldsymbol{\eta})}} \sum_{x \in X} d^p(x_0, x) \exp\left(\tfrac{1}{2} \sum_{i=1}^{s} \eta_i T_x^i\right) h_x < \infty,
\end{aligned}
$$

where $\lambda(\boldsymbol{\eta})$ denotes the Laplace transform. The first inequality is due to the fact that $h_x \geq 1$ for all $x \in X$ and the second is a result of the facts that the natural parameter space is closed with respect to multiplication with $\tfrac{1}{2}$ and that the p-th moment w.r.t. d exist. $\qquad\square$

The following examples show, that all three conditions in Theorem 2.10 are necessary. If the p-th moment does not exist, it is clear that condition (1.11) cannot be fulfilled as $\sqrt{x} \geq x$ for $x \in [0, 1]$.

Example 2.11. *Let X be the countable metric space $X = \left\{\tfrac{1}{k}\right\}_{k \in \mathbb{N}}$ and let $\boldsymbol{r}$ be the measure with probability mass function*

$$
r_{1/k} = \frac{1}{\zeta(\eta)} \frac{1}{k^{\eta}}
$$

with respect to the counting measure. Here, $\zeta(\eta)$ denotes the Riemann zeta function. This is an SEF with natural parameter η, natural statistic $-\log(k)$ and natural parameter space $\mathcal{N} = (1, \infty)$. We choose the Euclidean distance as the distance d on our space X and set $x_0 = 1$. It holds

$$
\sum_{k=1}^{\infty} \left|1 - \frac{1}{k}\right|^p \frac{1}{\zeta(\eta)} \frac{1}{k^{\eta}} \leq \sum_{k=1}^{\infty} \frac{1}{\zeta(\eta)} \frac{1}{k^{\eta}} = 1 < \infty \quad \forall \eta \in \mathcal{N}
$$

and hence all moments exist for all η in the natural parameter space. Furthermore, $h_{1/k} \equiv 1$. However, the natural parameter space is not closed with respect to multiplication by $\tfrac{1}{2}$ and therefore,

$$
\sum_{k=1}^{\infty} \left|1 - \tfrac{1}{k}\right|^p \frac{1}{\zeta(\eta)} \frac{1}{k^{\eta/2}} \geq \frac{1}{2^p} \sum_{k=2}^{\infty} \frac{1}{\sqrt{\zeta(\eta)}} \frac{1}{k^{\eta/2}} = \infty \quad \forall \eta \in (1, 2],
$$

i.e., condition (1.11) is not fulfilled.

The next example shows, that we cannot omit condition 1.) in Theorem 2.10.

Example 2.12. *Consider $X = \mathbb{N}$ with the metric $d(k, l) = \sqrt{|k! - l!|}$. The family of Poisson distributions constitute an SEF with natural parameter space $\mathcal{N} = (-\infty, \infty)$ which satisfies condition 2.) in Theorem 2.10, i.e., closed with respect to multiplication by $\frac{1}{2}$. The first moment with respect to this metric exists and $h_k < 1$ for all $k \geq 2$. Condition (1.11) for $p = 1$ with $x_0 = 0$ reads*

$$\sum_{k=1}^{\infty} \sqrt{k!} \sqrt{\frac{\eta^k}{k!} \exp(-\eta)} = \sum_{k=1}^{\infty} \eta^{k/2} \exp(-\eta/2) = \infty$$

for all $\eta > 1$, i.e., the summability condition (1.11) is not fulfilled.

2.1.3 Approximation of continuous distributions

In this section we investigate to what extent we can approximate continuous measures by its discretization such that condition (1.11) remains valid. Let $X = \left(\frac{k}{M}\right)_{k \in \mathbb{Z}}$ with $M \in \mathbb{N}$ be a discretization of $\mathbb{R}$ and X a real-valued random variable with c.d.f. F which is continuous and has a Lebesgue density f. We take d to be the Euclidean distance and $x_0 = 0$. For $k \in \mathbb{Z}$ we define

$$r_k := F\left(\frac{k+1}{M}\right) - F\left(\frac{k}{M}\right). \tag{2.14}$$

Now, (1.11) can be estimated as follows.

$$
\begin{aligned}
&\sum_{k=-\infty}^{\infty} \left|\frac{k}{M}\right|^p \sqrt{F\left(\frac{k+1}{M}\right) - F\left(\frac{k}{M}\right)} \\
&= \sum_{k=-\infty}^{\infty} \left|\frac{k}{M}\right|^p \frac{1}{\sqrt{M}} \sqrt{M \int_{k/M}^{(k+1)/M} f(x)dx} \\
&\geq \sum_{k=-\infty}^{\infty} \left|\frac{k}{M}\right|^p \sqrt{M} \int_{k/M}^{(k+1)/M} \sqrt{f(x)}dx \\
&\geq \sqrt{M} \sum_{k=-\infty}^{\infty} \frac{1}{2^p} \int_{k/M}^{(k+1)/M} |x|^p \sqrt{f(x)}dx \\
&= \sqrt{M} \frac{1}{2^p} \int_{\mathbb{R}} |x|^p \sqrt{f(x)}dx,
\end{aligned}
$$

where the first inequality is due to Jensen's inequality. As the r.h.s. tends to infinity with rate $\sqrt{M}$ as $M \to \infty$, condition (1.11) does not hold in the limit. Hence, in general our method of proof cannot be extended in an obvious way to continuous measures.

The one-dimensional case D = 1 For the rest of this section we consider $X = \mathbb{R}$ and want to put condition (1.11) in relation to the condition (del Barrio et al., 1999b)

$$\int_{-\infty}^{\infty} \sqrt{F(t)(1 - F(t))}dt < \infty, \tag{2.15}$$

where $F(t)$ denotes the cumulative distribution function, which is sufficient and necessary for the empirical 1-Wasserstein distance on $\mathbb{R}$ to satisfy a limit law (see also Corollary 1 in Jain (1976) in a more general context).

Condition (1.11) is stronger than (2.15) as the following shows. Let X be a countable subset of $\mathbb{R}$ such that it can be ordered indexed by $\mathbb{Z}$. Furthermore, let $d(x, y) = |x - y|$ be the Euclidean distance on X. For any measure r with cumulative distribution function F on X it holds

$$\int_{-\infty}^{\infty} \sqrt{F(t)(1 - F(t))}dt$$

$$= \sum_{k \in \mathbb{Z}} d(x_k, x_{k+1}) \sqrt{\sum_{j \leq k} r_j} \sqrt{\sum_{j > k} r_j}$$

$$\leq \sum_{k=0}^{\infty} d(x_k, x_{k+1}) \sqrt{\sum_{j > k} r_j} + \sum_{k=-\infty}^{-1} d(x_k, x_{k+1}) \sqrt{\sum_{j \leq k} r_j}$$

$$\leq \sum_{k=0}^{\infty} d(x_k, x_{k+1}) \sum_{j > k} \sqrt{r_j} + \sum_{k=-\infty}^{-1} d(x_k, x_{k+1}) \sum_{j \leq k} \sqrt{r_j}$$

$$= \sum_{k=0}^{\infty} d(x_0, x_k) \sqrt{r_k} + \sum_{k=-\infty}^{-1} d(x_0, x_k) \sqrt{r_k}.$$

Hence, if condition (1.11) holds, (2.15) is also fulfilled. However, the conditions are not equivalent as the following example shows.

Example 2.13. *Let $X = \mathbb{N}$ and $d(x, y) = |x - y|$ the Euclidean distance and r a power-law, i.e., $r_n = \frac{1}{\zeta(s)} \frac{1}{n^s}$, where $\zeta(s)$ is the Riemann zeta function. In this case (2.15) reads*

$$\int_{-\infty}^{\infty} \sqrt{F(t)(1 - F(t))}dt = \frac{1}{\zeta(s)} \sum_{k=1}^{\infty} \sqrt{\sum_{j=1}^{k} \frac{1}{j^s} \sum_{j=k+1}^{\infty} \frac{1}{j^s}}$$

$$\leq \frac{1}{\zeta(s)} \sum_{k=1}^{\infty} \sqrt{\sum_{j=k}^{\infty} \frac{1}{j^s}} \lesssim \frac{1}{\zeta(s)} \sum_{k=1}^{\infty} \sqrt{\frac{s}{k^{s-1}}}$$

and this is finite if and only if $s > 3$. On the other hand, condition (1.11) reads as

$$\sum_{k=1}^{\infty} (k - 1) \sqrt{\frac{1}{\zeta(s)} \frac{1}{k^s}} \leq \frac{1}{\sqrt{\zeta(s)}} \sum_{k=1}^{\infty} \frac{1}{k^{s/2-1}}.$$

This is finite if and only if s > 4. Hence, for s > 3 condition (2.15) *is fulfilled while condition* (1.11) *is only valid for s > 4.*

For $p = 2$ in dimension $D = 1$ there is no such easy condition anymore in the case of continuous measures, see del Barrio et al. (2005). Already for the normal distribution one needs to subtract a term that tends sufficiently fast to infinity to get a distributional limit (which was originally proven by de Wet and Venter (1972)). Nevertheless, for a fixed discretization of the normal distribution via binning as in (2.14) condition (1.11) is fulfilled and Theorems 2.1 and 2.4 are valid.

2.1.4 Bounded diameter of $\mathcal{X}$

A tremendous simplification is obtained once $\mathcal{X}$ is assumed to have bounded diameter. First and most important, we do not need to introduce the spaces $\ell^1_{d^p_{x_0}}(\mathcal{X})$ and its dual $\ell^\infty_{d^{-p}_{x_0}}(\mathcal{X})$ in this case. This is due to the fact, that as the diameter of the space with respect to the metric d is bounded all moments of probability measures on this space exist. Hence, we do not need to restrict to probability measures that have finite p-th moment to guarantee that the linear program (2.23) defining the Wasserstein distance has a finite value. Thus, we can operator on $\mathcal{P}(\mathcal{X})$ which is a subset of $\ell^1(\mathcal{X})$. This simplifies the summability condition (1.11) to

$$\sum_{x \in \mathcal{X}} \sqrt{r_x} < \infty$$

as we get directional Hadamard differentiability with respect to the $\| \cdot \|_1$-norm.

2.2 Limiting distribution for tree metrics

2.2.1 Explicit limits

In this subsection we give an explicit expression for the limiting distribution under equality of measures, i.e., in the case $r = s$, given in (2.6) and (2.9) when the metric is generated by a weighted tree. Further, we assume that r has full support (otherwise see Rem. 2.15).

Throughout this section, we assume that the underlying ground space $\mathcal{X}$ is a tree. More precisely, we assume $\mathcal{X}$ to be the vertices of an undirected, connected graph $\mathcal{T} = (\mathcal{X}, E)$ that has no cycles. Here, $E \subset \mathcal{X} \times \mathcal{X}$ denotes the set of edges, i.e., connections between two elements in $\mathcal{X}$. The non-existence of cycles means that we cannot find edges $e_1, \cdots, e_n$ such that there exists a path from $x \in \mathcal{X}$ to itself. Further, we assume that the

edges are weighted by a weight function

$$w : E \to \mathbb{R}_+.$$

Additionally, we assume that our tree $\mathcal{T}$ is rooted at $\text{root}(\mathcal{T}) \in \mathcal{X}$. Then, the parent $(\text{parent}(x) \in \mathcal{X})$ of $x \in \mathcal{X}$ with $x \neq \text{root}(\mathcal{T})$ is given as the immediate neighbor of x in the unique path connecting x and $\text{root}(\mathcal{T})$. For the root, we set $\text{parent}(\text{root}(\mathcal{T})) = \text{root}(\mathcal{T})$. Furthermore, $\text{children}(x)$ is defined as the set of vertices $x' \in \mathcal{X}$ such that there exists a sequence $x' = x_1, \ldots, x_n = x \in \mathcal{X}$ with $\text{parent}(x_j) = x_{j+1}$ for $j = 1, \ldots, n-1$. Note that with this definition $x \in \text{children}(x)$. Observe that $\text{children}(x)$ can consist of countably many elements, but the path joining x and $x' \in \text{children}(x)$ is still finite. Let $A_0 = \{x \in \mathcal{X} : x = \text{root}(\mathcal{T})\}$, $A_1 = \{x \in \mathcal{X} : \text{parent}(x) \in A_0\} \setminus \{\text{root}(\mathcal{T})\}$ and $A_k = \{x \in \mathcal{X} : \text{parent}(x) \in A_{k-1}\}$ for $k \geq 2 \in \mathbb{N}$. By the definition of the A_k, these sets are disjoint and it follows $\bigcup_{k=0}^{\infty} A_k = \mathcal{X}$. Now let $x, x' \in \mathcal{X}$, then there exist k_1 and k_2 such that $x \in A_{k_1}$ and $x' \in A_{k_2}$. Then, there is a sequence of $k_1 + k_2 + 1$ vertices connecting x and x'. Hence, the unique path joining x and x' has at most $k_1 + k_2$ edges.

The metric $d_{\mathcal{T}}$ on $\mathcal{X}$ is given by the length of the unique path joining two elements x, x' in $\mathcal{X}$. More precisely,

$$d_{\mathcal{T}}(x, x') = \sum_{j=1}^{n} w(e_j),$$

where $e_1, \ldots, e_n \in E$ is the unique path in $\mathcal{T}$ joining x and x'. This metric is well defined, since the unique path joining x and x' is finite as we have shown above. Additionally, define

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

and

$$Z_{\mathcal{T},p}(u) = \left\{ \sum_{x \in \mathcal{X}} |(S_{\mathcal{T}} u)_x| d_{\mathcal{T}}(x, \text{parent}(x))^p \right\}^{\frac{1}{p}} \tag{2.16}$$

for $u \in \ell^1_{d^p_{x_0}}(\mathcal{X})$ and we set w.l.o.g. $x_0 = \text{root}(\mathcal{T})$. The main result of this section is the following.

Theorem 2.14. *Let $r \in \mathcal{P}_p(\mathcal{X})$ be a probability distribution on $\mathcal{X}$ that has full support and fulfills condition (1.11). Further, let the empirical measures $\hat{r}_n$ and $\hat{s}_m$ be generated by independent random variables $X_1, \ldots, X_n$ and $Y_1, \ldots Y_m$ independent of the X_i's, respectively, all drawn from $r = s$.*

Then, with a Gaussian process $G \sim \mathcal{N}(0, \Sigma(r))$ with $\Sigma(r)$ as defined in (2.5) we have the following.

(a) full tree   (b) tree reduced to support
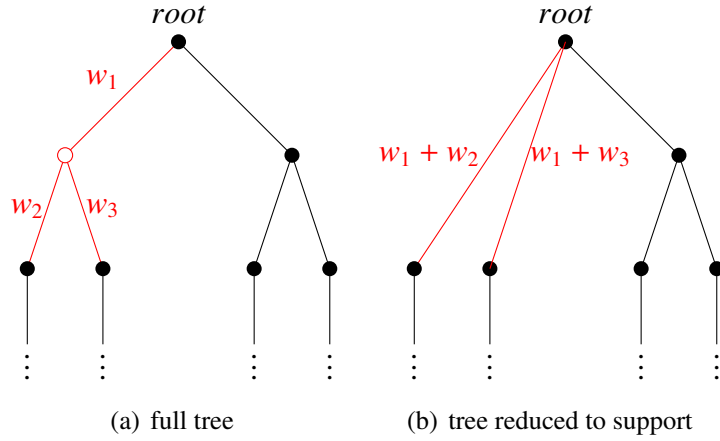
Figure 2.1: Scheme for the reduction of $\mathcal{X}$ to the support of r. Solid circles indicate support points, hollow circles elements which are not in the support.

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

$$n^{\frac{1}{2p}} W_p(\hat{\mathbf{r}}_n, r) \xrightarrow{\mathscr{D}} Z_{\mathcal{T},p}(\mathbf{G}) \tag{2.17}$$

b) *(Two sample)* If $n \wedge m \to \infty$ and $n/(n+m) \to \alpha \in [0,1]$ we have

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2p}} W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \xrightarrow{\mathscr{D}} Z_{\mathcal{T},p}(\mathbf{G}). \tag{2.18}$$

A rigorous proof of Theorem 2.14 is given in Section 2.3.4.

The same result was derived in Sommerfeld and Munk (2018) for finite spaces. For $\mathcal{X}$ countable we require a different technique of proof. Simplifying the set of dual solutions in the same way, the second step of rewriting the target function with a summation and difference operator does not work in the case of measures with countable support, since the inner product of the operators applied to the parameters is no longer well defined. For this setting we need to introduce a new basis in $\ell^1_{d^p_{x_0}}(\mathcal{X})$ and for each element $r \in \ell^1_{d^p_{x_0}}(\mathcal{X})$ a sequence which has only finitely many non-zeros that converges to r in order to obtain an upper bound on the optimal value. Then, we define a feasible solution for which this upper bound is attained.

Remark 2.15. *In case that the support is not full we can generate a weighted tree for the support points in the following way. If x is not in the support of r we delete x and connect* parent(x) *to all nodes in the set* $A_{+1}(x) = \{x' \in \mathcal{X}: \text{parent}(x') = x\}$ *with edges that have the length of the sum of the edge joining x and* parent(x) *and the edge joining* $x' \in A_{+1}$ *and x. Then, we can use the same arguments as in the case of full support to derive the explicit limit on the restricted tree. This is an upper bound of the limiting*

distribution on the full tree with non full support. See Figure 2.1 for an illustration.

2.2.2 Distributional bound for the limiting distribution

In this section we use the explicit formula on the r.h.s. of (2.17) for the case of tree metrics to stochastically bound the limiting distribution on a general space $\mathcal{X}$ which is not a tree. As mentioned in the preface this distributional bound was derived by Max Sommerfeld (Sommerfeld, 2017).

For a finite space $\mathcal{X} = \{x_1, \ldots, x_N\}$ a spanning tree $\mathcal{T}$ is a rooted tree with vertices given by the elements of $\mathcal{X}$ and edges such that any two vertices are connected by exactly one path. The length of this path defines the tree metric $d_{\mathcal{T}}$.

Theorem 2.16 (Sommerfeld (2017), Theorem 10). *Let $r, s \in \mathcal{P}(\mathcal{X})$ and let $\hat{r}_n$, $\hat{s}_m$ be generated by i.i.d. $X_1, \ldots, X_n \sim r$ and $Y_1, \ldots, Y_m \sim s$, respectively. Let further $\mathcal{T}$ be a spanning tree of $\mathcal{X}$. Then, if $r = s$ we have, as n and m approach infinity such that $n \wedge m \to \infty$ and $n/(n + m) \to \alpha \in [0, 1]$, that*

$$\limsup_{n,m\to\infty} P\left[\left(\frac{nm}{n+m} \right)^{1/2p} W_p(\hat{r}_n, \hat{s}_m) \geq z \right] \leq P\left[Z_{\mathcal{T},p}(\mathbf{G}) \geq z \right], \qquad (2.19)$$

where $\mathbf{G} \sim \mathcal{N}(0, \Sigma(r))$ (multivariate Gaussian) with $\Sigma(r)$ as defined in (2.5).

In order to investigate the accuracy of this upper bound, we simulate the true limiting distribution and the upper bound for three different spanning trees on an equidistant grid in $[0, 1]^2$ for different grid sizes.

The spanning trees considered are the dyadic partition, the chain and the fork, see Figure 2.2. For the dyadic partition a grid with 2^k points on each side is chosen. Further, for $0 \leq l \leq k$ let P_l be the natural partition of the grid $\mathcal{X}$ into 2^{2l} squares with $2^{2k}/2^{2l}$ points. We enlarge the space $\mathcal{X}$ by the center points of all P_l and identify the center points in P_k with the points in $\mathcal{X}$. The enlarged space is denoted by $\mathcal{X}'$. A probability measure r on $\mathcal{X}$ can be naturally extended to a probability measure on $\mathcal{X}'$ by giving zero mass to all center points that are not in $\mathcal{X}$. A tree on $\mathcal{X}'$ can be build as follows. The parent of a center point $C \in P_l$ is the center point of the unique set in P_{l-1} that contains C. The root of this tree is the center point of all points in $\mathcal{X}$, see Figure 2.2 (c). The spanning tree 'Chain' (see Figure 2.2 (a)) is constructed by taking the top left pixel as the root and going in a slalom through all other nodes. The top left pixel is also the root of the spanning tree 'Fork' (see Figure 2.2 (b)), having only branchings at each node in the most left column of the grid $\mathcal{X}$.

We investigate the behavior of the upper bound (2.19) for five different probability measures, the uniform measure, a random measure, i.e., a realization of the Dirichlet
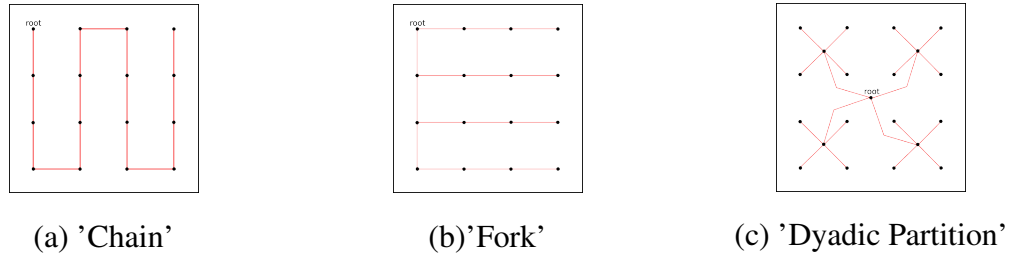
(a) 'Chain'   (b)'Fork'   (c) 'Dyadic Partition'

Figure 2.2: **Three different spanning trees on a** 4×4 **grid.** The black rectangle depicts the unit square $[0, 1]^2$, the dots indicate the locations which represent the pixels, i.e., $\mathcal{X}$ (the vertices of the tree) and the red lines indicate the edges.

distribution $Dir(1)$ and three versions of a discretized bivariate Gaussian. For the discretized bivariate Gaussian we took as weights the density of the bivariate Gaussian at the points in the grid normalized so that their sum is one. For the first discretized Gaussian we choose mean $\mu = (0.5, 0.5)$ and the identity as covariance matrix Σ_1. For the second and third the mean is again $\mu = (0.5, 0.5)$ but this time with covariance matrices

$$\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.2 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

The first Gaussian is just a shifted bivariate standard normal, i.e., the directions are independent. In the second case this probability measure is highly concentrated in the y-direction, still both directions are independent. In the third case the directions are highly correlated, see Figure 2.3.

(a) Gaussian 1   (b) Gaussian 2   (c) Gaussian 3
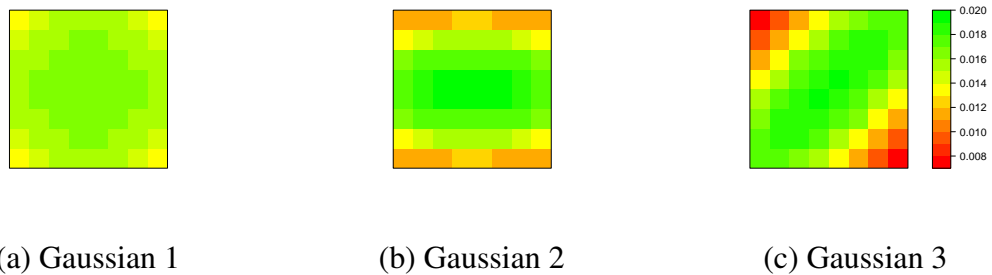
Figure 2.3: **Discretized Gaussians.** The probability weights of three discretized Gaussians with mean $\mu = (0.5, 0.5)$ and covariances Σ_1, Σ_2 and Σ_3 as explained above.

All limiting distributions and their upper bounds based on the different spanning trees are simulated by a sample of 1000 realizations for each grid size $8 \times 8, 16 \times 16, 32 \times 32$, respectively.

In Figure 2.4 we show the resulting densities for all five different considered measures on the 8×8 grid. Surprisingly, we notice that the density of the upper bound based

on the spanning tree 'Fork' is the closest to the true limiting density no matter which measure is underlying. The density of the upper bound with the dyadic partition is more concentrated but farther away from the true distribution than for the chain tree.

As displayed exemplary for the uniform distribution in Figure 2.5 the upper bounds get farther away from density of the true limiting distribution if the gird size increases. This is due to the fact, that for increasing grid size there are more and more nodes that have a farther distance in the tree metric than in the Euclidean distance on the grid. Note that the tree metric is defined via the length of the edges and this is given by the Euclidean distance between the nodes, more precisely the distance in the tree metric is alway greater or equal to the Euclidean distance. We detect the same pattern for all other measures.

(a) Uniform

(b) Random

(c) Gaussian 1

(d) Gaussian 2

(e) Gaussian 3

Figure 2.4: **The densities of the limiting distribution and the upper bounds based on the three different spanning trees for different measures**. The densities were estimated by a kernel density estimator with Gaussian kernel and bandwidth according to Silverman's rule of thumb.
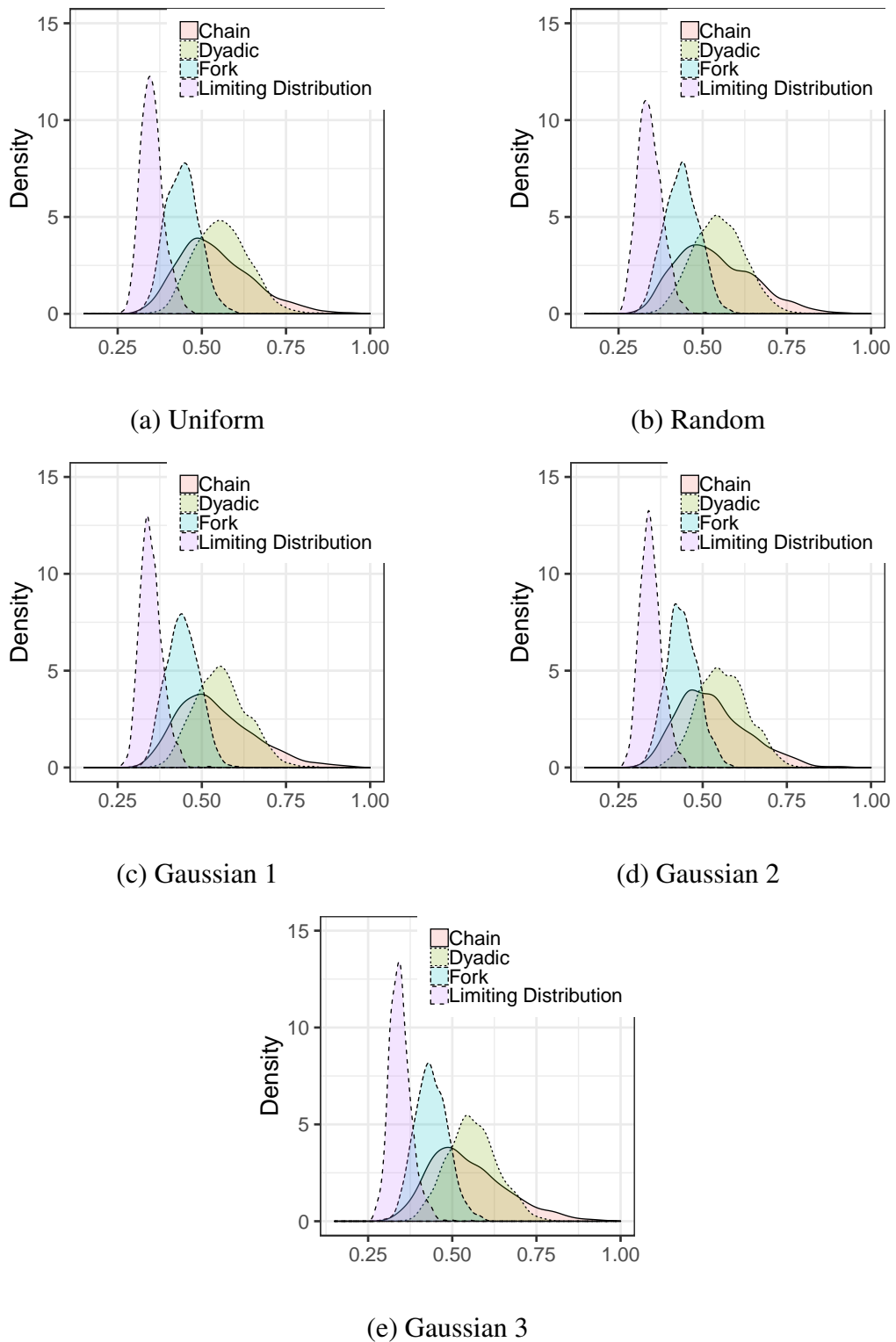
(a) 8×8 grid

(b) 16×16 grid

(c) 32×32 grid

Figure 2.5: **The densities of the limiting distribution and the upper bounds based on the three different spanning trees for the uniform distribution on three different grid sizes.** The densities were estimated by a kernel density estimator with Gaussian kernel and bandwidth according to Silverman's rule of thumb.
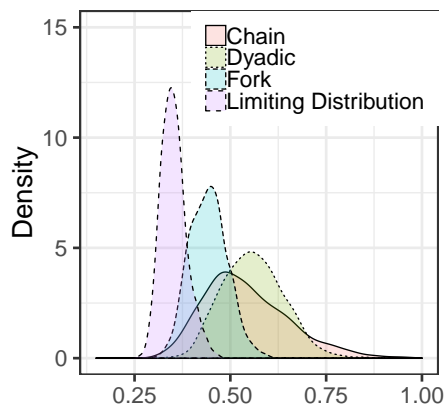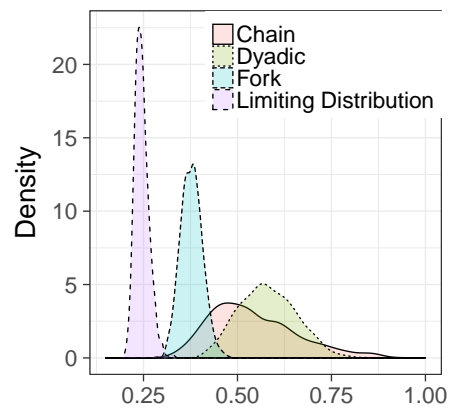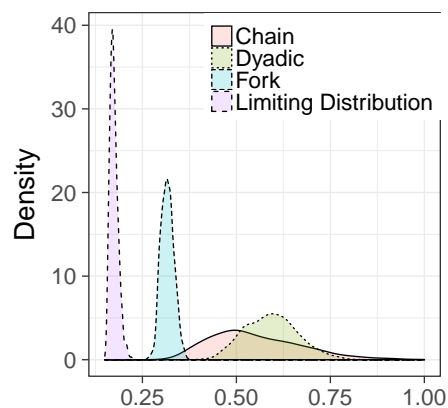
2.3   Proofs

2.3.1   The weighted ℓ^1 space and its dual

In this section, we investigate for $p \geq 1$ and $x_0 \in X$ the space $\ell^1_{d^p_{x_0}}$

$$\ell^1_{d^p_{x_0}} = \left\{ \boldsymbol{r} \in \mathbb{R}^X : \sum_{x \in X} d^p(x_0, x) \, |r_x| + |r_{x_0}| < \infty \right\}.$$

in more detail. Observe, that the space $\ell^1_{d^p_{x_0}}$ is in general not independent of the choice of x_0 as the following example demonstrates.

Example 2.17. *Let* $X = \left\{ \frac{1}{n} \right\}_{n \in \mathbb{N}} \cup \{0\}$ *and* $d(x, y) = |x - y|$. *Then,* $\boldsymbol{r}$ *given by*

$$r_x = \begin{cases} 0, & \text{if } x = 0, \\ \frac{1}{n}, & \text{else,} \end{cases}$$

is an element of $\ell^1_{d^p_0}$ *($p \geq 1$), i.e.,* $x_0 = 0$, *since*

$$\sum_{x \in X} d^p(x_0, x) \, |r_x| + |r_{x_0}| = \sum_{n=1}^{\infty} \frac{1}{n^{p+1}} < \infty.$$

However, for $x_0 = 1$ *it holds*

$$\sum_{x \in X} d^p(x_0, x) \, |r_x| + |r_{x_0}| = 1 + \sum_{n=2}^{\infty} \frac{(n-1)^p}{n^{p+1}} = \infty.$$

Hence, $\boldsymbol{r}$ *is not an element of* $\ell^1_{d^p_1}$.

Nevertheless, we can give conditions on X under which the space $\ell^1_{d^p_{x_0}}$ is independent of the choice of x_0.

Lemma 2.18. *Let* X *have no accumulation point with respect to d and let the unit ball* $B_1(x)$ *around each element* $x \in X$ *be totally bounded. Then, the space* $\ell^1_{d^p_{x_0}}$ *is independent of the choice of* x_0.

Proof. Let $\boldsymbol{r} \in \ell^1_{d^p_{x_0}}$. Further, let $\tilde{x}_0$ be another element in X. Then, it holds

$$\sum_{x \in X} d^p(\tilde{x}_0, x) \, |r_x| + |r_{\tilde{x}_0}| \leq 2^{-p} \sum_{x \in X} d^p(\tilde{x}_0, x_0) \, |r_x| + 2^{-p} \sum_{x \in X} d^p(x_0, x) \, |r_x| + |r_{\tilde{x}_0}|.$$

The second sum is finite by assumption. For the first sum it holds

$$\sum_{x \in X} |r_x| = \sum_{x \in B_1(x_0)} |r_x| + \sum_{x \in B_1^C(x_0)} |r_x| \leq \sum_{x \in B_1(x_0)} |r_x| + \sum_{x \in B_1^C(x_0)} d^p(x_0, x) |r_x|.$$

The sum over $x \in B_1^C(x_0)$ is again finite by assumption. The unit ball $B_1(x_0)$ is totally bounded by assumption and since X contains no accumulation point we can conclude that $|B_1(x_0)| < \infty$. Hence, the sum over $x \in B_1(x_0)$ is finite. From this, it follows that r is also an element of $\ell^1_{d^p_{x_0}}$. □

The rest of this section is devoted to the proof of the fact that $\ell^\infty_{d^{-p}_{x_0}}$ is the dual of $\ell^1_{d^p_{x_0}}$. Note that, for the following considerations it is not important whether the space $\ell^1_{d^p_{x_0}}$ is independent of x_0 or not. The space $\ell^1_{d^p_{x_0}}$ can be viewed as ℓ^1 with weights. For a general sequence of weights $(g_x)_{x \in X}$, i.e., $g_x \geq 0$ for all $x \in X$, we define the subset of X on which the weights are strictly positive $P(g) = \{x \in X \colon g_x > 0\}$. The space ℓ^1_g is then defined as

$$\ell^1_g = \left\{ r \in \mathbb{R}^X \colon \sum_{x \in P(g)} g_x |r_x| + \sum_{x \in P^C(g)} |r_x| < \infty \right\},$$

where $P^C(g)$ denotes the complement of $P(g)$ in X. More precisely, the space ℓ^1_g consists of all elements $r \in \mathbb{R}^X$ with finite

$$\|r\|_{\ell^1_g} = \sum_{x \in P(g)} g_x |r_x| + \sum_{x \in P^C(g)} |r_x|$$

norm. Our aim is now to calculate the dual of ℓ^1_g explicitly. As it will turn out, the dual space is isomorphic to a weighted ℓ^∞ space, i.e.,

$$\ell^\infty_{g^{-1}} = \left\{ \lambda \in \mathbb{R}^X \colon \sup_{x \in X} \left| \mathbb{1}_{\{g_x > 0\}} g_x^{-1} \lambda_x + \mathbb{1}_{\{g_x = 0\}} \lambda_x \right| < \infty \right\}.$$

Observe, that this is a normed space with norm

$$\|\lambda\|_{\ell^\infty_{g^{-1}}} = \sup_{x \in X} \left| \mathbb{1}_{\{g_x > 0\}} g_x^{-1} \lambda_x + \mathbb{1}_{\{g_x = 0\}} \lambda_x \right|.$$

Remark 2.19. *In case that $P^C(g) = \{x_0\}$ is a singleton the $\ell^\infty_{g^{-1}}$-norm can be written as*

$$\|\lambda\|_{\ell^\infty_{g^{-1}}} = \max \left(\sup_{x_0 \neq x \in X} \left| g_x^{-1} \lambda_x \right|, \left| \lambda_{x_0} \right| \right).$$

The first step is to verify that the space ℓ^1 is isomorphic to ℓ^1_g. This holds true as the

function

$$\phi \colon \ell^1 \to \ell_g^1, \quad \phi(\boldsymbol{r})_x = \begin{cases} g_x^{-1} r_x, & \text{if } x \in P(\boldsymbol{g}), \\ r_x, & x \in P^C(\boldsymbol{g}). \end{cases} \tag{2.20}$$

is an isometric isomorphism between ℓ^1 and its weighted version ℓ_g^1. It is well known that a basis of ℓ^1 is given by e^x with

$$(e^x)_y = \begin{cases} 1, & \text{if } y = x, \\ 0, & \text{if } y \neq x. \end{cases}$$

Hence, $\tilde{e}^x = \phi(e^x)$ is a basis of ℓ_g^1 with $\phi(\cdot)$ defined in (2.20). More precisely,

$$(\tilde{e}^x)_y = \begin{cases} \frac{1}{g_y}, & \text{if } y = x \text{ and } y \in P(\boldsymbol{g}), \\ 1, & \text{if } y = x \text{ and } y \in P^C(\boldsymbol{g}), \\ 0, & \text{if } y \neq x. \end{cases}$$

The second step is to show that each $f \in \left(\ell_g^1 \right)^*$ can be written as

$$f(\boldsymbol{r}) = \sum_{x \in \mathcal{X}} \tilde{r}_x \lambda_x \tag{2.21}$$

with $\boldsymbol{\lambda} = (\lambda_x)_{x \in \mathcal{X}} \in \ell_{g^{-1}}^\infty$. Here, $\tilde{r}_x$ are the coefficients of the basis representation, i.e., $\boldsymbol{r} = \sum_{x \in \mathcal{X}} \tilde{r}_x \tilde{e}^x$.

Let $\lambda_x = f(\tilde{e}^x)$ be the value of the continuous linear functional at the basis element $\tilde{e}^x$. For $\boldsymbol{r} \in \ell_g^1$ it holds by the continuity and linearity of f

$$f(\boldsymbol{r}) = \sum_{x \in \mathcal{X}} \tilde{r}_x f(\tilde{e}^x) = \sum_{x \in \mathcal{X}} \tilde{r}_x \lambda_x.$$

Hence, for the representation given in (2.21) it remains to show that $\boldsymbol{\lambda}$ with the choice $\lambda_x = f(\tilde{e}^x)$ is an element of $\ell_{g^{-1}}^\infty$. Therefore, let for $x \in \mathcal{X}$

$$\boldsymbol{r}^x = \left(\mathbb{1}_{\{g_x > 0\}} g_x^{-1} sign(\lambda_x) + \mathbb{1}_{\{g_x = 0\}} sign(\lambda_x) \right) \tilde{e}^x,$$

where $sign(\cdot)$ denotes the sign function. For each $x \in \mathcal{X}$ the sequence $\boldsymbol{r}^x$ is an element of ℓ_g^1 as

$$\|\boldsymbol{r}^x\|_{\ell_g^1} = |sign(\lambda_x)| \leq 1.$$

Further, it holds

$$f(\boldsymbol{r}^x) = \mathbb{1}_{\{g_x > 0\}} g_x^{-1} |\lambda_x| + \mathbb{1}_{\{g_x = 0\}} |\lambda_x|$$

and hence, we can deduce that

$$\sup_{x \in \mathcal{X}} \left\{ \mathbb{1}_{\{g_x > 0\}} g_x^{-1} |\lambda_x| + \mathbb{1}_{\{g_x = 0\}} |\lambda_x| \right\} \le \sup_{x \in \mathcal{X}} f(\boldsymbol{r}^x) \le \|f\|_{(\ell_g^1)^*} \le \infty.$$

This proves that $\boldsymbol{\lambda} = (\lambda_x)_{x \in \mathcal{X}}$ with $\lambda_x = f(\tilde{\boldsymbol{e}}^x)$ is an element of $\ell_{g^{-1}}^\infty$.

As the last step we show that for all $\boldsymbol{\lambda} \in \ell_{g^{-1}}^\infty$ the functional

$$L_{\boldsymbol{\lambda}}(\boldsymbol{r}) = \sum_{x \in \mathcal{X}} \lambda_x \tilde{r}_x$$

is an element of $\left(\ell_g^1 \right)^*$ and that it holds $\|L_{\boldsymbol{\lambda}}\|_{(\ell_g^1)^*} = \|\boldsymbol{\lambda}\|_{\ell_{g^{-1}}^\infty}$. Clearly, $L_{\boldsymbol{\lambda}}(\cdot)$ is a linear functional and it is bounded as

$$|L_{\boldsymbol{\lambda}}(\boldsymbol{r})| \le \sum_{x \in \mathcal{X}} |\lambda_x \tilde{r}_x| \le \|\boldsymbol{\lambda}\|_{\ell_{g^{-1}}^\infty} \|\boldsymbol{r}\|_{\ell_g^1}$$

and hence, it holds by the definition of the dual norm

$$\|L_{\boldsymbol{\lambda}}\|_{(\ell_g^1)^*} \le \|\boldsymbol{\lambda}\|_{\ell_{g^{-1}}^\infty}.$$

The equality $\|L_{\boldsymbol{\lambda}}\|_{(\ell_g^1)^*} = \|\boldsymbol{\lambda}\|_{\ell_{g^{-1}}^\infty}$ can be deduced from the fact that we have proven the reversed inequality in the second step. Altogether, we can conclude that the dual space of ℓ_g^1 is isomorphic to $\ell_{g^{-1}}^\infty$.

2.3.2 Hadamard directional differentiability

In this section we follow mainly Shapiro (1991) and Römisch (2004). Let $\mathcal{U}$ and $\mathcal{Y}$ be normed spaces.

Definition 2.20 (cf. Shapiro (1991); Römisch (2004))**.**

a) *Hadamard directional differentiability*

 A mapping $f \colon D_f \subset \mathcal{U} \to \mathcal{Y}$ is said to be Hadamard directionally differentiable at $u \in D_f$ if for any sequence h_n that converges to h and any sequence $t_n \searrow 0$ such that $u + t_n h_n \in D_f$ for all $n \in \mathbb{N}$ the limit

$$f_u'(h) = \lim_{n \to \infty} \frac{f(u + t_n h_n) - f(u)}{t_n} \tag{2.22}$$

 exists.

b) *Hadamard directional differentiability tangentially to a set*

Let K be a subset of $\mathcal{U}$, f is directionally differentiable tangentially to K in the sense of Hadamard at u if the limit (2.22) exists for all sequences h_n of the form $h_n = t_n^{-1}(k_n - u)$ where $k_n \in K$ and $t_n \searrow 0$ that converge to h. This derivative is defined on the contingent (Bouligand) cone to K at u

$$T_K(u) = \left\{ h \in \mathcal{U} : h = \lim_{n \to \infty} t_n^{-1}(k_n - u), k_n \in K, t_n \searrow 0 \right\}.$$

Note that this derivative is not required to be linear in h, but it is still positively homogeneous. Moreover, the directional Hadamard derivative $f'_u(\cdot)$ is continuous if u is an interior point of D_f (Römisch, 2004).

Hadamard directional differentiability of the Wasserstein distance on countable metric spaces   For $r, s \in \mathcal{P}_p(X)$ the p-th power of the p-th Wasserstein distance is the optimal value of an infinite dimensional linear program. We use this fact to verify that the p-th power of the Wasserstein distance (1.9) on the countable metric spaces X is directionally Hadamard differentiable with methods of sensitivity analysis for optimal values in linear programming.

The p-th power of the Wasserstein distance on countable metric spaces is the optimal value of the following infinite dimensional linear program

$$
\begin{aligned}
\min_{w \in \ell^1_{d^p_{x_0}}(X \times X)} &\sum_{x,x' \in X} d^p(x, x') w_{x,x'} \\
\text{subject to } &\sum_{x' \in X} w_{x,x'} = r_x \quad \forall x \in X, \\
&\sum_{x \in X} w_{x,x'} = s_{x'}, \quad \forall x' \in X, \\
&w_{x,x'} \geq 0, \quad \forall x, x' \in X.
\end{aligned}
\tag{2.23}
$$

Theorem 2.21. *W_p^p as a map from $(\mathcal{P}_p(X) \times \mathcal{P}_p(X), \| \cdot \|_{\ell^1(d^p_{x_0})})$ to $\mathbb{R}$, $(r, s) \mapsto W_p^p(r, s)$ is Hadamard directionally differentiable tangentially to $\mathcal{P}_p(X) \times \mathcal{P}_p(X)$. The contingent cone, on which the derivative is defined, is given by*

$$\mathcal{D}(r, s) = \mathcal{D}(r) \times \mathcal{D}(s)$$

with

$$\mathcal{D}(\boldsymbol{r}) := \left\{ \boldsymbol{d} \in \ell^1_{d^p_{x_0}}(\mathcal{X}) \backslash \{0\} : \sum_{x \in \mathcal{X}} d_x = 0, \ d_x \in [-r_x, 1 - r_x] \right\}$$

and the directional derivative is as follows

$$(\boldsymbol{d}_1, \boldsymbol{d}_2) \mapsto \sup_{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s})} -(\langle \boldsymbol{\lambda}, \boldsymbol{d}_1 \rangle + \langle \boldsymbol{\mu}, \boldsymbol{d}_2 \rangle), \tag{2.24}$$

where $\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s})$ is set of optimal solutions of the dual problem which is defined in (2.3).

Proof. We start the proof with stating the considered functions and the spaces on which they are defined. The *objective function* of the linear program that determines the p-th power of the p-th Wasserstein distance is given as $f \colon \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X}) \to \mathbb{R}, \boldsymbol{w} \mapsto \sum_{x,x' \in \mathcal{X}} d^p(x, x') w_{x,x'}$. The constraints are encoded by the *constraint function* $C \colon \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X}) \to \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X})$ with

$$C(\boldsymbol{w}, (\boldsymbol{r}, \boldsymbol{s})) = \begin{pmatrix} \boldsymbol{w} \\ \Sigma_1 \boldsymbol{w} - \boldsymbol{r} \\ \Sigma_2 \boldsymbol{w} - \boldsymbol{s} \end{pmatrix}, \tag{2.25}$$

here $\Sigma_1, \Sigma_2 \colon \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X}) \to \ell^1_{d^p_{x_0}}(\mathcal{X})$ are the summation operators over the first and the second component, i.e., $(\Sigma_1 \boldsymbol{w})_x = \sum_{x' \in \mathcal{X}} w_{x,x'}$ and $(\Sigma_2 \boldsymbol{w})_{x'} = \sum_{x \in \mathcal{X}} w_{x,x'}$. Furthermore, we need the closed convex set $K = \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X})_+ \times \{0\} \times \{0\}$, where $\ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X})_+$ are the elements in $\ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X})$ that have only non-negative entries. With these definitions the p-th power of the p-th Wasserstein distance is the optimal value of the abstract parametrized optimization problem

$$\min_{\boldsymbol{w} \in \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X})} f(\boldsymbol{w}) \text{ s.t. } C(\boldsymbol{w}, (\boldsymbol{r}, \boldsymbol{s})) \in K. \tag{2.26}$$

We will use Theorem 4.24 from Bonnans and Shapiro (2000). To this end, we need to check the following three conditions.

(i.) *Convexity and existence of optimal solution*

Problem (2.23) is convex, since the objective function f is convex and the constraint set $K = \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X})_+ \times \{0\} \times \{0\}$ is convex. It remains to show that the constraint function C in (2.25) is convex with respect to $-K$, i.e.,

$$\psi((\boldsymbol{w}, (\boldsymbol{r}, \boldsymbol{s})), (\tilde{\boldsymbol{w}}, (\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{s}}))) = I_K(C((\boldsymbol{w}, (\boldsymbol{r}, \boldsymbol{s}))) + (\tilde{\boldsymbol{w}}, (\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{s}})))$$

is convex for $(\boldsymbol{w}, (\boldsymbol{r}, \boldsymbol{s})), (\tilde{\boldsymbol{w}}, (\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{s}})) \in \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X})$, where I_K is the indicator function on K.

Let $(\boldsymbol{w}_1, \boldsymbol{r}_1, \boldsymbol{s}_1, \tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{r}}_1, \tilde{\boldsymbol{s}}_1)$ and $(\boldsymbol{w}_2, \boldsymbol{r}_2, \boldsymbol{s}_2, \tilde{\boldsymbol{w}}_2, \tilde{\boldsymbol{r}}_2, \tilde{\boldsymbol{s}}_2)$ be in $\ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X}) \times$
$\ell^1_{d^p_{x_0}}(\mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X})$ such that $C(\boldsymbol{w}_1, (\boldsymbol{r}_1, \boldsymbol{s}_1)) + (\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{r}}_1, \tilde{\boldsymbol{s}}_1)$
and $C(\boldsymbol{w}_2, (\boldsymbol{r}_2, \boldsymbol{s}_2)) + (\tilde{\boldsymbol{w}}_2, \tilde{\boldsymbol{r}}_2, \tilde{\boldsymbol{s}}_2)$ are in K. This yields for $i = 1, 2$ that

$$\boldsymbol{w}_i + \tilde{\boldsymbol{w}}_i \geq \boldsymbol{0},$$
$$\Sigma_1 \boldsymbol{w}_i - \boldsymbol{r}_i + \tilde{\boldsymbol{r}}_i = \boldsymbol{0}, \tag{2.27}$$
$$\Sigma_2 \boldsymbol{w}_i - \boldsymbol{s}_i + \tilde{\boldsymbol{s}}_i = \boldsymbol{0}.$$

Therefore, the convex combination

$$\lambda \cdot (C(\boldsymbol{w}_1, (\boldsymbol{r}_1, \boldsymbol{s}_1)) + (\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{r}}_1, \tilde{\boldsymbol{s}}_1)) + (1 - \lambda) \cdot (C(\boldsymbol{w}_2, (\boldsymbol{r}_2, \boldsymbol{s}_2)) + (\tilde{\boldsymbol{w}}_2, \tilde{\boldsymbol{r}}_2, \tilde{\boldsymbol{s}}_2))$$

is for all $\lambda \in [0, 1]$ an element of K.

Next, we want to show that the set of primal optimal solutions $\mathcal{S}(\boldsymbol{r}, \boldsymbol{s})$ is non-empty. Since $\mathcal{X}$ is countable, the space is separable. If we take the discrete topology on $\mathcal{X}$ that is induced by the discrete metric

$$d^D(x, y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y, \end{cases}$$

our space is complete and hence, $\mathcal{X}$ is a Polish space. By Theorem 4.1 in Villani (2008) the set of optimal solutions for (2.23) is non-empty for each $(\boldsymbol{r}, \boldsymbol{s}) \in \mathcal{P}_p(\mathcal{X}) \times \mathcal{P}_p(\mathcal{X})$ in the right hand side of the constraints of (2.23).

(ii.) *Directional regularity*

Set for some direction $(\boldsymbol{d}_1, \boldsymbol{d}_2) \in \mathcal{D}(\boldsymbol{r}, \boldsymbol{s}) \subset \ell^1_{d^p_{x_0}}(\mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X})$

$$\bar{C}(\boldsymbol{w}, t) = (\boldsymbol{w}, \boldsymbol{w}^T \mathbb{1} - \boldsymbol{r} - t\boldsymbol{d}_1, \boldsymbol{w}\mathbb{1} - \boldsymbol{s} - t\boldsymbol{d}_2, t).$$

The directional regularity condition is fulfilled at $\boldsymbol{w}_0$ in a direction $(\boldsymbol{d}_1, \boldsymbol{d}_2)$ if Robinson's constraint qualification is satisfied at the point $(\boldsymbol{w}_0, 0)$ for the mapping $\bar{C}(\boldsymbol{w}, t)$ with respect to the set $K \times \mathbb{R}_+$ (Bonnans and Shapiro, 2000, Def. 4.8). According to Theorem 4.9 in Bonnans and Shapiro (2000) the condition

$$0 \in int\left\{C(\boldsymbol{w}_0, (\boldsymbol{r}, \boldsymbol{s})) + DC(\boldsymbol{w}, (\boldsymbol{r}, \boldsymbol{s}))(\ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X}), \mathbb{R}_+(\boldsymbol{d}_1, \boldsymbol{d}_2)) - K\right\},$$

where $\mathbb{R}_+(d_1, d_2) = \{t(d_1, d_2), t \geq 0\}$, is necessary and sufficient for the directional regularity constraint. We are going to show that the directional regularity condition in a direction $(d_1, d_2) \in \mathcal{D}(r, s)$ holds for all primal optimal solutions $w_0 \in \mathcal{S}(r, s)$. For a primal optimal solution w_0 it is

$$C(w_0, (r, s)) = (w_0, 0, 0).$$

In the following, we prove that $C(w, (r, s))$ is bounded with respect to the product norm on the space $\ell^1_{d^p_{x_0}}(X \times X) \times \ell^1_{d^p_{x_0}}(X) \times \ell^1_{d^p_{x_0}}(X)$ as this together with the linearity of C in $(w, (r, s))$ yields that the derivative is the mapping itself. Let $\|\cdot\|$ denote the product norm on $\ell^1_{d^p_{x_0}}(X \times X) \times \ell^1_{d^p_{x_0}}(X) \times \ell^1_{d^p_{x_0}}(X)$. Then, we obtain

$$
\begin{aligned}
\|C(w, (r, s))\| &= \|w\|_{\ell^1(d^p_{x_0})} + \|\Sigma_1 w - r\|_{\ell^1(d^p_{x_0})} + \|\Sigma_2 w - s\|_{\ell^1(d^p_{x_0})} \\
&\leq \|w\|_{\ell^1(d^p_{x_0})} + \|\Sigma_1 w\|_{\ell^1(d^p_{x_0})} + \|r\|_{\ell^1(d^p_{x_0})} + \|\Sigma_2 w\|_{\ell^1(d^p_{x_0})} + \|s\|_{\ell^1(d^p_{x_0})} \\
&\leq 2\|w\|_{\ell^1(d^p_{x_0})} + \|r\|_{\ell^1(d^p_{x_0})} + \|s\|_{\ell^1(d^p_{x_0})} \\
&\leq 2\|(w, (r, s))\|.
\end{aligned}
$$

Hence, it holds that

$$DC(w_0, (r, s))(\ell^1_{d^p_{x_0}}(X \times X), \mathbb{R}_+(d_1, d_2)) = (w, \Sigma_1 w - td_1, \Sigma_2 w - td_2)$$

for $t \geq 0$ and the directional regularity condition reads

$$0 \in int\{(w_0, 0, 0) + (w, \Sigma_1 w - td_1, \Sigma_2 w - td_2) - K\}.$$

This set is just $\ell^1_{d^p_{x_0}}(X \times X) \times \ell^1_{d^p_{x_0}}(X) \times \ell^1_{d^p_{x_0}}(X)$ as $w \in \ell^1_{d^p_{x_0}}(X \times X)$ and hence the directional regularity constraint is fulfilled.

(iii.) *Stability of primal optimal solution*

We aim to verify that for perturbed measures of the form $r_n = r + t_n d_1 + o(t_n)$ and $s_n = s + t_n d_2 + o(t_n)$ with $t_n \searrow 0$, $r, s \in \mathcal{P}_p(X)$, $d_1 \in \mathcal{D}(r)$ and $d_2 \in \mathcal{D}(s)$ there exists a sequence of primal optimal solutions w_n that converges to the primal optimal solution w_0 of the unperturbed problem. For n large enough it holds $t_n \leq 1$. Hence, we can assume without loss of generality that $t_n \leq 1$ for all n. In this case r_n and s_n are probability measure with existing p-th moment, i.e., elements of $\mathcal{P}_p(X)$. Now, Theorem 5.20 in Villani (2008) yields the stability of the optimal solution as $\mathcal{P}_p(X)$ is a closed subset of $\ell^1_{d^p_{x_0}}(X)$.

So far, we checked all the assumptions of Theorem 4.24 in Bonnans and Shapiro (2000).

The rest of this section is devoted to the derivation of formula (2.24). The Lagrangian L of a parametrized optimization problem

$$\min_w f(w, u) \text{ s.t. } C(w, u) \in K$$

is given by

$$L(w, \lambda, u) = f(w, u) + \langle \lambda, C(w, u) \rangle,$$

where f is the objective function, u the parameter, C the constraint function and $\langle \cdot, \cdot \rangle$ the dual pairing (see for example Section 2.5.2 in Bonnans and Shapiro (2000)). We refer to λ as Lagrange multiplier. For the transport problem this yields with (r, s) being the parameter and the definition of the constraint function in (2.25)

$$L(w, (\nu, \lambda, \mu), (r, s))$$
$$= \sum_{x,x' \in X} d^p(x, x') w_{x,x'} + \langle \nu, w \rangle + \langle \lambda, \Sigma_1 w - r \rangle + \langle \mu, \Sigma_2 w - s \rangle.$$

Differentiating this in the Fréchet sense with respect to (r, s) and applying (d_1, d_2) to this linear operator results in

$$D_{(r,s)} L(w, (\nu, \lambda, \mu), (r, s))(d_1, d_2) = -(\langle \lambda, d_1 \rangle + \langle \mu, d_2 \rangle)$$

as the Lagrangian is linear and bounded in (r, s). As this derivative is independent of w and the set of Lagrange multipliers $\Lambda(r, s)$ equals the set of dual solutions $\mathcal{S}^*(r, s)$ in the case of a convex unperturbed problem (see section above Theorem 4.24 in Bonnans and Shapiro (2000)) it holds that the directional Hadamard derivative is given by

$$(d_1, d_2) \mapsto \inf_{w \in \mathcal{S}(r,s)} \sup_{(\lambda,\mu) \in \Lambda(r,s)} D_{(r,s)} L(w, (\nu, \lambda, \mu), (r, s))(d_1, d_2)$$
$$= \inf_{w \in \mathcal{S}(r,s)} \sup_{(\lambda,\mu) \in \Lambda(r,s)} -(\langle \lambda, d_1 \rangle + \langle \mu, d_2 \rangle)$$
$$= \sup_{(\lambda,\mu) \in \mathcal{S}^*(r,s)} -(\langle \lambda, d_1 \rangle + \langle \mu, d_2 \rangle).$$

$\square$

2.3.3 The limit distribution under equality of measures

First, observe that for the case $r = s$ the set of dual solutions $\mathcal{S}^*(r, r)$ in (2.3) is given by

$$
\begin{aligned}
\mathcal{S}^*(r, r) &= \Big\{ (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) \times \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) : \langle r, \boldsymbol{\lambda} \rangle + \langle r, \boldsymbol{\mu} \rangle = 0, \\
&\qquad\qquad\qquad\qquad\qquad\qquad \lambda_x + \mu_{x'} \le d^p(x, x') \quad \forall x, x' \in \mathcal{X} \Big\} \\
&= \Big\{ (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) \times \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) : \sum_{x \in \mathcal{X},\, \text{s.t.} r_x > 0} r_x(\lambda_x + \mu_x) = 0, \\
&\qquad\qquad\qquad\qquad\qquad\qquad \lambda_x + \mu_{x'} \le d^p(x, x') \quad \forall x, x' \in \mathcal{X} \Big\} \\
&= \Big\{ (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) \times \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) : \lambda_x = -\mu_x \text{ for } x \in \operatorname{supp}(r), \\
&\qquad\qquad\qquad\qquad\qquad\qquad \lambda_x + \mu_{x'} \le d^p(x, x') \quad \forall x, x' \in \mathcal{X} \Big\}.
\end{aligned}
$$

The equality follows as for $x = x'$ the inequality condition gives $\lambda_x + \mu_x \le 0$ and all r_x in the sum are non-negative. The conjunction of these two conditions yields $\lambda_x + \mu_x = 0$. This set is a subset of the set given in (2.4), but changing $\mathcal{S}^*(r, r)$ to $\mathcal{S}^*(r)$ does not change the optimal value of the linear programs in Theorem 2.1 and 2.4 as the Gaussian process $\mathbf{G}$ is zero at all $x \notin \operatorname{supp}(r)$.

In the case, that the support of r, i.e., $\{x \in \mathcal{X}: r_x > 0\}$, is the whole ground space $\mathcal{X}$, the set $\mathcal{S}^*(r)$ is independent of r and hence given by

$$
\mathcal{S}^* = \Big\{ \boldsymbol{\lambda} \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) : \lambda_x - \lambda_{x'} \le d^p(x, x') \quad \forall x, x' \in \mathcal{X} \Big\}.
$$

2.3.4 Proof of Theorem 2.14

Simplify the set of dual solutions $\mathcal{S}^*$ As a first step, we rewrite the set of dual solutions $\mathcal{S}^*$ given in definition (2.8) in our tree notation as

$$
\mathcal{S}^* = \Big\{ \boldsymbol{\lambda} \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X}) : \lambda_x - \lambda_{x'} \le d_{\mathcal{T}}(x, x')^p, \quad x, x' \in \mathcal{X} \Big\}. \tag{2.28}
$$

The key observation is that in the condition $\lambda_x - \lambda_{x'} \le d_{\mathcal{T}}(x, x')^p$ 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 $x = x_1, \ldots, x_n = x'$ the sequence of vertices defining the unique path joining x and x', such that $(x_j, x_{j+1}) \in E$ for $j = 1, \ldots, n - 1$. That this path contains only a finite number

of edges, was proven in Chapter 2.2.1. Then,

$$\lambda_x - \lambda_{x'} = \sum_{j=1}^{n-1}(\lambda_{x_j} - \lambda_{x_{j+1}}) \le \sum_{j=1}^{n-1} d_{\mathcal{T}}(x_j, x_{j+1})^p \le \left(\sum_{j=1}^{n-1} d_{\mathcal{T}}(x_j, x_{j+1})\right)^p \le d_{\mathcal{T}}(x, x')^p,$$

such that (2.28) is satisfied for all $x, x' \in X$. Noting that if two vertices are joined by an edge then one has to be the parent of the other, we can write the set of dual solutions as

$$S^* = \left\{\lambda \in \ell^\infty_{d_{x_0}^{-p}}(X) : |\lambda_x - \lambda_{\text{parent}(x)}| \le d_{\mathcal{T}}(x, \text{parent}(x))^p, \quad x \in X\right\}. \tag{2.29}$$

Rewrite the target function   To rewrite the target function we need to make several definitions. Let

$$\tilde{e}_y^{(x)} = \begin{cases} \frac{1}{d^p(x, x_0)} & \text{if } y = x, \\ -\frac{1}{d^p(x, x_0)} & \text{if } y = \text{parent}(x), \\ 0 & \text{else.} \end{cases}$$

Furthermore, we define for $r \in \ell^1_{d_{x_0}^p}(X)$

$$s_x = \sum_{x' \in \text{children}(x)} d^p(x, x_0) r_{x'}$$

and

$$r_n = \sum_{x \in A_{\le n}\backslash \text{root}(\mathcal{T})} s_x \tilde{e}^{(x)} = r\mathbb{1}_{A_{<n}} + \sum_{x \in A_{=n}} \frac{1}{d^p(x, x_0)} s_x e(x),$$

here, $A_{\le n} = \{x \in X : \text{level of } x \le n, x \text{ is within the first } n \text{ vertices of its level}\}$, $A_{=n} = \{x \in X : \text{level of } x = n, x \text{ is within the first } n \text{ vertices of its level}\}$, $A_{>n} = \{x \in X : \text{level of } x > n \text{ or } x \text{ is not within the first } n \text{ vertices of its level}\}$ and $e(x)$ the sequence which is 1 at x and 0 everywhere else. For this sequence r_n it holds

$$\|r - r_n\|_{\ell^1(d_{x_0}^p)} = \sum_{x \in X} d^p(x, x_0) \left|r\mathbb{1}_{A_{>n}} - \sum_{\tilde{x} \in A_{=n}} \frac{1}{d^p(\tilde{x}, x_0)} s_{\tilde{x}} e^{(\tilde{x})}\right|_x$$

$$\le \|r\mathbb{1}_{A_{>n}}\|_{\ell^1(d_{x_0}^p)} + \left|\sum_{x \in A_{=n}} s_x\right|.$$

As $n \to \infty$, the first part tends to zero as $r \in \ell^1_{d_{x_0}^p}(X)$, and

$$\left|\sum_{x \in A_{=n}} s_x\right| \le \sum_{x \in A_{=n}} \sum_{x' \in \text{children}(x)} |r_{x'}| d^p(x', x_0) \le \sum_{x \in A_{\ge n}} |r_x| d^p(x, x_0) \xrightarrow{n \to \infty} 0.$$

Hence, our target function for $r \in \ell^1_{d^p_{x_0}}(\mathcal{X})$ and $\lambda \in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X})$ can be rewritten in the following way

$$
\begin{aligned}
\langle r, \lambda \rangle &= \lim_{n \to \infty} \langle r_n, \lambda \rangle \\
&= \lim_{n \to \infty} \sum_{x \in A_{\leq n}} s_x \langle \tilde{e}^{(x)}, \lambda \rangle \\
&= \lim_{n \to \infty} \sum_{x \in A_{\leq n}} \sum_{x' \in \text{children}(x)} r_{x'}(\lambda_x - \lambda_{\text{parent}(x)}) \\
&\leq \lim_{n \to \infty} \sum_{x \in A_{\leq n}} \left| \sum_{x' \in \text{children}(x)} r_{x'} \right| \left| \lambda_x - \lambda_{\text{parent}(x)} \right| \\
&= \lim_{n \to \infty} \sum_{x \in A_{\leq n}} |(S_{\mathcal{T}} r)_x| \left| \lambda_x - \lambda_{\text{parent}(x)} \right|
\end{aligned}
\tag{2.30}
$$

Observe that for $\lambda \in \mathcal{S}^*$ it holds

$$
\left| \lambda_x - \lambda_{\text{parent}(x)} \right| \leq d^P(x, \text{parent}(x)).
\tag{2.31}
$$

By condition (1.11) $G \sim \mathcal{N}(0, \Sigma(r))$ is an element of $\ell^1_{d^p_{x_0}}(\mathcal{X})$. For $\lambda \in \mathcal{S}^*$ we get with (2.30) and (2.31) that

$$
\langle G, \lambda \rangle \leq \lim_{n \to \infty} \sum_{x \in A_{\leq n}} |(S_{\mathcal{T}} G)_x| \, d_{\mathcal{T}}(x, \text{parent}(x))^p.
\tag{2.32}
$$

Therefore, $\max_{\lambda \in \mathcal{S}^*} \langle G, \lambda \rangle$ is bounded by $\lim_{n \to \infty} \sum_{x \in A_{\leq n}} |(S_{\mathcal{T}} G)_x| d_{\mathcal{T}}(x, \text{parent}(x))^p$. We can define the sequence $\nu \in \ell^\infty_{d^{-p}}(\mathcal{X})$ by

$$
\begin{aligned}
\nu_{\text{root}} &= 0 \\
\nu_x - \nu_{\text{parent}(x)} &= \text{sign}((S_{\mathcal{T}} G)_x) d_{\mathcal{T}}(x, \text{parent}(x))^p
\end{aligned}
\tag{2.33}
$$

From (2.29) and the fact that $d^p(x, \text{parent}(x)) \leq d^p(x, \text{root}(\mathcal{T}))$ we see that $\nu \in \mathcal{S}^*$ and by plugging ν into equation (2.32) we can conclude that $\langle G, \nu \rangle$ attains the upper bound in (2.32).

As the last step of our proof, we verify that the limit in (2.32) exists. Therefore, we rewrite condition (1.11) in terms of the edges and recall that $x_0 = \text{root}(\mathcal{T})$

$$
\sum_{x \in \mathcal{X}} d_{\mathcal{T}}(x, x_0)^p \sqrt{r_x} \geq \sum_{x \in \mathcal{X}} \sum_{x' \in \text{children}(x)} d_{\mathcal{T}}(x, \text{parent}(x))^p \sqrt{r_{x'}}.
\tag{2.34}
$$

The first moment of the limiting distribution can be bounded in the following way:

$$
\mathbb{E}\left[\sum_{x \in \mathcal{X} \setminus \{\text{root}(\mathcal{T})\}} |(S_{\mathcal{T}} \boldsymbol{G})_x| d_{\mathcal{T}}(x, \text{parent}(x))^p\right]
$$

$$
\leq \sum_{x \in \mathcal{X}} d_{\mathcal{T}}(x, \text{parent}(x))^p \sqrt{(S_{\mathcal{T}} r)_x (1 - (S_{\mathcal{T}} r)_x)}
$$

$$
\leq \sum_{x \in \mathcal{X}} \sum_{x' \in \text{children}(x)} d_{\mathcal{T}}(x, \text{parent}(x))^p \sqrt{r_{x'}}
$$

$$
< \infty
$$

due to Hölder's inequality and (2.34). This bound shows that the limit in (2.32) is almost surely finite and hence, concludes the proof.

CHAPTER 3

Colocalization

In this chapter we propose a new method to analyze spatial proximity - colocalization - in coordinate-targeted super-resolution images.

The methods based on pixel intensity correlation are well-suited for the analysis of diffraction limited data. To be more precise, two images $Img1$ and $Img2$ each with $N = N_x \times N_y$ pixels where N_x, N_y are the number of pixels in x- and y- direction, respectively, are considered as a data set $(Img1_1, Img2_1), \ldots, (Img1_N, Img2_N)$ for $i = 1, \ldots, N$. The pixel intensity correlation methods are based on the correlation between these data sets. These methods are very sensitive to the resolution of the images to be compared. With increasing resolution the correlative nature, i.e., the actual signal overlap, of colocalization decreases as it is more likely that two neighboring proteins are imaged in two different pixels (see Figure 3.1). On the scale of a true single molecule
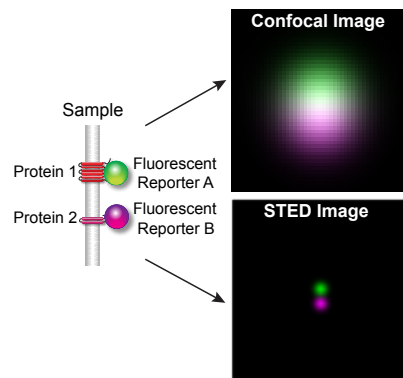

Figure 3.1: **Simulation of Confocal and STED images** of two proteins which are located at a distance of 45 nm. The resolution of the confocal image is 244 nm and for the STED image it is 40 nm.

resolution all these methods would be zero or even negative and hence would indicate no colocalization at all.

A further challenge to be faced in the analysis of super-resolution data is that the data sets that are generated by the two types of super-resolution microscopy methods are fundamentally different. Coordinate-stochastic nanoscopy provides lists containing molecule coordinates where coordinate-targeted nanoscopy yields pixel images containing fluorescence intensities. For the lists with molecule coordinates obtained by coordinate-stochastic nanoscopy methods from spatial statistics (object-based methods) are widely used and show a good performance.

These object-based methods are not directly applicable to STED data sets as they are pixel images. One could estimate the coordinates from the images using mathematical approaches (see e.g. Blom et al. (2012)), but this goes along with the introduction of a statistical error and the loss of the pixel intensity information. Consequently, there is the need for a direct pixel-based method that is able to quantify colocalization in coordinate-targeted nanoscopy.

3.1 Optimal transport colocalization

To reformulate the colocalization problem in terms of optimal transport, we consider the set of $N_x \times N_y = N$ pixels of the pixel image as our ground space $\mathcal{X} = \{x_1, \ldots, x_N\}$. Here, N_x and N_y denote the number of pixels in the x- and the y-direction, respectively. More precisely, $x_{(j-1) \cdot N_y + i} = P_{ij}$, where P_{ij} for $i = 1, \ldots, N_y$ and $j = 1, \ldots, N_x$ denotes pixel i, j. Furthermore, the intensities generated by STED nanoscopy themselves are viewed as probability measure supported on a subset of $\mathbb{R}^2$ by rescaling the intensities such that they sum up to one, i.e., they are elements of $\mathcal{P}(\mathcal{X})$. More precisely, for a pixel size of l nanometers we consider the image as an probability measure on an equidistant grid in $[0, N_x \cdot l] \times [0, N_y \cdot l]$ which represents the pixels.

Let w^* be the optimal solution of the linear program defining the Wasserstein distance in (1.9) with $p = 2$ and the Euclidean distance as the metric on the set of pixels $\mathcal{X}$. We introduce the *Optimal Transport Colocalization (OTC)* at spatial size t between two probability measures r and s in $\mathcal{P}(\mathcal{X})$ as

$$OTC(t) := \sum_{i,j=1}^{N} \mathbb{1}\{\|x_i - x_j\| \le t\} w^*_{ij}, \tag{3.1}$$

where $\|\cdot\|$ denotes the Euclidean distance and $\mathbb{1}$ the indicator function, i.e.,

$$\mathbb{1}\{\|x_i - x_j\| \le t\} = \begin{cases} 1, & \text{if } \|x_i - x_j\| \le t \\ 0, & \text{else.} \end{cases}$$

Intuitively, $OTC(t)$ describes the amount of mass (intensity of an image) matched at distances not larger than t inherent in the optimal matching (of the intensity distributions) between r and s. See Figure 3.2(c) for a schematic representation.

By definition OTC is a value in [0, 1] for any t. The OTC curves are monotone increasing and they approach one. Due to the fact that OTC analysis provides curves and not a fixed value, it is a relative rather than an absolute measure for colocalization. Hence, we propose that it is most reasonable to compare two OTC curves where one can serve as a reference curve instead of interpreting a single OTC curve.

The biggest advantage of OTC is that it is able to detect colocalization on different scales simultaneously. Figure 3.2(a) illustrates the case of colocalized structures which require only a relatively small spatial adjustment to be matched. If we analyze the colocalization of Image 1 and Image 2 in Figure 3.2(a) we observe that these structures are perfectly colocalized at a distance of one pixel in the diagonal direction. This is displayed by the optimal transport plan which is indicated as light blue arrows in the right column of Figure 3.2(a). The structures in Image 1 and Image 3 are colocalized on different scales, i.e., the vertical line is shifted by one pixel and the horizontal part by two pixels, see the respective optimal transport plan shown as light blue arrows in the right column of Figure 3.2(a). The size of the pixels in the three images is set to 15 nm as in the following real data sets. Hence, the whole image is contained in $[0, 150]^2$. The OTC captures these different scales, see Figure 3.2(b). The red curve indicates that the structures in Image 1 and Image 2 are perfectly colocalized at a scale of 25 nm. The green dashed curve shows that roughly 37% of the objects in Image 1 and 3 (i.e., the vertical part of the structure) are colocalized at a scale of 15 nm and perfect colocalization appears on a scale of 30 nm. Manders' M1 and M2 coefficients equal zero in both settings and Pearson's correlation between Image 1 and Image 2 is -0.12 as well as between Image1 and Image 3. Hence, the conventional coefficients detect no colocalization at all for these structures.

3.1.1 Computational aspects of OTC

For all computational tasks we use **R** (R Core Team, 2018). The optimal transport plan is calculated with the **R**-package *transport* (Schuhmacher et al., 2017). More precisely, we use the shielding algorithm (Schmitzer, 2016) to solve the optimal transport problem. To solve the optimal transport problem is a computational bottleneck as standard solvers have a runtime of $O(N^3 \log(N))$ (Pele and Werman, 2009). This renders many practical real world problems computational infeasible (Schrieber et al., 2017) including to calculate OTC for the whole STED images which are usually of size 1000×1000 or even larger. Therefore, we propose a uniform random sampling scheme to select image

sections of size 128×128 pixels from bigger images to gain computational speed. More precisely, we select uniformly the upper left edge of the 128×128 image section and test if the proportion of pixels that are nonzero is at least as large as the proportion in the whole image. If this is not the case, we reject this section and repeat the process. An additional advantage of this approach is that we analyze the colocalization in image sections that are not only background. The OTC is robust to a varying size of randomly selected image sections (see Supplementary Figure A.6). For smaller image sections the OTC curves increase faster as the maximal distance on which mass can be transported is smaller. Here, we consider the fact that changing the size of the image sections does not affect the ordering of the OTC curves as robustness.

3.2 Statistical analysis of mean OTC curves

If we have access to several OTC curves for each combination of proteins, we propose to use the mean OTC curve. A pointwise confidence band can then be generated by bootstrap methods. It is necessary to use bootstrap methods as we cannot assume that our data is normally distributed at each fixed $t_0 \in [0, d_{max}]$, where $d_{max} = \max_{1 \leq i,j \leq N} \|x_i - x_j\|$ is the maximal distance between any two pixels in the image. There are several methods to derive confidence bands via bootstrap (Efron and Tibshirani, 1994). We will compare the bootstrap-t confidence bands (Efron, 1981), the bootstrap percentile confidence bands (Efron, 1979) and the BC_α confidence bands (Efron, 1987) exemplarily on $n = 10$ STED images of the protein Tom20 and the protein complex MICOS (exemplary images are displayed in Figure 3.3).

For each fixed threshold t_0, we have n observations of $OTC(t_0)$ which we use to calculate the mean OTC value $\overline{OTC}(t_0) = \frac{1}{n} \sum_{i=1}^{n} OTC(t_0)_i$. Here, the subscript i denotes the value of the i-th OTC curve evaluated at t_0. To derive the bootstrap-t confidence bands we resample from these n observations B times n observations $OTC(t_0)^*_1, \ldots, OTC(t_0)^*_n$ with replacement and evaluate for each of these samples the statistic

$$t_b = \frac{\overline{OTC}^*(t_0) - \overline{OTC}(t_0)}{s^* / \sqrt{n}}, \quad b = 1, \ldots, B, \tag{3.2}$$

where $\overline{OTC}^*(t_0)$ is the mean of the bootstrap sample and s^* its standard deviation. The $1 - \alpha$ bootstrap-t confidence interval is now given by

$$\left[\overline{OTC} - \frac{1}{\sqrt{n}} t^{(1-\alpha/2)} s, \overline{OTC} - \frac{1}{\sqrt{n}} t^{(\alpha/2)} s \right]. \tag{3.3}$$
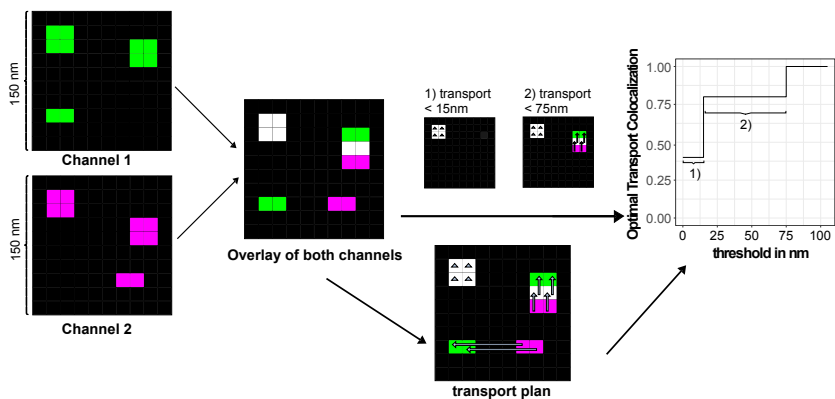
Here, s is the standard deviation of $OTC(t_0)_1, \ldots, OTC(t_0)_n$ and $t^{(1-\alpha/2)}$ is the $1 - \alpha/2$
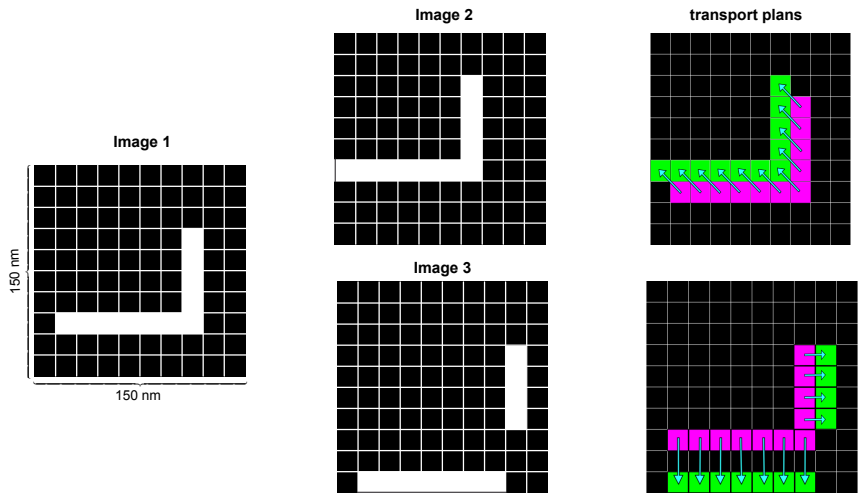
(a) Exemplary pixel images with structures that are colocalized on different scales. In the right column the transport plans of Image1 with Image2 and Image1 with Image3 are displayed via the light blue arrows.
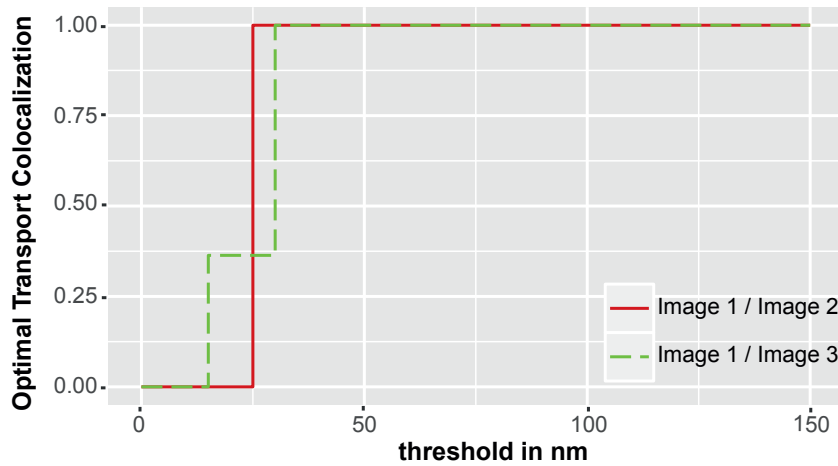
(b) OTC analysis of Image1/Image2 and Image1/Image3.

(c) Schematic for protocol to calculate OTC curves from data through optimal transport plan. White pixels indicate overlap between both channels. Arrows in the middle part represent the transport plan. In the case that the arrow is only a triangle the mass stays at that pixel.

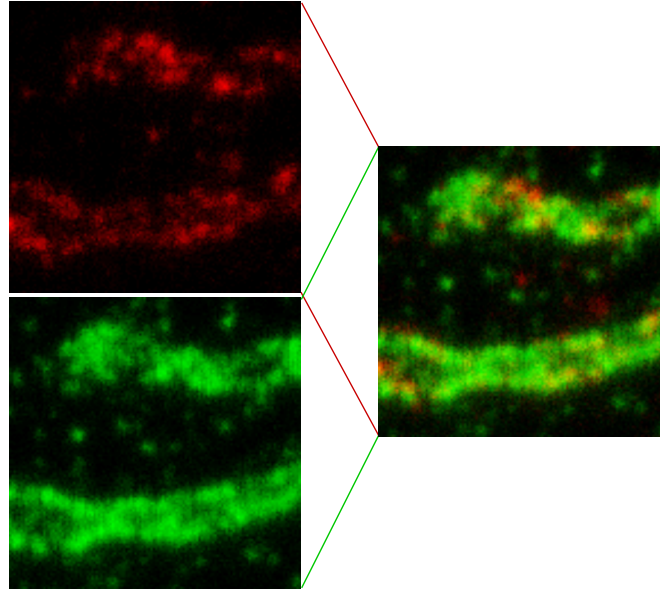Figure 3.2: **Optimal Transport Colocalization (OTC) analysis.**

Figure 3.3: **Exemplary STED images.** Left: STED image of Tom20 (red) and MICOS (green). Right: Overlay of both channels.

percentile of the distribution of the statistic t in (3.2). To get a pointwise confidence band for the whole curve, we apply this procedure for every $t_0 \in \{30, 45, \ldots, 240\}$.

For the bootstrap percentile confidence bands we generate B bootstrap replications of the mean and then take the $\alpha/2$ and $1 - \alpha/2$ percentile of this distribution. To be more precise, let $\hat{F}^{-1}$ be the empirical quantile function of the means. Then, the confidence interval is given by

$$\left[\hat{F}^{-1}\left(\tfrac{\alpha}{2}\right), \hat{F}^{-1}\left(1 - \tfrac{\alpha}{2}\right)\right]. \tag{3.4}$$

As for the bootstrap percentile confidence bands we generate B bootstrap replications of the OTC mean for the BC_α method. The confidence interval for this method given as

$$\left[\hat{F}^{-1}(\alpha_1), \hat{F}^{-1}(\alpha_2)\right], \tag{3.5}$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha/2)})}\right)$$

and

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha/2)})}\right).$$

Here, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal and $z^{(\alpha)}$
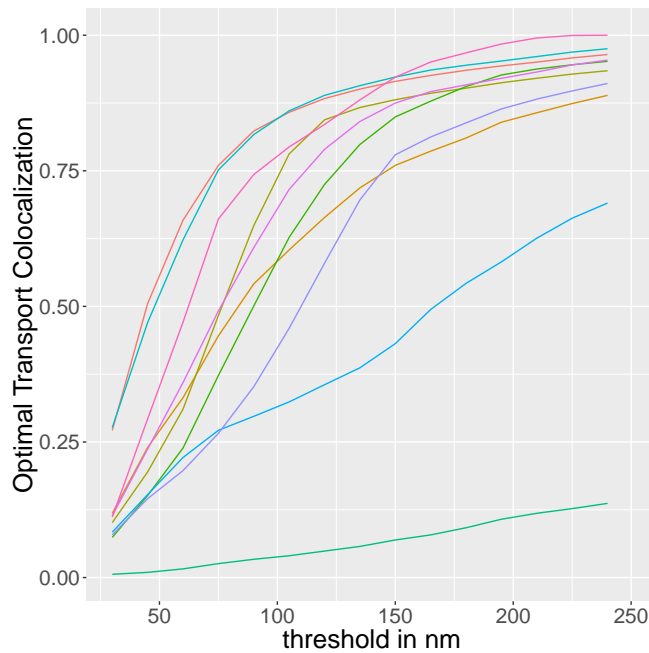
Figure 3.4: **OTC curves of** $n = 10$ **different STED image sections of Tom20 and MICOS.** Exemplary STED image sections are displayed in Figure 3.3.

the α percentile point of the standard normal. Further, $\hat{z}_0$ is defined as

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\overline{OTC}^* < \overline{OTC}\right\}}{B}\right)$$

with $\Phi^{-1}(\cdot)$ being the quantile function of the standard normal. Let $\overline{OTC}_{(i)}$ be the i-th jackknife value, i.e, the mean over the $n-1$ observations leaving out the i-th, and $\overline{OTC}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\overline{OTC}_{(i)}$ the mean jackknife value. Then, $\hat{a}$ is given by

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\overline{OTC}_{(\cdot)} - \overline{OTC}_{(i)}\right)^3}{6\left(\sum_{i=1}^{n}\left(\overline{OTC}_{(\cdot)} - \overline{OTC}_{(i)}\right)^2\right)^{3/2}}.$$

As mentioned we have $n = 10$ STED images of Tom20 and MICOS. All OTC curves are displayed in Figure 3.4. It is clearly visible that there are two curves which have a completely different shape compared to the rest. We generate for each of the above described methods $B = 1000$ bootstrap replications and set $\alpha = 0.05$. The three different pointwise confidence bands are depicted in Figure 3.5. We observe that the confidence bands generated by the bootstrap-t method are not range preserving and hence, drastically influenced by the two outliers. Furthermore, the percentile method gives smooth confidence bands compared to the BC_α method. These two methods yield
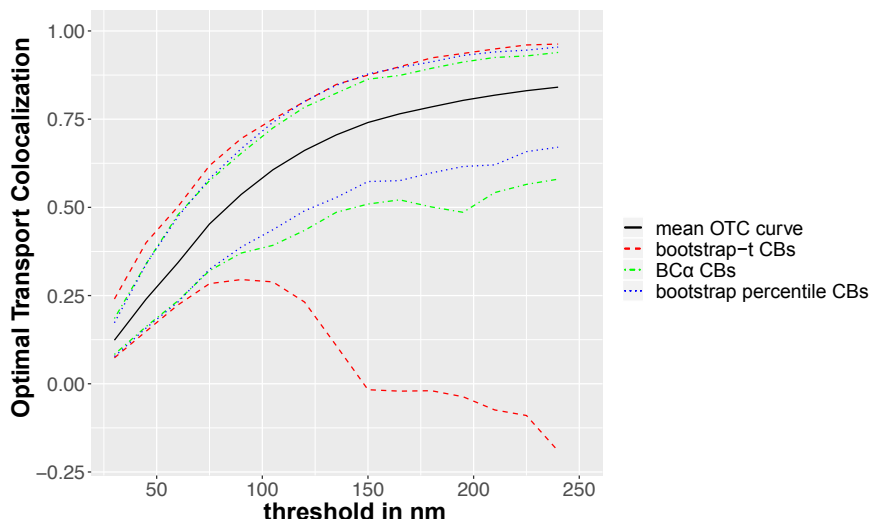
Figure 3.5: **Pointwise bootstrap confidence bands for mean OTC curve.** Confidence bands (CBs) generated from $B = 1000$ bootstrap replications, $\alpha = 0.05$. Displayed are the mean OTC curve (black solid line), the bootstrap-t CBs (red dashed lines), the bootstrap percentile CBs (blue dotted lines) and the BC_α CBs (green dash dotted lines).

range preserving confidence bands. As the BC_α method provides accurate coverage probability in contrast to the percentile method (Efron and Tibshirani, 1994), we suggest to use this method to evaluate the mean OTC curves statistically.

3.3 Results for different data applications

In this section, we analyze data sets with different characteristics by our OTC method. We begin with an analysis of confocal and STED data. Next, as a proof of concept and to validate our method we use data where the spatial proximity is known. Moreover, we investigate whether OTC curves are robust against background and last we evaluate if better results are achievable with the application of a 3D STED PSF compared to the application of a 2D STED PSF. For all data sets we use the BC_α method described in Section 3.2 to derive pointwise confidence bands for the mean OTC curves.

3.3.1 Comparison of OTC and conventional colocalization methods on STED and confocal data

To compare the performance of OTC with conventional colocalization coefficient (Manders' and Pearson's) we utilize confocal and STED images recorded on immunolabeled human cells. We label the cells for a protein in the mitochondrial outer membrane (Tom20) and one in the mitochondrial inner membrane (Mic60) (Figure 3.6(a)). A

simple overlay of the two confocal images reveals many white areas, i.e., areas where both signals are superimposed, suggesting that the labeled mitochondrial proteins are colocalized at this level of resolution. As expected, in the overlay of the STED images, however, only a few white dots remain, due to the higher resolution of around 40 nm. This dependence of resolution and colocalization as seen by visual inspection of the images, renders a quantitative determination of colocalization of STED image data challenging. We manually select sections from the image data sets as well as relied on the random selection mechanism described in Subsection 3.1.1. The manually selected sections contain mitochondria and are chosen based on structural preservation and signal to noise ratio. As the random selection mechanism singles out image regions containing fluorescence signals, the manually selected sections have a large overlap with the randomly selected sections (Suppl. Fig. A.1 and A.2). Both data sets are analyzed by Manders' M1/M2 and the two versions of Pearson's (usual Pearson's correlation and a thresholded version) (Figure 3.6(b)). For the thresholded Pearson's correlation coefficient the background of the images is first removed by setting all pixel intensities below a predetermined quantile to zero. This quantile is derived by iteratively setting all intensities above a given quantile to zero until the remaining pixel intensities are no longer correlated. We first determine on the manually selected sections the amount of colocalization with the three conventional colocalization methods and compare the findings for the confocal and STED data sets recorded on corresponding regions. The mean of Manders' M1 over 10 different manually selected image sections of the confocal data is 0.94 and the mean value of M2 is 0.9. This suggest that Mic60 and Tom20 are almost perfectly colocalized. The averaged Pearson's correlation coefficient is 0.73 and Pearson's correlation with threshold is 0.72 on average. The Manders' analysis of the same STED sections yields an amount of colocalization of 0.78 in M1 and 0.77 in M2. This value is considerably smaller than the amount of colocalization in the confocal data. The difference is even more drastic for the Pearson's methods showing a coefficient of 0.39 on the STED data. Hence, compared to the confocal images we move from the highly colocalized regime to a regime with almost no colocalization as the coefficients based on pixel intensity correlation can only detect the actual signal overlap. However, the two proteins are still in close proximity compared at the cellular scale. The considerably smaller values in the analysis of the STED data illustrate that the conventional colocalization methods which are based on pixel-intensity correlation are not well-suited for colocalization analysis of nanoscopy data sets. The analysis of the randomly chosen data sets yields comparable results (an average over 100 different randomly selected sections was analyzed). Together, the analysis reveals that the conventional colocalization methods report very different amounts of colocalization

when applied on diffraction limited or super-resolved data sets. Next, we analyze the data sets with our OTC method (Figure 3.6(c)). As described in Section 3.1, OTC is a curve that increases from zero to one. Over the range of thresholds t in the interval $[0, 2000]$ we find a maximal difference between the average OTC curves over 10 manually selected sections of 0.12 at a threshold of 105 nm. For the OTC analysis of 100 randomly selected sections we find a difference of 0.11 at a threshold of 90 nm. Hence, in contrast to the pixel intensity correlation based methods, the OTC analysis reveals only a slight difference between the manually and the randomly selected sections. In the zoomed region (inset Figure 3.6(c)) the OTC curves are displayed for thresholds t between 30 nm and 240 nm, which represents the characteristic range between the obtained resolution in the STED images and the resolution of the confocal images and hence, is the most interesting regime. Therefore, we will restrict the evaluation of the OTC curves to the range between 30 nm and 240 nm in all following analysis. Over this entire regime we find a higher amount of colocalization in the confocal recordings than in the STED images. The higher value in the confocal images is due to the blurring caused by diffraction as there are more pixels that contain mass and hence the transport takes place on smaller scales. Contrary to the established colocalization methods OTC can quantify spatial proximity even if fluorophores are detected in different pixels. Additionally, we display the 95% pointwise confidence band of the mean OTC curves for the manually selected image sections as well as for the randomly selected image sections. Both pointwise confidence bands are generated by the BC_α method (see Section 3.2) based on $B = 1000$ bootstrap replications. Interestingly, the pointwise confidence bands for the manually selected image section show that we can not find a statistically significant difference between the OTC curves of the confocal and the STED data. In contrast, for the randomly selected image sections we deduce a significant difference ($\alpha = 0.05$) on each individual spatial scale. To sum up, the conventional colocalization coefficients are not well suited for the analysis of nanoscopy data sets. Furthermore, OTC analysis reveals also a difference in the degree of colocalization of confocal data compared to STED data. This is due to the fact that the blurring in the confocal images leads to transport on smaller scales compared to the STED images.

3.3.2 Proof of concept on real STED data

To evaluate the OTC analysis for the quantification of colocalization in STED data, we recorded dual-color STED images from yeast mitochondria labeled for the mitochondrial protein Tom40 paired with the mitochondrial proteins Tom20, Cbp3 and Mrpl4 (Figure 3.7(a)) whose sub-mitochondrial distributions were previously investigated by cryo-electron microscopy, generating a ground truth data set (Stoldt et al., 2018). From
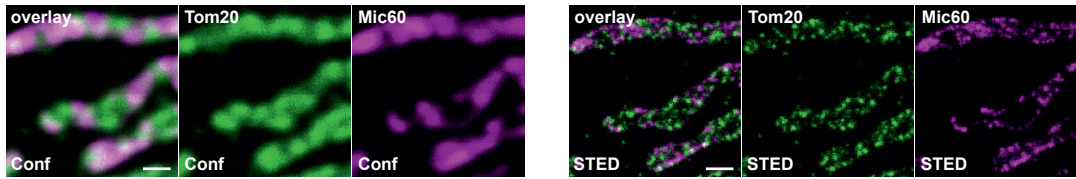
(a) Exemplary confocal and STED nanoscopy images of Tom20 and Mic60 in adult human dermal fibroblasts (HDFa) (Scale bar, 500 nm).
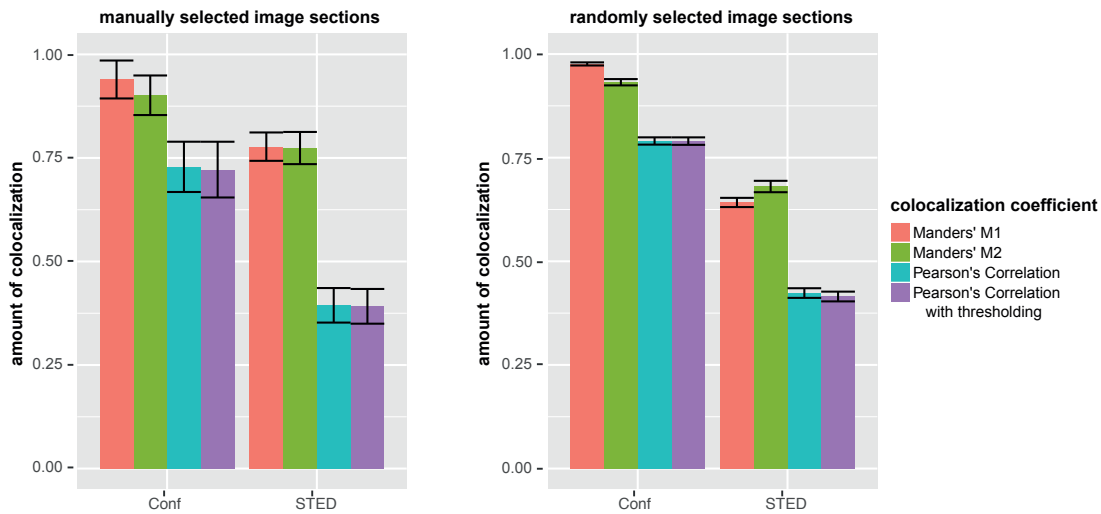
(b) Colocalization analysis of Tom20/Mic60 in HDFa cells using the Manders' and Pearson's correlation coefficients. Bars represent the mean (over 10 manually selected sections and 100 randomly selected sections), error bars are given in black.
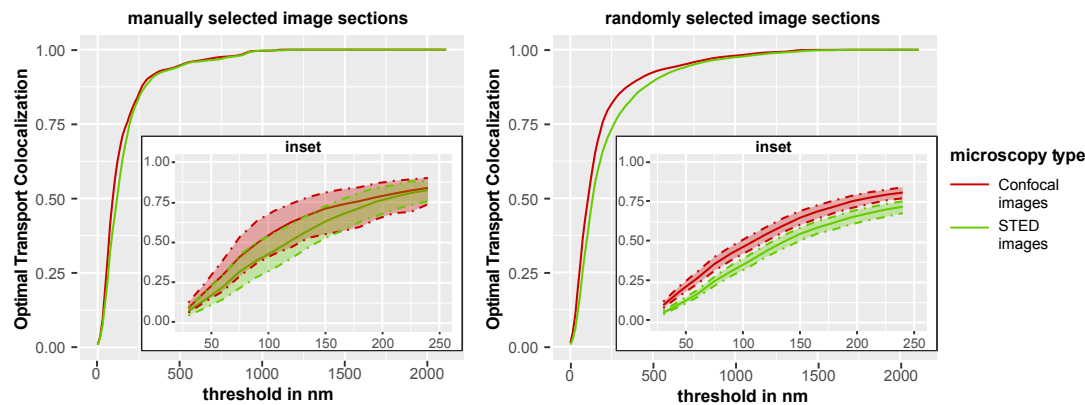
(c) OTC analysis of Tom20/Mic60 in HDFa Cells. 95% pointwise CBs based on $B = 1000$ bootstrap replications are depiced as colored areas.
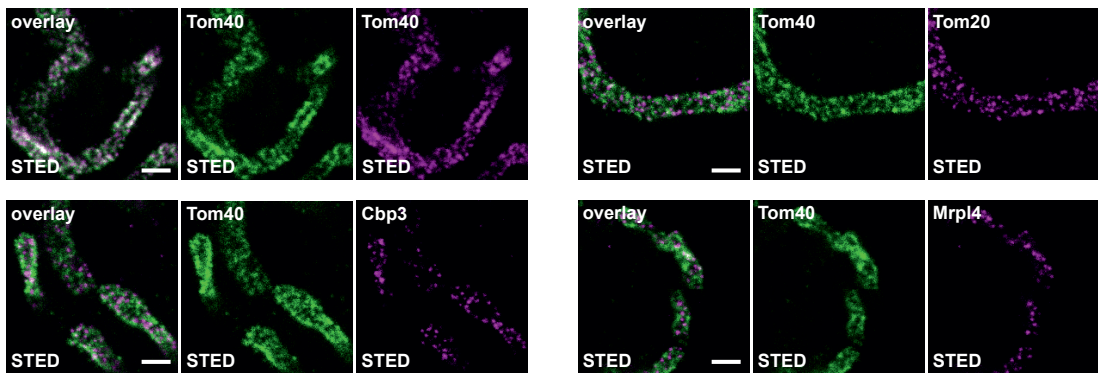
Figure 3.6: **Conventional methods for colocalization analysis versus OTC analysis of STED and confocal data.** In (b) and (c) the left part displays the analysis of manually selected sections and the right the analysis of the randomly selected sections.

this data set it is known that Tom40 and Tom20 have the highest spatial proximity, whereas Tom40 and Mrpl4 have the least proximity and Tom40 and Cbp3 are at an intermediate proximity range. In addition, as a control experiment, which represents highest empirically colocalization, we labeled Tom40 with two different markers. An analysis with the conventional methods was performed on manually and randomly selected image sections of the STED data. As we found again comparable results for the manually selected and the randomly selected sections, we will only describe the findings for the manually selected image sections. We found that the conventional colocalization coefficients report that Tom40/Tom40 have a higher degree of colocalization (M1: 0.9, M2: 0.96, Pearson's and Pearson's with threshold: 0.82) than the other three pairs (Figure 3.7(b)). It seems to be especially difficult to distinguish between the range of proximity of the pairs Tom40/Cbp3 and Tom40/Mrpl4. The M1 coefficient is 0.51 and 0.57, respectively. M2 yields colocalization degrees of 0.64 and 0.71. Hence, from Manders' colocalization coefficient we would deduce the wrong ordering of the proximity behavior. The two versions of Pearson's correlation are also not able to distinguish between the proximity of these two protein pairs. All conventional methods find a slightly higher amount of colocalization of Tom40/Tom20 compared to Tom40/Cbp3 and Tom40/Mrpl4 (M1: 0.71, M2: 0.79, Pearson's: 0.49, Pearson's with threshold: 0.48), which corresponds to the ground truth. In contrast, the OTC analysis on the same data sets reveals a difference between all four labeled pairs; the order of spatial proximity detected by the mean OTC curves matches the ground truth (Figure 3.7(c)). However, also the mean OTC curves show that the difference between Tom40/Mrpl4 and Tom40/Cbp3 is only marginal (in a range between 0.01 and 0.09). The difference increases for thresholds larger than 150 nm. Comparing the OTC analysis of the manually selected and the randomly selected image sections we find that the randomly selected sections give a better representation over the whole regime. We can deduce from these mean OTC curves that the difference in the degree of colocalization for the pair Tom40/Tom40 compared to the three other pairs is much larger (difference to mean OTC curve of Tom40/Tom20 in range between 0.06 and 0.23). The differences between the mean OTC curves of Tom40/Tom20 and Tom40/Cbp3 are between 0.03 and 0.07 and for Tom40/Cbp3 and Tom40/Mrpl4 between 0.007 and 0.055. Especially for the small scales ($t \in [30, 75]$) the degree of colocalization of Tom40/Mrpl4 is almost not distinguishable from the degree of colocalization of Tom40/Cbp3. In Supplementary Figure A.3 we display the mean OTC curves together with the 95% pointwise confidence bands generated by the BC_α method based on $B = 1000$ bootstrap replications. As the sample size $n = 10$ for the manually selected image sections is quite small, the pointwise confidence bands are rather large and hence, there is no significant difference between

the mean OTC curves considering all at once. However, there is a significant difference ($\alpha = 0.05$) between the mean OTC curve of Tom40/Tom40 and Tom40/Mrpl4 on each individual spatial scale. In contrast, for the randomly selected image sections there are several pointwise significant differences. The mean OTC curve of Tom40/Tom40 is at a significance level of 95% at each individual $t \in \{30, 45, \ldots, 240\}$ higher than the colocalization of all other protein pairs. Furthermore, the degree of colocalization of Tom40/Tom20 is also for each fixed t significantly higher than the one of Tom40/Mrpl4. In line with the findings of the mean curves there is no significant difference between Tom40/Mrpl4 and Tom40/Cbp3. In conclusion, the mean OTC curves represent the known spatial proximity correctly. However, especially in the case of manually selected sections the differences found are not significant.

3.3.3   Robustness against background

A common challenge of immunofluorescence microscopy is an unspecific background signal, which often complicates the analysis of the images. We labelled adult human dermal fibroblasts (HDFa) with antibodies against the mitochondrial proteins Tom20 and Mic60 or with antibodies against Tom20, Mic60, and Mic27. The antibody against Mic27 binds to the Mic60 interacting protein Mic27, and, in addition to unspecific structures in the cells. As a result, the cells labeled with the Mic27 antibody show a stronger background signal (Figure 3.8(a)). We ask the question if OTC can be used also to analyze such noisy data sets. Manually as well as randomly selected sections from the noisy data sets and the data sets that have a low background are analyzed with OTC (Figure 3.8(b)). Additionally, we display the 95% pointwise confidence bands for each curve which is generated by the BC_α method with $B = 1000$ bootstrap replications. For the manually selected section we find that the colocalization in the recordings with high background is a little higher than in the recordings with low background (maximal difference: 0.12). In contrast, for the randomly selected sections there is almost no difference (maximal difference: 0.02). From the pointwise confidence bands we can deduce that there is no significant difference in the low and high background setting at each individual spatial scale. Where Figure 3.7(b) already indicates a slightly better performance of OTC with randomly selected sections, we find a big evidence that the random selection mechanism performs better in the case of noisy data. The robustness of OTC against background is in line with the robustness of the conventional methods (see Suppl. Fig. A.4).

(a) Exemplary STED nanoscopy images of protein pairs with a known decrease of colocalization (Scale bar, 500 nm).
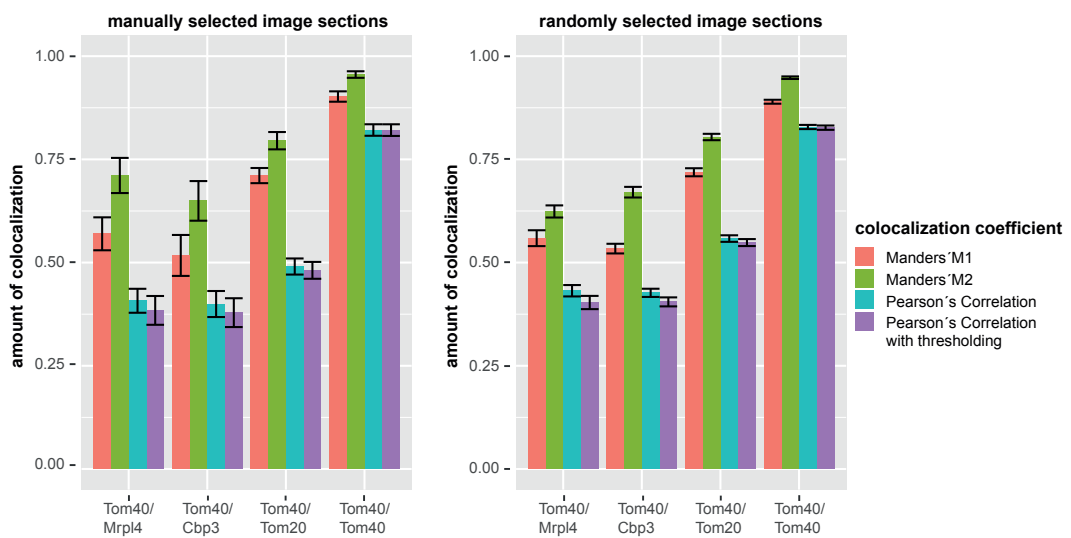
(b) Conventional colocalization analysis of the data represented in (a) using the Manders' and Pearson's colocalization coefficients.
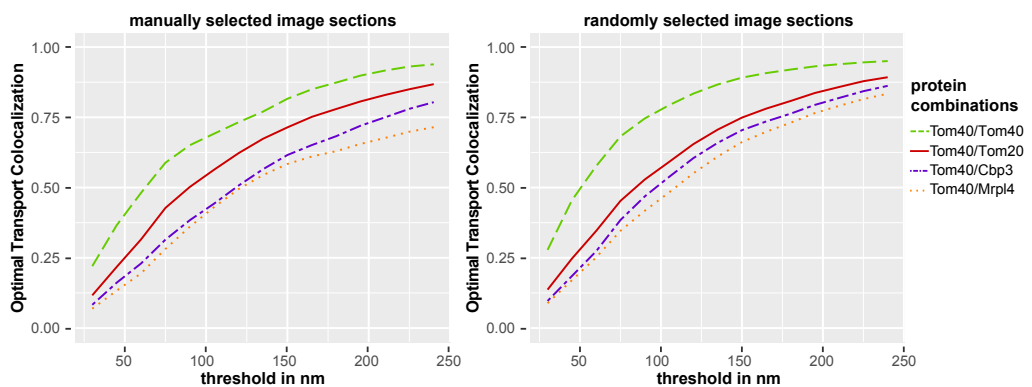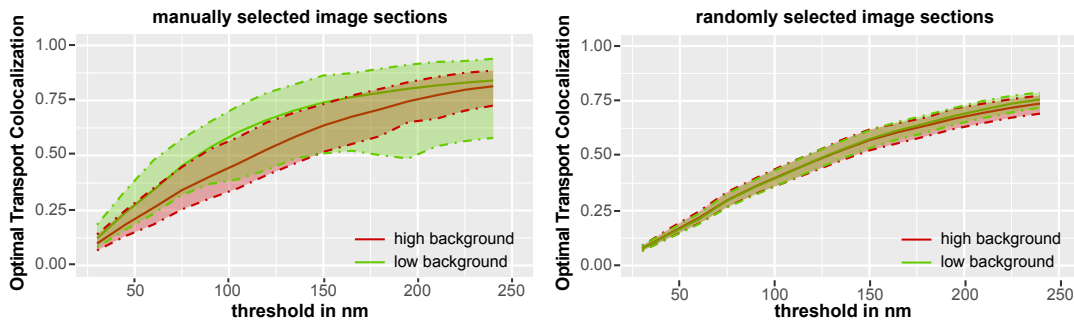

(c) OTC analysis of the data sets represented in (a).

Figure 3.7: **OTC analysis of protein pairs with known varying proximities to each other in yeast mitochondria.** In (b) and (c) the left part displays the analysis of manually selected sections and the right the analysis of the randomly selected sections.
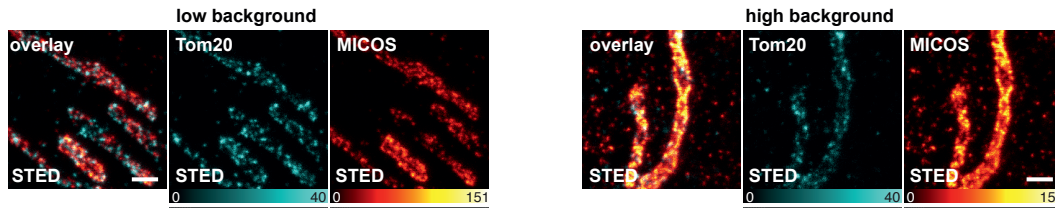
(a) Exemplary STED nanoscopy images of Tom20 and Mic60 (low background) or Tom20 and Mic60 plus Mic27 (high background) (Scale bar, 500nm).
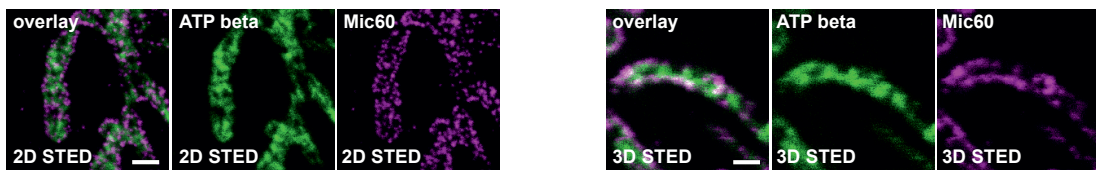

(b) OTC analysis of the data sets represented in (a). Left: Analysis of manually picked image slices. Right: Analysis of randomly selected image slices. 95% pointwise confidence bands based on $B = 1000$ bootstrap replications are depicted as colored areas.

Figure 3.8: **OTC offers a robust colocalization analysis even under suboptimal conditions.**

3.3.4 OTC analysis of images generated with 2D and 3D STED techniques

So far, we have analyzed STED images with a xy resolution of about 40 nm, as this method gives the highest optical resolution. Here, we compare OTC analysis of data sets generated with 2D and 3D STED techniques. We imaged human cells labeled for the inner membrane proteins Mic60 and a beta subunit of the F_1F_O-ATP synthase (ATP beta) with a STED microscope providing an almost uniform 3D resolution of ~80 nm in all room directions and in the 2D mode providing ~40 nm lateral resolution and ~500 nm axial resolution (Figure 3.9(a)). Mic60 is enriched at the cristae junctions, whereas the ATP beta is primarily localized in the cristae membrane. Therefore, Mic60 is localized at the rim of the tubular mitochondria, whereas the ATP beta is preferentially distributed in the organelle's interior. As 2D STED inherently makes a 2D projection of the mitochondrion, this might lead to erroneous high colocalization values. Contrary to visual impression of the 2D and 3D STED images, the colocalization between the 2D STED images should be higher than the colocalization between the 3D STED images. We analyzed manually and randomly selected sections from both data sets with OTC. Again, we also display the 95% pointwise confidence bands for all curves based on

$B = 1000$ bootstrap replications (Figure 3.9(b)). For the manually selected sections we find a small difference (~0.05) between the colocalization in the 2D and the 3D images for thresholds smaller than 175 nm. For thresholds between 175 nm and 250 nm we cannot deduce a difference. Here, the OTC analysis with randomly selected sections performs better as the difference in the colocalization between the 2D and 3D STED images is clearly visible. The difference between the curves ranges from 0.04 to 0.15. The pointwise confidence bands for the manually selected sections show that this difference is not significant. In contrast, the found difference for the randomly selected sections is significant ($\alpha = 5\%$) on each individual $t \in \{30, 45, \ldots, 240\}$. To sum up, the usage of the 3D STED PSF enhances the OTC colocalization analysis in this setting where proteins in a relatively thick organelle are imaged. On the contrary, the usage of the 3D PSF does not improve the colocalization analysis with conventional methods (see Suppl. Fig. A.5).

(a) Exemplary STED nanoscopy images of ATP beta and Mic60. A 2D STED PSF (superior resolution along the X and Y axes but no improvement along the Z axis) was utilized to generate the images shown on the left side. A 3D STED PSF (isotropic resolution improvement) was applied to generate the images shown on the right side (Scale bar, 500nm).
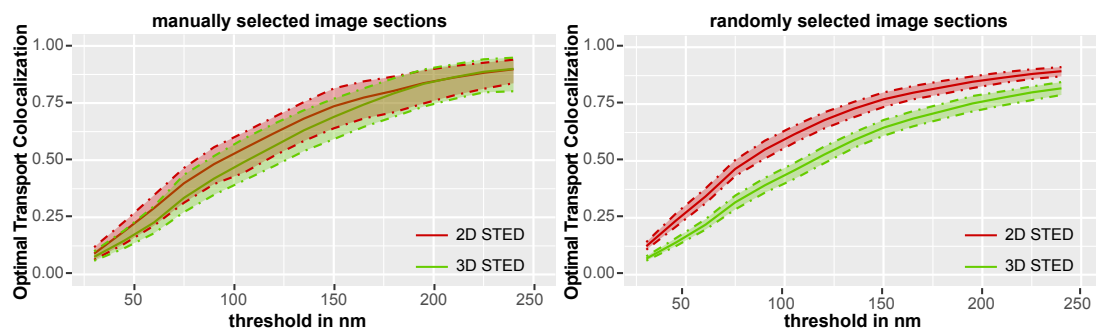
(b) OTC analysis of the data sets represented in (a). Analysis of manually picked image sections (left) and analysis of randomly selected image sections (right). 95% pointwise confidence bands based on $B = 1000$ bootstrap replications are depicted as colored areas.

Figure 3.9: **The application of a 3D STED PSF enhances colocalization analysis.**

3.4 Discussion

As nanoscopy provides much higher spatial resolution compared to conventional light microscopy the areas with high intensity overlap are drastically reduced. Therefore, pixel intensity correlation based colocalization coefficients, which mainly measure the actual signal overlap, yield only small values and are not well suited for detailed and meaningful analysis of nanoscopy data. In contrast, OTC analysis provides a curve which displays the amount of relative proximity in terms of the minimal effort to match two given protein distributions as a function of spatial distance. This is especially useful for STED data because the raw data is given by pixel images as the recording depends on pixel-scanning steps. This renders OTC analysis also a useful tool for other scanning-based nanoscopy methods, like for example RESOLFT. We speculate that OTC analysis performs also well on preprocessed data from coordinate-stochastic nanoscopy. If one wants to apply one of the object-based methods developed for colocalization analysis of coordinate-stochastic super-resolution data on STED data, one first needs to estimate the locations of the proteins from the raw data. This introduces a statistical error. If one applies, e.g. k-nearest neighbors directly on the pixels this does not incorporate the pixel intensities. OTC takes care of the intensities as it matches intensity distributions in an optimal way. Furthermore, for a chosen k the k-nearest neighbor method takes only one scale of magnitude k into account, whereas the OTC is evaluated over all spatial scales and is able to match corresponding pixels *across scales* (see Figure 3.2(c)). Bearing in mind the challenges of colocalization analysis in super-resolution light microscopy it seems to be prudent to reassess the concept of colocalization for nanoscopy in general. Whereas in diffraction limited light microscopy colocalization is seen as a spatial correlation of pixel intensities, it seems to be more appropriate to speak of relative spatial proximities of protein distributions when it comes to nanoscopy data sets. Additionally, OTC analysis offers the possibility to set distance thresholds depending on a priori biological knowledge. Therefore, the solution of the optimal transport problem in the form of OTC analysis represents a promising new approach.

Recalling the results from the proof of concept section and the robustness section we can deduce that OTC is able to distinguish the degree of colocalization for very similar protein pairs (see Figure 3.6) and that this method is robust against background (see Figure 3.8). In contrast to conventional methods (see Suppl. Figure A.5), the performance of OTC analysis is even enhanced by the application of a 3D STED PSF (see Figure 3.9) and therefore enables users to utilize the full potential of modern 3D nanoscopy methods like 3D STED.
OTC can be calculated fully automated. The automated selection of image sections for

OTC performed even better than the manual selection. Figures 3.6 - 3.9 demonstrate strong robustness of the concept. We selected image sections of 128×128 pixels, which takes a few seconds on a standard laptop with the shielding algorithm from Schmitzer (2016) implemented in the *transport* package (Schuhmacher et al., 2017). Our random selection scheme overcomes the computational burden to evaluate the optimal transport problem on the full image as it selects randomly parts of the image of a computable size.

CHAPTER 4

Discussion and outlook

This thesis considered two different topics related to optimal transport.

First, we derived limit laws for the empirical optimal transport distance on countable spaces $\mathcal{X}$. The empirical Wasserstein distance $W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m)$ is the plug in estimate of the empirical measures generated by i.i.d. data $X_1, \ldots, X_n \sim \mathbf{r}$ and $Y_1, \ldots, Y_m \sim \mathbf{s}$ independent of the X_i's. Limit results were derived in the one-sample case as well as in the two-sample case. In both cases we were able to give results under the assumption of equality of measures, i.e., $\mathbf{r} = \mathbf{s}$ and for different underlying measures, i.e., $\mathbf{r} \neq \mathbf{s}$. We found a necessary and sufficient condition on the underlying measures that has to hold for these distributional limits if the diameter of the ground space is not finite. This condition was extensively discussed. Further, we found that the scaling rate under equality of measures is not the same as the scaling rate for different underlying measures.

All derived limit distributions are given implicitly via a maximization problem. We were able to calculate this maximum explicitly for $\mathcal{X}$ being a tree in the case that the underlying measures are the same. This explicit limit was used to upper bound the limiting distribution on general spaces. Simulation studies investigated the accuracy of this upper bound.

Second, we proposed a method based on optimal transport - OTC - to measure colocalization in super-resolution images. For the OTC method we presented methods to evaluate these curves statistically and validated correctness of the results deduced by the OTC curves by a proof of concept on STED data of proteins with known proximity. We further investigated the robustness of OTC against background and the benefits of using a 3D STED PSF instead of an 2D PSF.

In the following we discuss further possible research questions that could be investigated.

Distributional limits in higher dimensions   Distributional limits for the empirical
Wasserstein distance for general continuous probability measures in higher dimensions
under equality of measures are still an open research question. We already observed
different scaling rates under equality of measures and under different underlying
measures. This finding is in line with the findings of Ajtai et al. (1984), who get
a rate of $\sqrt{n \log(n)}$ for the uniform measure on the unit square in $\mathbb{R}^2$ and a sample from
this measure, which shows the intrinsic difficulty of this problem. Another point that
suggests, that this problem is quite hard to solve is the fact that already on countable
spaces a careful calibration of the norm was needed to get both, differentiability and
weak convergence.

Distributional limits for the optimal transport plan   Another interesting area would
be to derive limit laws for the empirical optimal transport plan on finite and countable
spaces. These results would give access to confidence statements for each entry in the
optimal transport plan and hence, we would be able to derive simultaneous confidence
bands for the OTC curves. To investigate this in more detail, recall, the definition of
OTC in (3.1)

$$OTC(t) = \sum_{i,j=1}^{N} \mathbb{1}\{\|x_i - x_j\| \leq t\} w_{ij}^*,$$

where w^* is the optimal solution, i.e., optimal transport plan, of (1.9) with the Euclidean
distance and $p = 2$. OTC can be seen as an operator from $\mathbb{R}^{N \times N}$ to the càdlàg space
$\mathcal{D}[0, d_{max}]$ (Billingsley, 2013) on $[0, d_{max}]$, where $d_{max} = \max_{1 \leq i,j \leq N} \|x_i - x_j\|$ is the
maximal distance between any two pixels in the images. More precisely, we can write

$$OTC \colon \mathbb{R}_+^{N \times N} \to \mathcal{D}[0, d_{max}], \quad w \mapsto \sum_{i,j=1}^{N} \mathbb{1}_{\{\|x_i - x_j\| \leq \cdot\}} w_{ij}$$

The operator OTC is linear and Lipschitz continuous with constant at most one.
Therefore, we would be able to derive distributional limit results for the empirical
counterpart of OTC by the continuous mapping theorem if limit results for the empirical
optimal transport plan are known.

Wasserstein barycenter   Wasserstein barycenters, i.e., the Fréchet mean in the Was-
serstein space, are a widely used tool for the analysis of complex data set. For this
reason, deriving limit results for empirical Wasserstein barycenters is an interesting area
of research to use the barycenters in any statistical application. The barycenter problem
on finite spaces can also be written as the optimal solution of a linear program. So one
hopes to be able to apply similar techniques as in the case of distributional limits for the

optimal transport plan.

Quantization   The upper bound of the limiting distribution by the explicit results for the trees falls into the field of quantization. Going deeper into this research direction, we could ask for theoretical bounds to evaluate the accuracy of this upper bound to complement the results from the simulation study.

CHAPTER A

Appendix

A.1   Supplementary figures to Chapter 3
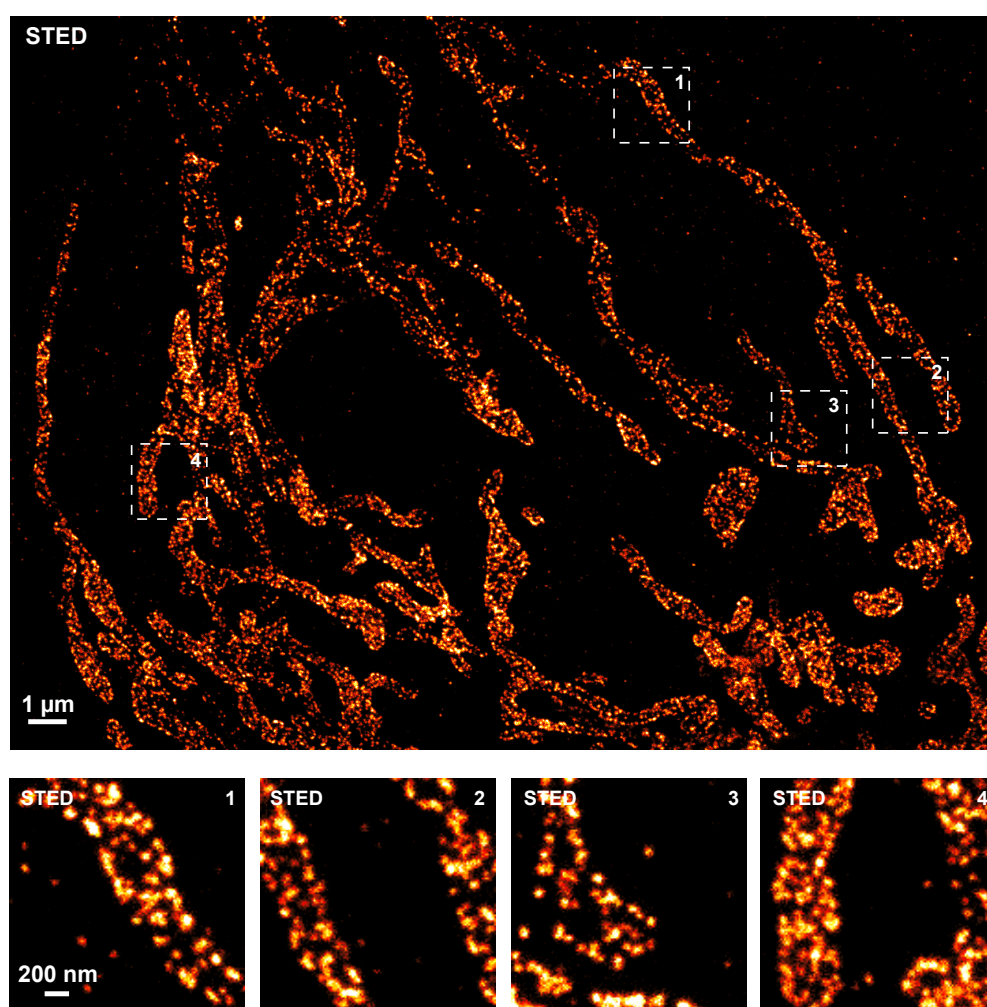
Supplementary Fig. A.1: Example for **manually selected** image sections.
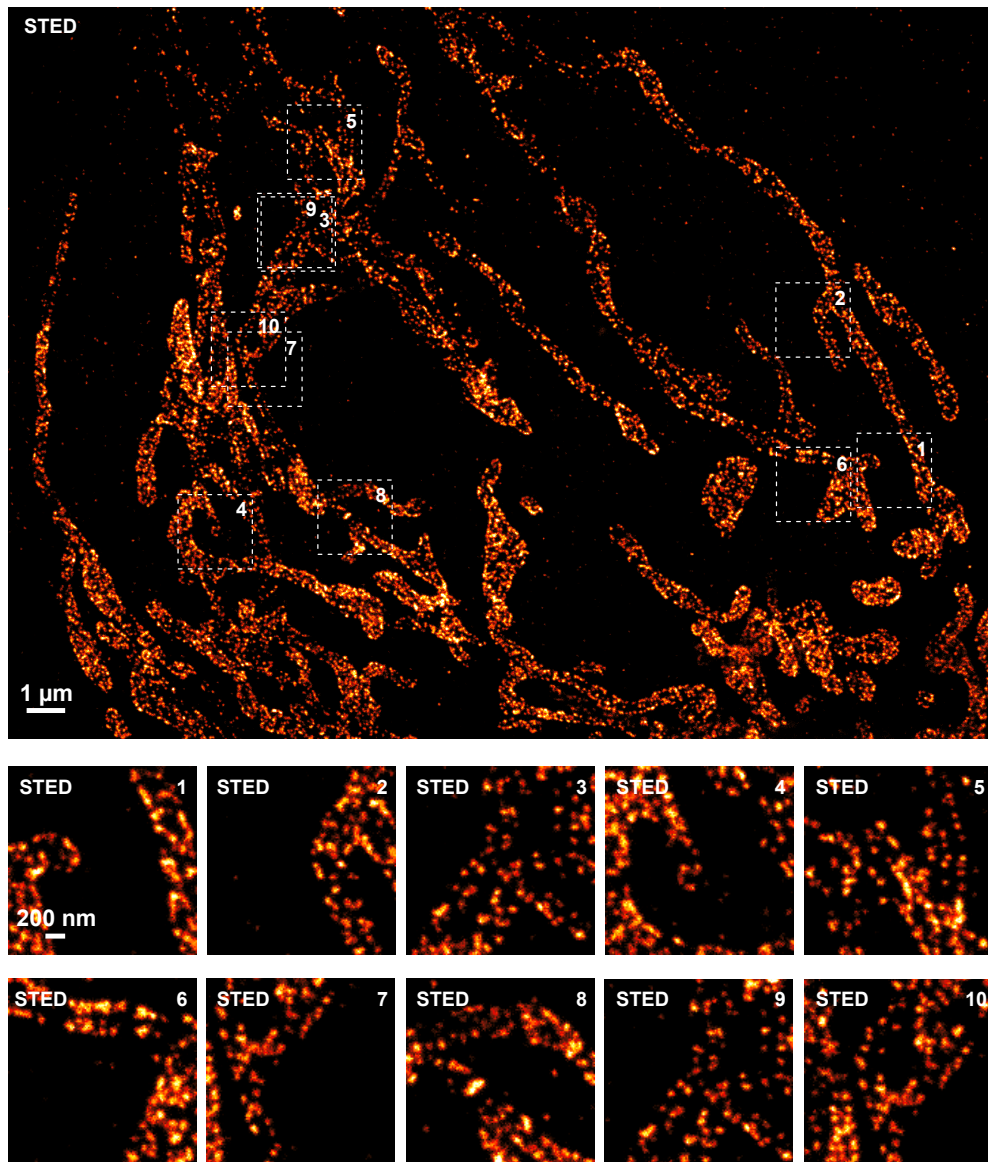
Supplementary Fig. A.2: Example for **randomly selected** image sections.
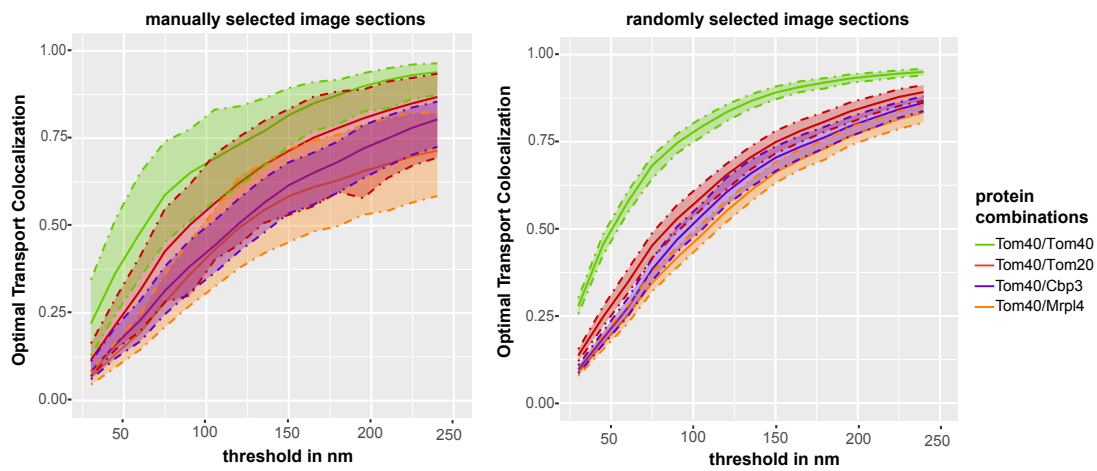
Supplementary Fig. A.3: **Mean OTC curves from Figure 3.7 together with** 95% **pointwise CBs** based on $B = 1000$ bootstrap replications.
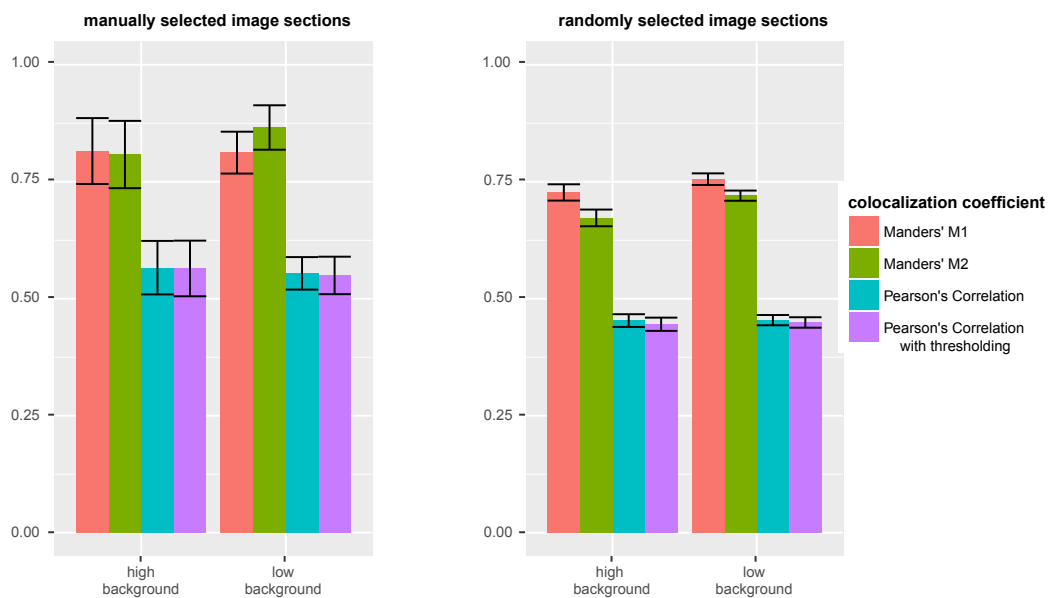
Supplementary Fig. A.4: **Conventional colocalization analysis** of the data with low and high background from Figure 3.8.
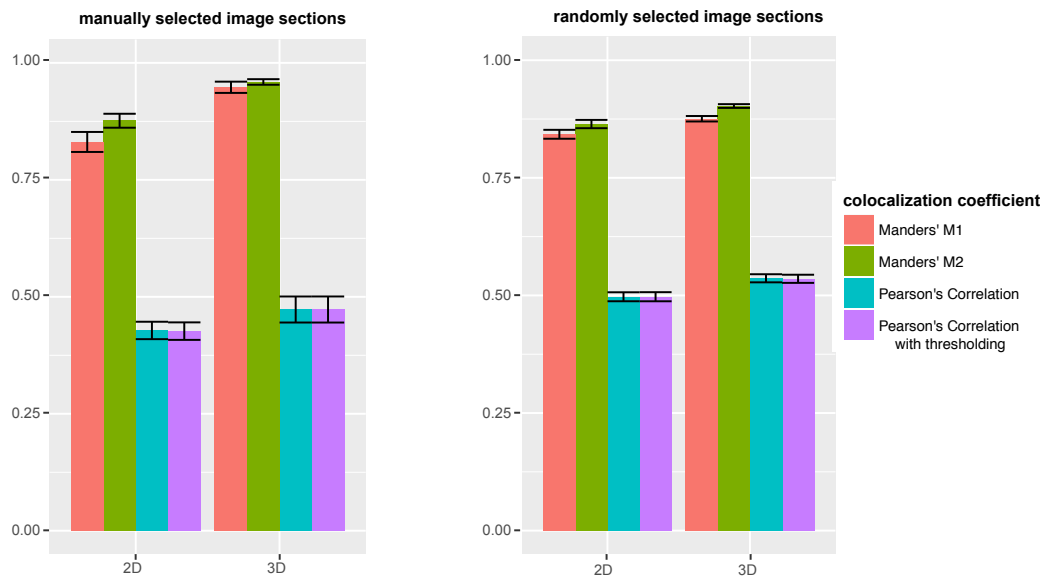
Supplementary Fig. A.5: **Conventional colocalization analysis** of the data recorded with 2D and 3D STED PSF from Figure 3.9.
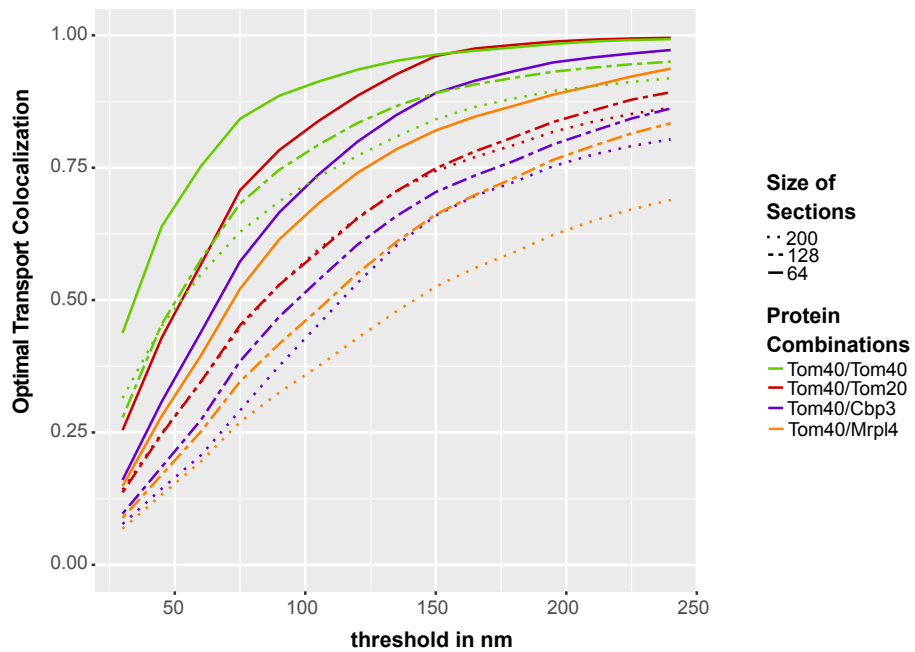
Supplementary Fig. A.6: Optimal Transport Colocalization analysis of yeast data for **different section sizes**.

Bibliography

Adler, J. and Parmryd, I. (2010). Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytometry Part A*, 77A(8):733–742.

Agnati, L. F., Fuxe, K., Torvinen, M., Genedani, S., Franco, R., Watson, S., Nussdorfer, G. G., Leo, G., and Guidolin, D. (2005). New methods to evaluate colocalization of fluorophores in immunocytochemical preparations as exemplified by a study on A2A and D2 receptors in chinese hamster ovary cells. *Journal of Histochemistry & Cytochemistry*, 53(8):941–953.

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

Ambrosio, L. (2003). Lecture notes on optimal transport problems. In *Mathematical Aspects of Evolving Interfaces*, volume 1812, pages 1–52. Springer Berlin Heidelberg, Berlin, Heidelberg.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv:1701.07875 [cs, stat]*.

Barbour, A. D. and Brown, T. C. (1992). Stein's method and point process approximation. *Stochastic Processes and their Applications*, 43(1):9–31.

Barlow, A. L., MacLeod, A., Noppen, S., Sanderson, J., and Guérin, C. J. (2010). Colocalization analysis in fluorescence micrographs: Verification of a more accurate calculation of Pearson's correlation coefficient. *Microscopy and Microanalysis*, 16(06):710–724.

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.

Billingsley, P. (2013). *Convergence of Probability Measures*. Wiley.

Blom, H., Rönnlund, D., Scott, L., Spicarova, Z., Rantanen, V., Widengren, J., Aperia, A., and Brismar, H. (2012). Nearest neighbor analysis of dopamine D1 receptors and Na(+)-K(+)-ATPases in dendritic spines dissected by STED microscopy. *Microscopy Research and Technique*, 75(2):220–228.

Bobkov, S. and Ledoux, M. (2016). One-dimensional empirical measures, order statistics and Kantorovich transport distances. *Memoirs of the Amer. Math. Soc.* to appear.

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.

Bolte, S. and Cordelières, F. P. (2006). A guided tour into subcellular colocalization analysis in light microscopy. *Journal of Microscopy*, 224(3):213–232.

Bonnans, J. F. and Shapiro, A. (2000). *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research and Financial Engineering. Springer-Verlag, New York.

Coltharp, C., Yang, X., and Xiao, J. (2014). Quantitative analysis of single-molecule superresolution images. *Current Opinion in Structural Biology*, 28:112–121.

Costes, S. V., Daelemans, D., Cho, E. H., Dobbin, Z., Pavlakis, G., and Lockett, S. (2004). Automatic and quantitative measurement of protein-protein colocalization in live cells. *Biophysical Journal*, 86(6):3993–4003.

de Wet, J. H. and Venter, T. . (1972). Asymptotic distributions of certain test criteria of normality. *South African Statistical Journal*, 6(2):135–149.

Dede, S. (2009). An empirical central limit theorem in L1 for stationary sequences. *Stochastic Processes and their Applications*, 119(10):3494–3515.

Dedecker, J. and Merlevede, F. (2015). Behavior of the Wasserstein distance between the empirical and the marginal distributions of stationary alpha-dependent sequences. *arXiv preprint arXiv:1503.00113*.

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., Csörgö, S., Cuadras, C. M., de Wet, T., Giné, E., Lockhart, R., Munk, A., and Stute, W. (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test*, 9(1):1–96.

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., and Rodríguez-Rodríguez, J. M. (1999a). 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 Matrán, C. (1999b). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *The Annals of Probability*, 27(2):1009–1071.

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. and Loubes, J.-M. (2017). Central limit theorems for empirical transportation cost in general dimension. *arXiv:1705.01299 [math, stat]*.

Demandolx, D. and Davoust, J. (1997). Multicolour analysis and local image correlation in confocal microscopy. *Journal of Microscopy*, 185(1):21–36.

Dobrić, V. and Yukich, J. E. (1995). Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118.

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

Dunn, K. W., Kamocka, M. M., and McDonald, J. H. (2011). A practical guide to evaluating colocalization in biological microscopy. *American Journal of Physiology-Cell Physiology*, 300(4):C723–C742.

Eberle, A. (2014). Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337–377.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 9(2):139–158.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Number 57 in Monographs on statistics and applied probability. Chapman & Hall, New York.

Eggert, D., Rösch, K., Reimer, R., and Herker, E. (2014). Visualization and analysis of hepatitis C virus structural proteins at lipid droplets by super-resolution microscopy. *PLOS ONE*, 9(7):e102511.

Evans, L. C. and Gangbo, W. (1999). Differential equations methods for the Monge-Kantorovich mass transfer problem. *Memoirs of the American Mathematical Society*, 137(653):0–0.

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.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738.

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.

Georgieva, M., Cattoni, D. I., Fiche, J.-B., Mutin, T., Chamousset, D., and Nollmann, M. (2016). Nanometer resolved single-molecule colocalization of nuclear factors by two-color super resolution microscopy imaging. *Methods (San Diego, Calif.)*, 105:44–55.

Herce, H. D., Casas-Delucchi, C. S., and Cardoso, M. C. (2013). New image colocalization coefficient for fluorescence microscopy to quantify (bio-)molecular interactions. *Journal of Microscopy*, 249(3):184–194.

Hess, S. T., Girirajan, T. P. K., and Mason, M. D. (2006). Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical Journal*, 91(11):4258–4272.

Hitchcock, F. L. (1941). The distribution of a product from several sources to numerous localities. *Journal of Mathematics and Physics*, 20(1-4):224–230.

Hofmann, M., Eggeling, C., Jakobs, S., and Hell, S. W. (2005). Breaking the diffraction barrier in fluorescence microscopy at low light intensities by using

reversibly photoswitchable proteins. *Proceedings of the National Academy of Sciences*, 102(49):17565–17569.

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.

Huang, B., Bates, M., and Zhuang, X. (2009). Super-resolution fluorescence microscopy. *Annual Review of Biochemistry*, 78:993–1016.

Humpert, F., Yahiatène, I., Lummer, M., Sauer, M., and Huser, T. (2015). Quantifying molecular colocalization in live cell fluorescence microscopy. *Journal of Biophotonics*, 8(1-2):124–132.

Jain, N. C. (1976). Central limit theorem and related questions in Banach space. In *Probability (Proc. Sympos. Pure. Math., Vol. XXXI, Univ. Illinois, Urbana, Ill., 1976*, pages 55–65. American Mathematical Soc.

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

Kantorovich, L. (1958). On the translocation of masses. *Management Science*, 5(1):1–4.

Kantorovich, L. V. (1948). On a problem of Monge. *Uspekhi Mat. Nauk*, 3(2):225–226.

Kantorovich, L. W. and Rubinshtein, G. S. (1958). On a space of totally additive functions. *Vestnik Leningradskogo Universiteta*, 13(7):52–59.

Klar, T. A., Jakobs, S., Dyba, M., Egner, A., and Hell, S. W. (2000). Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Proceedings of the National Academy of Sciences*, 97(15):8206–8210.

Koopmans, T. C. (1949). Optimum utilization of the transportation system. *Econometrica*, 17:136.

Lagache, T., Sauvonnet, N., Danglot, L., and Olivo-Marin, J.-C. (2015). Statistical analysis of molecule colocalization in bioimaging. *Cytometry Part A*, 87(6):568–579.

Landmann, L. and Marbet, P. (2004). Colocalization analysis yields superior results after image restoration. *Microscopy Research and Technique*, 64(2):103–112.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Springer-Verlag, New York, 2 edition.

Lehmann, M., Rocha, S., Mangeat, B., Blanchet, F., Uji-i, H., Hofkens, J., and Piguet, V. (2011). Quantitative multicolor super-resolution microscopy reveals Tetherin HIV-1 interaction. *PLoS Pathogens*, 7(12):e1002456.

Li, Q., Lau, A., Morris, T. J., Guo, L., Fordyce, C. B., and Stanley, E. F. (2004). A Syntaxin 1, Gαo, and N-Type calcium channel complex at a presynaptic nerve terminal: Analysis by quantitative immunocolocalization. *Journal of Neuroscience*, 24(16):4070–4081.

Malkusch, S., Endesfelder, U., Mondry, J., Gelléri, M., Verveer, P. J., and Heilemann, M. (2012). Coordinate-based colocalization analysis of single-molecule localization microscopy data. *Histochemistry and Cell Biology*, 137(1):1–10.

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

Manders, E. M., Stap, J., Brakenhoff, G. J., van Driel, R., and Aten, J. A. (1992). Dynamics of three-dimensional replication patterns during the S-phase, analysed by double labelling of DNA and confocal microscopy. *Journal of Cell Science*, 103 (Pt 3):857–862.

Manders, E. M. M., Verbeek, F. J., and Aten, J. A. (1993). Measurement of co-localization of objects in dual-colour confocal images. *Journal of Microscopy*, 169(3):375–382.

Mason, D. M. (2016). A weighted approximation approach to the study of the empirical Wasserstein distance. In *High Dimensional Probability VII*, Progress in Probability, pages 137–154. Birkhäuser, Cham.

Maurey, B. (1972). Espaces de cotype p, 0 < p 2. In *Séminaire Maurey-Schwartz (année 1972–1973), Espaces Lp et applications radonifiantes, Exp. No. 7*, pages 1–11. Centre de Math., École Polytech., Paris.

Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale.

Moser, B., Hochreiter, B., Herbst, R., and Schmid, J. A. (2017). Fluorescence colocalization microscopy analysis can be improved by combining object-recognition with pixel-intensity-correlation. *Biotechnology Journal*, 12(1):1600332.

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.

Neumann, D., Bückers, J., Kastrup, L., Hell, S. W., and Jakobs, S. (2010). Two-color STED microscopy reveals different degrees of colocalization between hexokinase-I and the three human VDAC isoforms. *PMC Biophysics*, 3:4.

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.

Osterwald, S., Wörz, S., Reymann, J., Sieckmann, F., Rohr, K., Erfle, H., and Rippe, K. (2012). A three-dimensional colocalization RNA interference screening platform to elucidate the alternative lengthening of telomeres pathway. *Biotechnology Journal*, 7(1):103–116.

Panaretos, V. M. and Zemel, Y. (2016). Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812.

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

R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.r-project.org/.

Rachev, S. T. and Rüschendorf, L. (1994). On the rate of convergence in the CLT with respect to the Kantorovich metric. In Hoffmann-Jørgensen, J., Kuelbs, J., and Marcus, M. B., editors, *Probability in Banach Spaces, 9*, Progress in Probability, pages 193–207. Birkhäuser Boston.

Rachev, S. T., Stoyanov, S. V., and Fabozzi, F. J. (2011). *A Probability Metrics Approach to Financial Risk Measures*. John Wiley & Sons.

Ramdas, A., Trillos, N., Cuturi, M., Ramdas, A., Trillos, N. G., and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47.

Rippl, T., Munk, A., and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151(C):90–109.

Rolet, A., Cuturi, M., and Peyré, G. (2016). Fast dictionary learning with a smoothed Wasserstein loss. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 630–638.

Römisch, W. (2004). Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc.

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.

Rudolf, D. and Schweizer, N. (2015). Perturbation theory for Markov chains via Wasserstein distance. *arXiv:1503.04123 [math, stat]*.

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.

Sahl, S. J., Hell, S. W., and Jakobs, S. (2017). Fluorescence nanoscopy in cell biology. *Nature Reviews. Molecular Cell Biology*, 18(11):685–701.

Schmitzer, B. (2016). A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259.

Schrieber, J., Schuhmacher, D., and Gottschlich, C. (2017). DOTmark - a benchmark for discrete optimal transport. *IEEE Access*, 5:271–282.

Schuhmacher, D. (2009). Stein's method and Poisson process approximation for a class of Wasserstein metrics. *Bernoulli*, 15(2):550–568.

Schuhmacher, D., Bähre, B., Gottschlich, C., Heinemann, F., Schmitzer, B., and Wilm, T. (2017). Transport: Optimal Transport in Various Forms. R package version 0.9-4.

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.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

Sommerfeld, M. (2017). *Wasserstein Distance on Finite Spaces: Statistical Inference and Algorithms*. PhD thesis, Georg-August Universität, Göttingen.

Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238.

Stoldt, S., Wenzel, D., Kehrein, K., Riedel, D., Ott, M., and Jakobs, S. (2018). Spatial orchestration of mitochondrial translation and OXPHOS complex assembly. *Nature Cell Biology*, 20(5):528–534.

Sudakov, V. N. (1979). *Geometric Problems in the Theory of Infinite-Dimensional Probability Distributions*. American Mathematical Soc.

Talagrand, M. (1992). Matching random samples in many dimensions. *The Annals of Applied Probability*, 2(4):846–856.

Talagrand, M. (1994). The transportation cost from the uniform measure to the empirical measure in dimension >= 3. *The Annals of Probability*, 22(2):919–959.

Tameling, C. and Munk, A. (2018). Computational strategies for inference based on empirical optimal transport. *IEEE SigPort*.

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]*.

Tameling, C., Stoldt, S., Stephaen, T., Jakobs, S., and Munk, A. (2018). Statistical quantification of colocalization for coordinate-targeted super-resolution microscopy via optimal transport. *preprint*.

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

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

Villani, C. (2008). *Optimal Transport: Old and New*. Springer Science & Business Media.

Wang, S., Arena, E. T., Becker, J. T., Bement, W. M., Sherer, N. M., Eliceiri, K. W., and Yuan, M. (2017). Spatially adaptive colocalization analysis in dual-color fluorescence microscopy. *arXiv:1711.00069 [eess, q-bio, stat]*.

Wang, S., Arena, E. T., Eliceiri, K. W., and Yuan, M. (2018). Automated and robust quantification of colocalization in dual-color fluorescence microscopy: A nonparametric statistical approach. *IEEE Transactions on Image Processing*, 27(2):622–636.

Wang, S., Fan, J., Pocock, G., and Yuan, M. (2016). Structured correlation detection with application to colocalization analysis in dual-channel fluorescence microscopic imaging. *arXiv:1604.02158 [math.ST]*.

Weed, J. and Bach, F. (2017). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv:1707.00087 [math, stat]*.

Worz, S., Sander, P., Pfannmoller, M., Rieker, R. J., Joos, S., Mechtersheimer, G., Boukamp, P., Lichter, P., and Rohr, K. (2010). 3D geometry-based quantification of colocalizations in multichannel 3D microscopy images of human soft tissue tumors. *IEEE Transactions on Medical Imaging*, 29(8):1474–1484.

Xu, L., Rönnlund, D., Aspenström, P., Braun, L. J., Gad, A. K. B., and Widengren, J. (2016). Resolution, target density and labeling effects in colocalization studies - suppression of false positives by nanoscopy and modified algorithms. *The FEBS journal*, 283(5):882–898.

Zinchuk, V. and Grossenbacher-Zinchuk, O. (2009). Recent advances in quantitative colocalization analysis: Focus on neuroscience. *Progress in Histochemistry and Cytochemistry*, 44(3):125–172.

Zinchuk, V. and Grossenbacher-Zinchuk, O. (2014). Quantitative colocalization analysis of fluorescence microscopy images. *Current Protocols in Cell Biology*, 62(1):4.19.1–4.19.14.

Zinchuk, V. and Zinchuk, O. (2008). Quantitative colocalization analysis of confocal fluorescence microscopy images. *Current Protocols in Cell Biology*, Chapter 4:Unit 4.19.

Curriculum Vitae

Carla Tameling

Kreuzbergring 35
37075 Göttingen
Germany
✉carla.tameling@mathematik.uni-goettingen.de

▬▬▬ Education

Since 2015   **PhD Student**, *Georg-August-Universität Göttingen*, (Supervisor: Prof. Dr. A. Munk, Prof. Dr. A. Sturm).

2015   **M.Sc. in Mathematics**, *Westfälische Wilhelms-Universität Münster*, Thesis title: *Vector-Valued Multi-Bang control for Linearized Elasticity*, (Supervisor: Prof. Dr. Benedikt Wirth).

2013   **B.Sc. in Mathematics**, *Westfälische Wilhelms-Universität Münster*, Thesis title: *Turing-Instabilitäten in Reaktions-Diffusions-Systemen für die Morphogenese und numerische Lösungen mit Splitting Verfahren*, (Supervisor: Prof. Dr. Martin Burger).

2010   **Abitur**, *Arnold-Janssen-Gymnasium*, Neuenkirchen.

▬▬▬ Publications and Preprints

[1] **Tameling, C.**, Stoldt, S., Stephan, T., Jakobs, S. and Munk, A. (2018). Statistical Quantification of Colocalization for Coordinate-Targeted Super-Resolution Microscopy via Optimal Transport. *preprint*

[2] Klatt, M., **Tameling, C.** and Munk, A. (2018). Empirical Regularized Optimal Transport: Statistical Theory and Applications. *ArXiv181009880*

[3] **Tameling, C.** and Munk, A. (2018). Computational Strategies for Inference Based on Empirical Optimal Transport. *IEEE SigPort.*

[4] **Tameling, C.**, Sommerfeld, M. and Munk, A. (2017). Empirical Optimal Transport on Countable Metric Spaces: Distributional Limits and Statistical Applications. *ArXiv170700973*

[5] Clason, C., **Tameling, C.** and Wirth, B. (2016). Vector-Valued Multibang Control of Differential Equations. *SIAM J Control Optim.*

Conference Talks

2018 **Joint Statistical Meeting**, *Vancouver.*

2018 **IEEE Data Science Workshop**, *Lausanne.*

2017 **31st European Meeting of Statisticians**, *Helsinki.*

2016 **Latin American Congress of Probability and Mathematical Statistics**, *San José.*