The Annals of Applied Probability
2019, Vol. 29, No. 5, 2744–2781
https://doi.org/10.1214/19-AAP1463

EMPIRICAL OPTIMAL TRANSPORT ON COUNTABLE METRIC SPACES: DISTRIBUTIONAL LIMITS AND STATISTICAL APPLICATIONS[1]

BY CARLA TAMELING, MAX SOMMERFELD AND AXEL MUNK

University of Goettingen

We derive distributional limits for empirical transport distances between probability measures supported on countable sets. Our approach is based on sensitivity analysis of optimal values of infinite dimensional mathematical programs and a delta method for nonlinear derivatives. A careful calibration of the norm on the space of probability measures is needed in order to combine differentiability and weak convergence of the underlying empirical process. Based on this, we provide a sufficient and necessary condition for the underlying distribution on the countable metric space for such a distributional limit to hold. We give an explicit form of the limiting distribution for tree spaces.

Finally, we apply our findings to optimal transport based inference in large scale problems. An application to nanoscale microscopy is given.

1. Introduction. Optimal transport based distances between probability measures (see, e.g., Rachev and Rüschendorf (1998) or Villani (2009) for a comprehensive treatment), for example, the Wasserstein distance (Vasershtein (1969)), which is also known as Earth Movers distance (Rubner, Tomasi and Guibas (2000)), Kantorovich–Rubinstein distance (Kantorovič and Rubinšteĭn (1958)) or Mallows distance (Mallows (1972)), are of fundamental interest in probability and statistics, with respect to both theory and practice. The pth Wasserstein distance (WD) between two probability measures μ and ν on a Polish metric space $(\mathcal{X}, d)$ is given by

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

for $p \in [1, \infty)$, the infimum is taken over all probability measures π on the product space $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν.

The WD metrizes weak convergence of a sequence of probability measures on $(\mathcal{X}, d)$ together with convergence of its first p moments and has become a

Received September 2018; revised January 2019.

[1]Supported by the DFG Research Training Group 2088 Project A1, CRC 755 Project A6 and Cluster of Excellence MBExC.

MSC2010 subject classifications. Primary 60F05, 60B12, 62E20; secondary 90C08, 90C31, 62G10.

Key words and phrases. Optimal transport, Wasserstein distance, empirical process, limit law, statistical testing.

standard tool in probability, for example, to study limit laws (e.g., Johnson and Samworth (2005), Rachev and Rüschendorf (1994), Shorack and Wellner (1986)), to derive bounds for Monte Carlo computation schemes such as MCMC (e.g., Eberle (2014), Rudolf and Schweizer (2018)), for point process approximations (Barbour and Brown (1992), Schuhmacher (2009)), bootstrap convergence (Bickel and Freedman (1981)) or to quantify measures of risk (Rachev, Stoyanov and Fabozzi (2011)). Besides of its theoretical importance, the WD is used in many applications as an empirical measure to compare complex objects, for example, in image retrieval (Rubner, Tomasi and Guibas (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, Cuesta-Albertos and Matrán (2000)) and machine learning (Rolet, Cuturi and Peyré (2016)).

In such applications, the WD has to be estimated from a finite sample of the underlying measures. This raises the question how fast the *empirical* Wasserstein distance (EWD), that is, when either μ or ν (or both) are estimated by the empirical measures $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ (and $\hat{\nu}_m = \frac{1}{m}\sum_{i=1}^m \delta_{Y_i}$) approaches WD. Ajtai, Komlós and Tusnády (1984) investigated the rate of convergence of EWD for the uniform measure on the unit square, Talagrand (1992, 1994) extended this to higher dimensions. Horowitz and Karandikar (1994) then provided nonasymptotic bounds for the average speed of convergence for the empirical 2-Wasserstein distance. There are several refinements of these results, for example, Boissard and Le Gouic (2014), Fournier and Guillin (2015) and Weed and Bach (2017).

As a natural extension of such results, there is a long standing interest in distributional limits for EWD, 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 WD on the real line when $\mu \neq \nu$ whereas del Barrio, Giné and Matrán (1999), del Barrio et al. (1999) consider the empirical Wasserstein distance when μ belongs to a parametric family of distributions for the assessment of goodness-of-fit, for example, for a Gaussian location scale family. In a similar spirit, del Barrio, Giné and Utzet (2005) provided asymptotics for a weighted version of the empirical 2-Wasserstein distance in one dimension and Freitag and Munk (2005) derive limit laws for semiparametric models, still restricted to the univariate case. There are also several results for dependent data in one dimension, for example, Dede (2009), Dedecker and Merlevède (2017). For a recent survey, we refer to Bobkov and Ledoux (2014) 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) is known explicitly and can be expressed in terms of the quantile functions F^{-1} and G^{-1} of μ and ν, 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 can be used to derive limit

laws (Rippl, Munk and Sturm (2016)). Already for $D = 2$, Ajtai, Komlós and Tusnády (1984) indicate that the scaling rate for the limiting distribution of $W_1(\hat{\mu}_n, \mu)$ when μ is the uniform measure on $\mathcal{X} = [0, 1]^2$ (if it 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 EWD in general dimension with a scaling rate $\sqrt{n}$. This yields a (nondegenerate) normal limit in the case $\mu \neq \nu$, that is, 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 EWD with an expected EWD (whose value is typically unknown) instead of the true WD and requires μ and ν 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, see Villani (2003) for a definition and further results) and the Efron–Stein variance inequality. However, in the case $\mu = \nu$, 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 $\mathcal{X} = \{x_1, \dots, x_N\}$. They derive limit laws for the EWD for $\mu = \nu$ (and $\mu \neq \nu$), which requires a different scaling rate. In this paper, we extend their work to measures $\boldsymbol{r} = (r_x)_{x \in \mathcal{X}}$ that are supported on countable metric spaces $(\mathcal{X}, d)$. Our approach links the asymptotic distribution of the EWD on the one hand to the issue of weak convergence of the underlying multinomial process associated with $\hat{\mu}_n$ with respect to a weighted ℓ^1-norm (for fixed, but arbitrary $x_0 \in \mathcal{X}$)

$$\|\boldsymbol{r}\|_{\ell^1_{d^p_{x_0}}} = \sum_{x \in \mathcal{X}} d^p(x, x_0)|r_x| + |r_{x_0}|, \tag{2}$$

and on the other hand to infinite dimensional sensitivity analysis of the underlying linear program. Notably, we obtain a necessary and sufficient condition for such a limit law, which sheds some light on the limitation to approximate the WD between continuous measures for $D \geq 2$ by discrete random variables.

The outline of this paper is a follows. In Section 2, we give distributional limits for the EWD of measures that are supported on a countable metric space. In short, this limit can be characterized as the optimal value of an infinite dimensional linear program applied to a Gaussian process over the set of dual solutions. The main ingredients of the proof are the directional Hadamard differentiability of the Wasserstein distance on countable metric spaces and the delta method for nonlinear derivatives. We want to emphasize that the delta method for nonlinear derivatives is not a standard tool (see Römisch (2004), Shapiro (1991)). Moreover, for the delta method to work here weak convergence in the weighted ℓ^1-norm (2) of the underlying empirical process $\sqrt{n}(\hat{\mathbf{r}}_n - \boldsymbol{r})$ is required as the directional Hadamard differentiability is proven w.r.t. this norm. We cannot prove the directional Hadamard differentiability with our methods w.r.t. the ℓ^1-norm as the space

of probability measures with finite pth moment is not complete with respect to the ℓ^1-norm; see Section 2.5 for more details. We find that

$$\sum_{x\in\mathcal{X}} d^p(x, x_0)\sqrt{r_x} < \infty \tag{3}$$

is necessary and sufficient for weak convergence. This condition arises from Jain's CLT (Jain (1977)). Furthermore, we examine (3) in a more detailed way in Section 2.3. We give examples and counterexamples for (3), derive consequences for the moments of $\boldsymbol{r}$ from this condition and discuss whether the condition holds in case of an approximation of continuous measures. Further, we examine under which assumptions it follows that (3) holds for all $p' \leq p$ if it is fulfilled for p, and put it in relation to its one-dimensional counterpart; see del Barrio, Giné and Matrán (1999). We close this section by discussing simplifications for ground spaces $\mathcal{X}$ with bounded diameter.

In Section 3, we specify the case where the metric structure on the ground space is given by a rooted tree with weighted edges. In this case, we can calculate the optimal solution of the maximum defining the limit distribution explicitly. Furthermore, we use this explicit formula to derive a distributional upper bound for the limit distribution on general metric spaces via a spanning tree approximation of this general metric space.

In Section 4, we combine this with a well known lower bound (Pele and Werman (2009)) to derive a computationally efficient strategy to test for the equality of two measures $\boldsymbol{r}$ and $\boldsymbol{s}$ on a countable metric space. Furthermore, we derive an explicit formula of the upper bound from Section 3 in the case of the support of $\boldsymbol{r}$ being a regular grid.

An application of our results to data from single marker switching microscopy imaging is given in Section 5. As the number of pixels typically is of magnitude 10^5–10^6, this challenges the assumptions of a finite space underlying the limit law in Sommerfeld and Munk (2018) and our work provides the theoretical justification to perform EWD based inference in such a case. Finally, we stress that our results can be extended to many other situations, for example, the comparison of k samples and when the underlying data are dependent, as soon as a weak limit of the underlying empirical process w.r.t. the weighted ℓ^1-norm (2) can be shown.

2. Distributional limits.

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 $\boldsymbol{r}$ in

$$\mathcal{P}(\mathcal{X}) = \left\{\boldsymbol{r} = (r_x)_{x\in\mathcal{X}} : r_x \geq 0\ \forall x \in \mathcal{X} \text{ and } \sum_{x\in\mathcal{X}} r_x = 1\right\}.$$

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 $\boldsymbol{r} \in \mathcal{P}(\mathcal{X})$ is the union of points $x \in \mathcal{X}$ such that $r_x > 0$. The pth power of the pth Wasserstein distance ($p \geq 1$) then becomes

$$W_p^p(\boldsymbol{r}, \boldsymbol{s}) = \min_{\boldsymbol{w} \in \Pi(\boldsymbol{r}, \boldsymbol{s})} \sum_{x, x' \in \mathcal{X}} d^p(x, x') w_{x,x'}, \tag{4}$$

where

$$\Pi(\boldsymbol{r}, \boldsymbol{s}) = \Big\{ \boldsymbol{w} \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \sum_{x' \in \mathcal{X}} w_{x,x'} = r_x \text{ and } \sum_{x \in \mathcal{X}} w_{x,x'} = s_{x'} \ \forall x, x' \in \mathcal{X} \Big\}$$

is the set of all couplings between $\boldsymbol{r}$ and $\boldsymbol{s}$. Furthermore, let

$$\mathcal{P}_p(\mathcal{X}) = \Big\{ \boldsymbol{r} \in \mathcal{P}(\mathcal{X}) : \sum_{x \in \mathcal{X}} d^p(x, x_0) r_x < \infty \Big\}$$

be the set of probability measures on the countable metric space $\mathcal{X}$ with finite pth moment w.r.t. d. Here, $x_0 \in \mathcal{X}$ is arbitrary and we want to mention that the space is independent of the choice of x_0. 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 (2) as in this case the set of probability measures with finite pth moment is a closed subset, and hence complete itself. This will play a crucial role in the proof of the directional Hadamard differentiability (see Appendix A.1). The weighted ℓ^1-norm (2) can be extended in the following way to sequences on $\mathcal{X} \times \mathcal{X}$ and hence to $\mathcal{P}_p(\mathcal{X} \times \mathcal{X})$:

$$\begin{aligned} \|\boldsymbol{w}\|_{\ell^1_{d^p_{x_0}}} = &\sum_{x, x' \in \mathcal{X}} d^p(x_0, x) |w_{x,x'}| + |w_{x_0,x'}| \\ &+ \sum_{x, x' \in \mathcal{X}} d^p(x_0, x') |w_{x,x'}| + |w_{x,x_0}|. \end{aligned}$$

In contrast to $\mathcal{P}_p(\mathcal{X})$, the space $\ell^1_{d^p_{x_0}}(\mathcal{X})$ depends on $x_0 \in \mathcal{X}$.

2.2. *Main results.* Before we can state the main results, we need a few definitions. Define the empirical measure generated by i.i.d. random variables $X_1, \ldots, X_n$ from the measure $\boldsymbol{r}$ as

$$\hat{\mathbf{r}}_n = (\hat{r}_{n,x})_{x \in \mathcal{X}} \quad \text{where } \hat{r}_{n,x} = \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{X_k = x\}}, \tag{5}$$

and $\hat{\mathbf{s}}_m$ is defined in the same way by $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} \boldsymbol{s}$. In the following, we will denote weak convergence by $\overset{\mathscr{D}}{\to}$, and furthermore, let

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

and

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

Finally, we also require a weighted version of the ℓ^∞-norm to characterize the set of dual solutions:

$$\|a\|_{\ell^\infty_{d^{-p}_{x_0}}} = \max\Big(|a_{x_0}|, \sup_{x\neq x_0\in\mathcal{X}} \big|d^{-p}(x, x_0)a_x\big|\Big),$$

for $p \geq 1$. The space $\ell^\infty_{d^{-p}_{x_0}}(\mathcal{X})$ contains all elements which have a finite $\|\cdot\|_{\ell^\infty_{d^{-p}_{x_0}}}$-norm. In the following, $\langle \boldsymbol{r}, \boldsymbol{\lambda}\rangle = \sum_{x\in\mathcal{X}} r_x\lambda_x$ denotes the dual paring between an element $\boldsymbol{r} \in \ell^1_{d^p_{x_0}}(\mathcal{X})$ and an element $\boldsymbol{\lambda} \in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X})$. Note that all continuous linear functionals on $\ell^1_{d^p_{x_0}}(\mathcal{X})$ can be represented by elements in $\ell^\infty_{d^{-p}_{x_0}}(\mathcal{X})$.

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

$$\begin{aligned}\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s}) = \big\{&(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) \times \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) : \langle \boldsymbol{r}, \boldsymbol{\lambda}\rangle + \langle \boldsymbol{s}, \boldsymbol{\mu}\rangle = W_p^p(\boldsymbol{r}, \boldsymbol{s})\\ &\lambda_x + \mu_{x'} \leq d^p\big(x, x'\big)\ \forall x, x' \in \mathcal{X}\big\}\end{aligned} \tag{6}$$

and

$$\mathcal{S}^*(\boldsymbol{r}) = \big\{\boldsymbol{\lambda} \in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) : \lambda_x - \lambda_{x'} \leq d^p\big(x, x'\big)\ \forall x, x' \in \operatorname{supp}(\boldsymbol{r})\big\}, \tag{7}$$

with $\operatorname{supp}(\boldsymbol{r}) = \{x \in \mathcal{X} : r_x > 0\}$. The set $\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s})$ is the set of dual solutions to the (infinite dimensional) minimization problem defined in (4). We refer the reader to Bonnans and Shapiro (2000) for the general concept of duality in mathematical programming. The general definition of the set of dual solution is given via an arg max formulation. That there is instead of the arg max an equality in our definition of the set of dual solutions is due to the Kantorovich duality (duality theory for the optimal transport problem) (Villani (2009)). For a further discussion of $\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s})$ and $\mathcal{S}^*(\boldsymbol{r})$, we refer the reader to Appendix A.2. For our limiting distributions, we define the following (multinomial) covariance structure

$$\Sigma(\boldsymbol{r}) = \begin{cases} r_x(1 - r_x) & \text{if } x = x',\\ -r_x r_{x'} & \text{if } x \neq x'.\end{cases} \tag{8}$$

THEOREM 2.1. *Let $(\mathcal{X}, d)$ be a countable metric space and $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_p(\mathcal{X})$, $p \geq 1$, and $\hat{\mathbf{r}}_n$ be generated by i.i.d. samples $X_1, \ldots, X_n \sim \boldsymbol{r}$. Furthermore, let $\mathbf{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ be a Gaussian process with $\Sigma(\boldsymbol{r})$ as defined in (8). Assume (3) for some $x_0 \in \mathcal{X}$. Then*

(a)

$$n^{\frac{1}{2}} W_p^p(\hat{\mathbf{r}}_n, \boldsymbol{r}) \xrightarrow{\mathscr{D}} \max_{\boldsymbol{\lambda}\in\mathcal{S}^*(\boldsymbol{r})} \langle \mathbf{G}, \boldsymbol{\lambda}\rangle \quad \text{as } n \to \infty. \tag{9}$$

(b) *In the case where* $\boldsymbol{r} \neq \boldsymbol{s}$, *it holds for* $n \to \infty$

$$n^{\frac{1}{2}}\big(W_p^p(\hat{\mathbf{r}}_n, \boldsymbol{s}) - W_p^p(\boldsymbol{r}, \boldsymbol{s})\big) \xrightarrow{\mathscr{D}} \max_{(\boldsymbol{\lambda},\boldsymbol{\mu})\in\mathcal{S}^*(\boldsymbol{r},\boldsymbol{s})} \langle \mathbf{G}, \boldsymbol{\lambda}\rangle. \tag{10}$$

Before, we give the proof, we will discuss the random variable on the right-hand side of (9) (and (10)) in more detail. In particular, we will investigate the situation in which the limit law in the case of equal marginals (9) will degenerate, that is, when the variance of

$$\max_{\boldsymbol{\lambda}\in\mathcal{S}^*(\boldsymbol{r})} \langle \mathbf{G}, \boldsymbol{\lambda}\rangle \tag{11}$$

is zero. Note that, $\langle \mathbf{G}, \boldsymbol{\lambda}\rangle = \sum_{x\in\mathcal{X}} G_x\lambda_x$ is a centered Gaussian random variable with variance

$$\sum_{x\in\mathcal{X}} \lambda_x^2 r_x - \Big(\sum_{x\in\mathcal{X}} \lambda_x r_x\Big)^2.$$

For $\boldsymbol{\lambda} \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X})$, it holds $|\lambda_x| \leq K d^p(x, x_0)$ for all $x \neq x_0 \in \mathcal{X}$ and some constant K. Hence, the variance of $\langle \mathbf{G}, \boldsymbol{\lambda}\rangle$ can be bounded by $K^2 \sum_{x\in\mathcal{X}} d^{2p} r_x$ which is finite if condition (3) holds (see Lemma 2.8).

Finally, it is worth to note that the stochastic process entering the max in (9)

$$\{\mathcal{G}_{\boldsymbol{\lambda}} := \langle \mathbf{G}, \boldsymbol{\lambda}\rangle, \boldsymbol{\lambda} \in \ell^\infty_{d_{x_0}^{-p}}(\mathcal{X})\}$$

with $\mathbf{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ and $\Sigma(\boldsymbol{r})$ given in (8) is a centered Gaussian process with covariance function

$$K(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{x\in\mathcal{X}} \lambda_x\mu_x r_x - \Big(\sum_{x\in\mathcal{X}} \lambda_x r_x\Big)\Big(\sum_{x\in\mathcal{X}} \mu_x r_x\Big).$$

REMARK 2.2 (The case $\boldsymbol{r} = \boldsymbol{s}$). (a) In general, if the space $\mathcal{X}$ contains isolated points the limit will be nondegenerate. In the case that $\boldsymbol{r}$ has full support, the limit law in (9) degenerates to a point mass at 0 if $\mathcal{S}^*(\boldsymbol{r})$ contains only constant elements, that is, for a $c \in \mathbb{R}$ $\lambda_x = c$ for all $x \in \mathcal{X}$. Then the right-hand side in (9) becomes zero. $\mathcal{S}^*(\boldsymbol{r})$ contains only constant elements if and only if the space $\mathcal{X}$ has no isolated point.

Specifying $\mathcal{X}$ to be a subset of the real line $\mathbb{R}$ that has no isolated point it follows from Theorem 7.11 in Bobkov and Ledoux (2014) that scaling the EWD with $\sqrt{n}$ provides then a nondegenerate limit law. On the other hand, as soon as $\mathcal{X} \subset \mathbb{R}$ contains an isolated point our rate $n^{1/2p}$ (see Remark 2.7) coincides with the rate given in Bobkov and Ledoux (2014).

(b) In the case of $\boldsymbol{r} = \boldsymbol{s}$, the set $\mathcal{S}^*(\boldsymbol{r})$ will always contain more than one element contrary to the case $\boldsymbol{r} \neq \boldsymbol{s}$, and hence, the limit cannot be Gaussian.

(c) In this case, the limiting distribution can also be written as

$$\max_{\boldsymbol{\lambda}\in\mathcal{S}^*(\boldsymbol{r})} \langle \mathbf{G}, \boldsymbol{\lambda}\rangle = \inf_{\boldsymbol{z}(\boldsymbol{r})\in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X})} W_p^p\big(\mathbf{G}^+ + \boldsymbol{z}(\boldsymbol{r}), \mathbf{G}^- + \boldsymbol{z}(\boldsymbol{r})\big),$$

where $\mathbf{G}^+$ and $\mathbf{G}^-$ denotes the (pathwise) decomposition of the Gaussian process $\mathbf{G}$, such that $\mathbf{G} = \mathbf{G}^+ - \mathbf{G}^-$ and $\boldsymbol{z}(\boldsymbol{r})$ is related to $\boldsymbol{r}$ in the sense that $z_x = 0$ for that $x \in \mathcal{X}$ such that $r_x = 0$. Further, we would like to emphasize that the set of dual solutions $\mathcal{S}^*(\boldsymbol{r})$ is independent of $\boldsymbol{r}$, if the support of $\boldsymbol{r}$ is full, that is,

$$\mathcal{S}^* = \big\{\boldsymbol{\lambda} \in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) : \lambda_x - \lambda_{x'} \le d^p\big(x, x'\big)\ \forall x, x' \in \mathcal{X}\big\}. \tag{12}$$

This offers a universal strategy to simulate the limiting distribution for fully supported measures on trees independent of $\boldsymbol{r}$. For more details, see Appendix A.2.

REMARK 2.3 (Some comments for $\boldsymbol{r} \neq \boldsymbol{s}$). (a) Note, that in Theorem 2.1(b) where the measures are not the same the objective function in (10) is independent of the second component $\boldsymbol{\mu}$ of the feasible set $\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s})$. This is due to the fact that in $W_p(\hat{\mathbf{r}}_n, \boldsymbol{s})$ the second component is not random.

(b) Observe that the limit in (10) is normally distributed if the set $\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s})$ is a singleton up to a constant shift. In the case of finite $\mathcal{X}$, conditions for $\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s})$ to be a singleton up to a constant shift are known (Hung, Rom and Waren (1986), Klee and Witzgall (1968)).

(c) Parallel to our work, del Barrio and Loubes (2017) showed asymptotic normality of the quadratic EWD in general dimensions for the case $\boldsymbol{r} \neq \boldsymbol{s}$. Their result requires the measures to have moments of order $4 + \delta$ for some $\delta > 0$ and positive density on their convex support. Their proof relies on the uniqueness and stability of the optimal transportation potential and the Efron–Stein variance inequality. In the case $\boldsymbol{r} = \boldsymbol{s}$, the limiting distribution is degenerated, in contrast to Theorem 2.1(a).

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

THEOREM 2.4. *Under the same assumptions as in Theorem 2.1 and with* $\hat{\mathbf{s}}_m$ *generated by* $Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} \boldsymbol{s}$, *independently of* $X_1, \dots, X_n$ *and* $\mathbf{H} \sim \mathcal{N}(0, \Sigma(\boldsymbol{s}))$, *which is independent of* $\mathbf{G}$, *and the extra assumption that* $\boldsymbol{s}$ *also fulfills* (3) *the following holds.*

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

$$\rho_{n,m} W_p^p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \xrightarrow{\mathscr{D}} \max_{\boldsymbol{\lambda}\in\mathcal{S}^*(\boldsymbol{r})} \langle \mathbf{G}, \boldsymbol{\lambda}\rangle. \tag{13}$$

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

$$\begin{aligned} &\rho_{n,m}\big(W_p^p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) - W_p^p(\boldsymbol{r}, \boldsymbol{s})\big) \\ &\quad \xrightarrow{\mathscr{D}} \max_{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s})} \sqrt{\alpha}\langle \mathbf{G}, \boldsymbol{\lambda}\rangle + \sqrt{1-\alpha}\langle \mathbf{H}, \boldsymbol{\mu}\rangle. \end{aligned} \tag{14}$$

REMARK 2.5. In the case of dependent data analogous results to Theorems 2.1 and 2.4 will hold, as soon as the weak convergence of the underlying empirical process $\sqrt{n}(\hat{\boldsymbol{r}}_n - \boldsymbol{r})$ w.r.t. the $\|\cdot\|_{\ell^1_{d^p_{x_0}}}$-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.

PROOFS OF THEOREM 2.1 AND THEOREM 2.4. To prove these two theorems, we use the delta method Theorem A.2. Therefore, we need to verify (1) directional Hadamard differentiability of $W_p^p(\cdot, \cdot)$ and (2) weak convergence of $\sqrt{n}(\hat{\mathbf{r}}_n - \boldsymbol{r})$. 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. In Appendix A.1, Theorem A.3 directional Hadamard differentiability of W_p^p is shown with respect to the $\|\cdot\|_{\ell^1_{d^p_{x_0}}}$-norm (2).

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

LEMMA 2.6. *Let* $X_1, \ldots, X_n \sim \boldsymbol{r}$ *be i.i.d. taking values in a countable metric space* $(\mathcal{X}, d)$ *and let* $\hat{\mathbf{r}}_n$ *be the empirical measure as defined in* (5). *Then*

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

with respect to the $\|\cdot\|_{\ell^1_{d^p_{x_0}}}$*-norm, where* $\mathbf{G}$ *is a Gaussian process with mean* 0 *and covariance structure*

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

as given in (8) *if and only if condition* (3) *is fulfilled.*

PROOF OF LEMMA 2.6. The weighted ℓ^1-space $\ell^1_{d^p_{x_0}}$ is according to Proposition 3, Maurey (1973) of cotype 2, hence $\sqrt{n}(\hat{\mathbf{r}}_n - \boldsymbol{r})$ converges weakly w.r.t. the $\ell^1_{d^p_{x_0}}$-norm by Corollary 1 in Jain (1977) if and only if the summability condition (3) is fulfilled. □

The proof of Theorem 2.4 works analogously. Note that under the assumptions of the theorem it holds ($\boldsymbol{r} = \boldsymbol{s}$):

$$\begin{aligned}
&\rho_{n,m}\big((\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) - (\boldsymbol{r}, \boldsymbol{s})\big) \\
&\quad = \left(\sqrt{\frac{m}{n+m}}\sqrt{n}(\hat{\mathbf{r}}_n - \boldsymbol{r}), \sqrt{\frac{n}{n+m}}\sqrt{m}(\hat{\mathbf{s}}_m - \boldsymbol{s})\right) \\
&\quad \xrightarrow{\mathscr{D}} (\sqrt{\alpha}\mathbf{G}, \sqrt{1-\alpha}\mathbf{G}')
\end{aligned} \tag{15}$$

with $\mathbf{G}' \overset{\mathcal{D}}{=} \mathbf{G}$. For further explanations, see Appendix A.2. □

REMARK 2.7. From Theorems 2.1 and 2.4, analogous results can be derived for the Wasserstein distance itself. For part (a) of both theorems, one needs to use the continuous mapping theorem for $f(x) = x^{1/p}$.

For part (b), one needs to apply the chain rule for directional Hadamard differentiability (Proposition 3.6(i), Shapiro (1990)).

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

2.3. *Examination of the summability condition* (3). According to Lemma 2.6, condition (3) is necessary and sufficient for the weak convergence with respect to the $\|\cdot\|_{\ell^1_{d^p_{x_0}}}$-norm defined in (2). As this condition is crucial for our main theorem and we are not aware of a comprehensive discussion, we will provide such in this section.

LEMMA 2.8. *If condition* (3) *holds for* $\boldsymbol{r}$, *then* $\boldsymbol{r}$ *has finite moments of order* $2p$.

PROOF. Condition (3) implies that $d^p(x, x_0)\sqrt{r_x} \leq 1$ for all $x \in \mathcal{X}$ besides at most a finite collection of points and denote this subset by $\mathcal{X}'$. Then

$$\sum_{x\in\mathcal{X}} d^{2p}(x, x_0) r_x \leq \sum_{x\in\mathcal{X}'} d^p(x, x_0)\sqrt{r_x} + \sum_{x\in\mathcal{X}\setminus\mathcal{X}'} d^{2p}(x, x_0) r_x < \infty.$$

□

Furthermore, the following question arises. "If the condition holds for p does it then also hold for all $p' \leq p$?" This is not true in general, but it is true if x_0 is not an accumulation point.

LEMMA 2.9. *Let* $x_0 \in \mathcal{X}$ *be an isolated point with respect to the metric* d. *If condition* (3) *holds for* p, *then it also holds for all* $1 \leq p' \leq p$.

PROOF. Let $x_0 \in \mathcal{X}$ be an isolated point, that is, there exists $\varepsilon > 0$ such that $d(x, x_0) > \varepsilon$ for all $x \neq x_0 \in \mathcal{X}$. Then

$$\sum_{x\in\mathcal{X}} d^p(x_0,x)\sqrt{r_x} = \varepsilon^p \sum_{x\in\mathcal{X}} \left(\frac{d(x_0,x)}{\varepsilon}\right)^p \sqrt{r_x}$$
$$\geq \varepsilon^{p/p'} \sum_{x\in\mathcal{X}} d^{p'}(x_0,x)\sqrt{r_x}.$$

□

Exponential families. As we will see, condition (3) is fulfilled for many well-known distributions including the Poisson distribution, geometric distribution or negative binomial distribution with the Euclidean distance as the ground measure d on $\mathcal{X} = \mathbb{N}$.

THEOREM 2.10. *Let $(\mathcal{P}_\eta)_\eta$ be an s-dimensional standard exponential family (SEF) with natural parameter space $\mathcal{N}$ (see Lehmann and Casella (1998), Section 1.5) of the form*

$$r_x^\eta = h_x \exp\left(\sum_{i=1}^s \eta_i T_x^i - A(\eta)\right). \tag{16}$$

The summability condition (3) is fulfilled if $(\mathcal{P}_\eta)_\eta$ satisfies:

(1) *$h_x \geq 1$ for all $x \in \mathcal{X}$,*

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

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

The proof of this theorem, as well as examples which show the necessity of all three conditions, can be found in Appendix B.

2.4. *Approximation of continuous distributions*. In this section, we investigate to what extent we can approximate continuous measures by its discretization such that condition (3) remains valid. Let $\mathcal{X} = (\frac{k}{M})_{k\in\mathbb{Z}}$ 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

$$r_k := F\left(\frac{k+1}{M}\right) - F\left(\frac{k}{M}\right). \tag{17}$$

Now, (3) can be estimated as follows:

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

where the first inequality is due to Jensen's inequality. As the right-hand side tends to infinity with rate $\sqrt{M}$ as $M \to \infty$, condition (3) 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 $\mathcal{X}$ to be a subset of $\mathbb{R}$ and want to put condition (3) in relation to the condition (del Barrio, Giné and Matrán (1999))

$$\int_{-\infty}^{\infty} \sqrt{F(t)\big(1 - F(t)\big)}\,dt < \infty, \tag{18}$$

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 (1977) in a more general context).

Condition (3) is under certain assumptions stronger than (18) as the following shows. Let $\mathcal{X}$ 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 $\mathcal{X}$. For any measure $\boldsymbol{r}$ with cumulative distribution function F on $\mathcal{X}$, it holds

$$\begin{aligned}
&\int_{-\infty}^{\infty} \sqrt{F(t)\big(1 - F(t)\big)}\,dt \\
&\quad = \sum_{k\in\mathbb{Z}} d(x_k, x_{k+1}) \sqrt{\sum_{j\leq k} r_j} \sqrt{\sum_{j>k} r_j} \\
&\quad \leq \sum_{k=0}^{\infty} d(x_k, x_{k+1}) \sqrt{\sum_{j>k} r_j} + \sum_{k=-\infty}^{-1} d(x_k, x_{k+1}) \sqrt{\sum_{j\leq k} r_j}
\end{aligned}$$

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

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

Hence, if condition (3) holds, (18) is also fulfilled. However, the conditions are not equivalent as the following example shows.

EXAMPLE 2.11. Let $\mathcal{X} = \mathbb{N}$ and $d(x, y) = |x - y|$ the Euclidean distance and $\boldsymbol{r}$ a power-law, that is, $r_n = \frac{1}{\zeta(s)}\frac{1}{n^s}$, where $\zeta(s)$ is the Riemann zeta function. In this case, (18) reads

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

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

and this is finite if and only if $s > 3$. On the other hand, condition (3) 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, condition (18) is fulfilled for $s \in (3, 4]$, but not (3).

For $p = 2$ in dimension $D = 1$, there is no such easy condition anymore in the case of continuous measures; see del Barrio, Giné and Utzet (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 (17) condition (3) is fulfilled and Theorems 2.1 and 2.4 are valid.

2.5. *Bounded diameter of* $\mathcal{X}$. For $\mathcal{X}$ with bounded diameter, further simplifications can be obtained.

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 pth moment to guarantee that the linear program (30) defining the

Wasserstein distance has a finite value. Thus, we can operate on $\mathcal{P}(\mathcal{X})$ which is a subset of $\ell^1(\mathcal{X})$. This simplifies the summability condition (3) to

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

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

3. Limiting distribution for tree metrics.

3.1. *Explicit limits.* In this subsection, we give an explicit expression for the limiting distribution in (9) and (13) in the case $\boldsymbol{r} = \boldsymbol{s}$ with full support (otherwise see Remark 3.3) when the metric is generated by a weighted tree. This extends Theorem 5 in Sommerfeld and Munk (2018) for finite spaces to countable spaces $\mathcal{X}$. In the following, we recall their notation.

Assume that the metric structure on the countable space $\mathcal{X}$ is given by a weighted tree, that is, an undirected connected graph $\mathcal{T} = (\mathcal{X}, E)$ with vertices $\mathcal{X}$ and edges $E \subset \mathcal{X} \times \mathcal{X}$ that contains no cycles. We assume the edges to be weighted by a function

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

Without imposing any further restriction on $\mathcal{T}$, we assume it to be rooted at $\operatorname{root}(\mathcal{T}) \in \mathcal{X}$, say. Then, for $x \in \mathcal{X}$ and $x \neq \operatorname{root}(\mathcal{T})$ we may define $\operatorname{parent}(x) \in \mathcal{X}$ as the immediate neighbor of x in the unique path connecting x and $\operatorname{root}(\mathcal{T})$. We set $\operatorname{parent}(\operatorname{root}(\mathcal{T})) = \operatorname{root}(\mathcal{T})$. We also define $\operatorname{children}(x)$ as the set of vertices $x' \in \mathcal{X}$ such that there exists a sequence $x' = x_1, \dots, x_n = x \in \mathcal{X}$ with $\operatorname{parent}(x_j) = x_{j+1}$ for $j = 1, \dots, n-1$. Note that with this definition $x \in \operatorname{children}(x)$. Furthermore, observe that $\operatorname{children}(x)$ can consist of countably many elements, but the path joining x and $x' \in \operatorname{children}(x)$ is still finite as explained below.

For $x, x' \in \mathcal{X}$, let $e_1, \dots, e_n \in E$ be the unique path in $\mathcal{T}$ joining x and x', then the length of this path,

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

defines a metric $d_{\mathcal{T}}$ on $\mathcal{X}$. This metric is well defined, since the unique path joining x and x' is finite as we show in the following. Let $A_0 = \{x \in \mathcal{X} : x = \operatorname{root}(\mathcal{T})\}$, $A_1 = \{x \in \mathcal{X} : \operatorname{parent}(x) = \operatorname{root}(\mathcal{T})\} \setminus \operatorname{root}(\mathcal{T})$ and $A_k = \{x \in \mathcal{X} : \operatorname{parent}(x) \in A_{k-1}\}$ 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.

Additionally, define

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

and

$$Z_{\mathcal{T},p}(\boldsymbol{u}) = \sum_{x \in \mathcal{X}} \big|(S_{\mathcal{T}}\boldsymbol{u})_x\big| d_{\mathcal{T}}\big(x, \text{parent}(x)\big)^p \tag{19}$$

for $\boldsymbol{u} \in \mathbb{R}^{\mathcal{X}}$ and we set w.l.o.g. $x_0 = \text{root}(\mathcal{T})$.

The main result of this section is the following.

THEOREM 3.1. *Let $\boldsymbol{r} \in \mathcal{P}_p(\mathcal{X})$, defining a probability distribution on $\mathcal{X}$ that fulfils condition (3) and let the empirical measures $\hat{\mathbf{r}}_n$ and $\hat{\mathbf{s}}_m$ be generated by independent random variables $X_1, \dots, X_n$ and $Y_1, \dots Y_m$, respectively, all drawn from $\boldsymbol{r} = \boldsymbol{s}$.*

Then, with a Gaussian vector $\boldsymbol{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ with $\Sigma(\boldsymbol{r})$ as defined in (8) we have the following:

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

$$\sqrt{n} W_p^p(\hat{\mathbf{r}}_n, \boldsymbol{r}) \xrightarrow{\mathscr{D}} Z_{\mathcal{T},p}(\mathbf{G}). \tag{20}$$

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

$$\sqrt{\frac{nm}{n+m}} W_p^p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \xrightarrow{\mathscr{D}} Z_{\mathcal{T},p}(\mathbf{G}). \tag{21}$$

A rigorous proof of Theorem 3.1 is given in Appendix A.3.

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 $\boldsymbol{\mu} \in \ell^1_{d^p_{x_0}}(\mathcal{X})$ a sequence which has only finitely many nonzeros that converges to $\boldsymbol{\mu}$ 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 3.2. Analogous results to Theorem 3.1 for the Wasserstein distance W_p on trees can be derived by the techniques described in Remark 2.7.

REMARK 3.3. 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

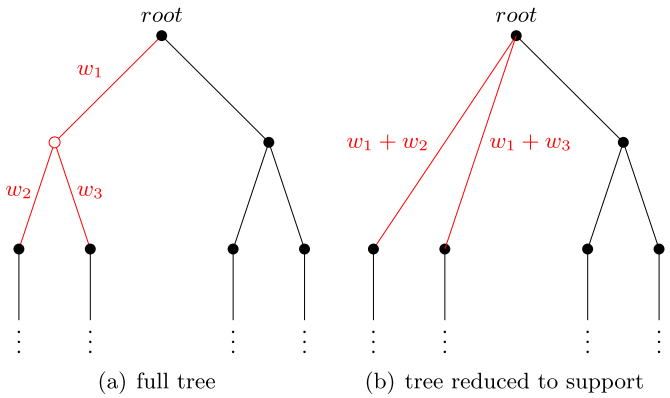


FIG. 1. *Schematic for the reduction of $\mathcal{X}$ to the support of $\boldsymbol{r}$. Solid circles indicate support points, hollow circles elements which are not in the support.*

$\boldsymbol{r}$, we delete x and connect parent(x) to all nodes in the set $A_{+1}(x) = \{x' \in \mathcal{X}\colon \text{parent}(x') = x\}$ 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 nonfull support. See Figure 1 for an illustration.

3.2. *Distributional bound for the limiting distribution.* In this section, we use the explicit formula on the r.h.s. of (20) for the case of tree metrics to stochastically bound the limiting distribution on a general space $\mathcal{X}$ which is not a tree.

This is based on the following simple observation: Let $\mathcal{T}$ be a spanning tree of $\mathcal{X}$ and $d_{\mathcal{T}}$ the tree metric generated by $\mathcal{T}$ and the weights $(x, x') \mapsto d(x, x')$ as described in Section 3.1. Then for any $x, x' \in \mathcal{X}$ we have $d(x, x') \le d_{\mathcal{T}}(x, x')$. Let $\mathcal{S}^*_{\mathcal{T}}$ denote the set defined in (7) with the metric $d_{\mathcal{T}}$ instead of d. Then $\mathcal{S}^* \subset \mathcal{S}^*_{\mathcal{T}}$, and hence

$$\max_{\boldsymbol{\lambda} \in \mathcal{S}^*} \langle \boldsymbol{v}, \boldsymbol{\lambda} \rangle \le \max_{\boldsymbol{\lambda} \in \mathcal{S}^*_{\mathcal{T}}} \langle \boldsymbol{v}, \boldsymbol{\lambda} \rangle$$

for all $\boldsymbol{v} \in \ell^1_{d^p_{x_0}}(\mathcal{X})$. It follows that

$$\max_{\boldsymbol{\lambda} \in \mathcal{S}^*} \langle \boldsymbol{v}, \boldsymbol{\lambda} \rangle \le Z_{\mathcal{T},p}(\boldsymbol{v}) \tag{22}$$

for all $\boldsymbol{v} \in \ell^1_{d^p_{x_0}}(\mathcal{X})$ and this proves the following main result of this subsection, which is stated for the case, when $\boldsymbol{r}$ and $\boldsymbol{s}$ are both estimated from data. The one-sample case is analogous.

THEOREM 3.4. *Let $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_p(\mathcal{X})$, assume that $\boldsymbol{r}, \boldsymbol{s}$ fulfill condition (3) and let $\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m$ be generated by i.i.d. $X_1, \dots, X_n \sim \boldsymbol{r}$ and $Y_1, \dots, Y_m \sim \boldsymbol{s}$, respectively. Let*

further $\mathcal{T}$ be a spanning tree of $\mathcal{X}$. Then, if $\boldsymbol{r} = \boldsymbol{s}$ we have, as n and m approach infinity such that $n \wedge m \to \infty$ and $n/(n+m) \to \alpha$, that

$$\limsup_{n,m\to\infty} P\left[\left(\frac{nm}{n+m}\right)^{1/2p} W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \geq z\right] \leq P\big[Z_{\mathcal{T},p}^{1/p}(\mathbf{G}) \geq z\big], \tag{23}$$

where $\mathbf{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$ with $\Sigma(\boldsymbol{r})$ as defined in (8).

REMARK 3.5. While the stochastic bound of the limiting distribution $Z_{\mathcal{T},p}$ is very fast to compute as it is explicitly given, the Wasserstein distance $W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m)$ in (23) is a computational bottleneck. Classical general-purpose approaches, for example, the simplex algorithm (Luenberger and Ye (2008)) for general linear programs or the auction algorithm for network flow problems (Bertsekas (1992, 2009)) were found to scale rather poorly to very large problems such as image retrieval (Rubner, Tomasi and Guibas (2000)).

Attempts to solve this problem include specialized algorithms (Gottschlich and Schuhmacher (2014)) and approaches leveraging additional geometric structure of the data (Ling and Okada (2007), Schmitzer (2016)). However, many practical problems still fall outside the scope of these methods (Schrieber, Schuhmacher and Gottschlich (2017)), prompting the development of numerous surrogate quantities which mimic properties of optimal transport distances and are amenable to efficient computation. Examples include Bonneel et al. (2015), Pele and Werman (2009), Shirdhonkar and Jacobs (2008) and the particularly successful entropically regularized transport distances (Cuturi (2013), Solomon et al. (2015)).

In the next section, we will discuss how to approximate the countable space $\mathcal{X}$ by a finite collection of points. Note that the distributional bound in Theorem 3.4 also holds on any finite collection of points. For a simulation study regarding this upper bound, see Tameling and Munk (2018).

4. Computational strategies for simulating the limit laws. If we want to simulate the limiting distributions in Theorems 2.1 and 2.4, we need to restrict to a finite number N of points, that is, we choose a subset I of $\mathcal{X}$ such that $\#I = N$. Let $\boldsymbol{r} \in \mathcal{P}_p(\mathcal{X})$ with full support (see Remark 4.1 for the general case), satisfying (3). For $\mathbf{G} \sim \mathcal{N}(0, \Sigma(\boldsymbol{r}))$, we define $\mathbf{G}^I = (G^I)_x = G_x \mathbb{1}_{\{x \in I\}}$. Then an upper bound for the difference between the exact limiting distribution and the limiting distribution on the finite set I in the one sample case for $\boldsymbol{r} = \boldsymbol{s}$ is given as (see (22))

$$\begin{aligned}
\Big|\max_{\boldsymbol{\lambda}\in\mathcal{S}^*}\langle \mathbf{G}^I, \boldsymbol{\lambda}\rangle - \max_{\boldsymbol{\lambda}\in\mathcal{S}^*}\langle \mathbf{G}, \boldsymbol{\lambda}\rangle\Big| &\leq \max_{\boldsymbol{\lambda}\in\mathcal{S}^*}\big|\langle \mathbf{G}^I, \boldsymbol{\lambda}\rangle - \langle \mathbf{G}, \boldsymbol{\lambda}\rangle\big| \\
&\leq \max_{\boldsymbol{\lambda}\in\mathcal{S}^*_{\mathcal{T}}}\big|\langle \mathbf{G}^I - \mathbf{G}, \boldsymbol{\lambda}\rangle\big| \\
&= \max\Big\{\max_{\boldsymbol{\lambda}\in\mathcal{S}^*_{\mathcal{T}}}\langle \mathbf{G}^I - \mathbf{G}, \boldsymbol{\lambda}\rangle, \max_{\boldsymbol{\lambda}\in\mathcal{S}^*_{\mathcal{T}}}\langle \mathbf{G} - \mathbf{G}^I, \boldsymbol{\lambda}\rangle\Big\} \qquad (24)
\end{aligned}$$

$$= \sum_{x\in\mathcal{X}} \left|\left(S_{\mathcal{T}}(\mathbf{G}^I - \mathbf{G})\right)_x\right| d_{\mathcal{T}}\big(x, \operatorname{parent}(x)\big)^p$$
$$= \sum_{x\notin I} |G_x| d_{\mathcal{T}}\big(x, \operatorname{root}(\mathcal{T})\big)^p.$$

For the last equality, one needs to construct the tree as follows: Choose I such that x_0 from condition (3) is an element of I and choose x_0 to be the root of the tree and let all other elements of $\mathcal{X}$ be direct children of the root, that is, $\operatorname{children}(x) = x$ for all $x \neq \operatorname{root}(\mathcal{T}) \in \mathcal{X}$. The upper bound can be made stochastically arbitrarily small as

$$\mathbb{E}\left[\sum_{x\notin I} |G_x| d_{\mathcal{T}}\big(x, \operatorname{root}(\mathcal{T})\big)^p\right] \leq \sum_{x\notin I} d_{\mathcal{T}}\big(x, \operatorname{root}(\mathcal{T})\big)^p \sqrt{r_x(1-r_x)}, \tag{25}$$

where we used Hölder's inequality and the definition of $\Sigma(\boldsymbol{r})$. As the root was chosen to be x_0, the sum above is finite as $\boldsymbol{r}$ fulfills condition (3) and becomes arbitrarily small for I large enough. Hence, (25) details that the speed of approximation by $\mathbf{G}^I$ depends on the decay of $\boldsymbol{r}$ and suggests to choose I such that most of the mass of $\boldsymbol{r}$ is concentrated on it.

REMARK 4.1. In case that the support of $\boldsymbol{r}$ is not full, we have to optimize over the set $\mathcal{S}^*(\boldsymbol{r})$ given in (7). In this case, we can derive the same upper bound as in (24) with the only change that we sum over all $x \in \operatorname{supp}(\boldsymbol{r})$ in the second last line of (24) and that our set I has to be a subset of the support of $\boldsymbol{r}$.

The computation of $\max_{\boldsymbol{\lambda}\in\mathcal{S}^*}\langle \mathbf{G}^I, \boldsymbol{\lambda}\rangle$ is a linear program with N^2 constraints and N variables. General purpose network flow algorithms such as the auction algorithm, Orlin's algorithm or general purpose LP solvers are required for the computation of this linear problem. These algorithms have at least cubic worst case complexity (Bertsekas (1981), Orlin (1993)) and quadratic memory requirement and its average runtime is much worse than $\mathcal{O}(N^2)$ empirically (Gottschlich and Schuhmacher (2014)). This renders a naive Monte Carlo approach to obtain quantiles computational infeasible for large N. In the following subsections, we therefore discuss possibilities to make the computation of the limit more accessible.

4.1. *Thresholded Wasserstein distance.* Following Pele and Werman (2009), we define for a thresholding parameter $t \geq 0$ the thresholded metric

$$d_t(x, x') = \min\{d(x, x'), t\}. \tag{26}$$

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

THEOREM 4.2. *The limiting distribution from Theorem 2.1 with the thresholded ground distance d_t instead of d can be computed in $\mathcal{O}(N^2 \log N)$ time with $\mathcal{O}(N)$ memory requirement, if each point in $\mathcal{X}$ has $\mathcal{O}(1)$ neighbors with distance smaller or equal to t. The limiting distribution can be calculated as the optimal value of the following network flow problem*:

$$\begin{aligned} &\min_{\boldsymbol{w} \in \ell^1_{d_{x_0}^p}(\mathcal{X}\times\mathcal{X})_+} - \sum_{x,x' \in \mathcal{X}} d_t^p(x,x') w_{x,x'} \\ &\text{subject to} \sum_{\tilde{x} \in \mathcal{X}, \tilde{x} \neq x} w_{\tilde{x},x} - \sum_{x' \in \mathcal{X}, x' \neq x} w_{x,x'} = G_x, \end{aligned} \tag{27}$$

where $\mathbf{G} = (G_x)_{x \in \mathcal{X}}$ is a Gaussian process with mean zero and covariance structure as defined in (8).

PROOF. We take a finite approximation $\boldsymbol{r}_N$ of $\boldsymbol{r}$ and reduce our space $\mathcal{X}$ to the support of $\boldsymbol{r}_N$ which should be exactly N points. If we take the thresholded distance as the ground distance similar as in Theorem 2.1, we obtain the limiting distribution as

$$\max_{\boldsymbol{\lambda} \in \mathcal{S}_t^*} \langle \mathbf{G}, \boldsymbol{\lambda} \rangle,$$

where now $\mathcal{S}_t^* = \{\boldsymbol{\lambda} \in \mathbb{R}^N : \lambda_x - \lambda_{x'} \leq d_t^p(x,x')\}$. The limiting distribution is again a finite dimensional linear program and since there is strong duality in this case, it is equivalent to solve (27). As the linear program (27) is a network flow problem, we can redirect all edges with length t through a virtual node without changing the optimal value. From the assumption that each point has $\mathcal{O}(1)$ neighbors with distance not equal to t, we can deduce that the number of edges (N^2 in the original problem) is reduced to $\mathcal{O}(N)$. According to Pele and Werman (2009), the new linear program with the virtual node can be solved in $\mathcal{O}(N^2 \log N)$ time with $\mathcal{O}(N)$ memory requirement. □

REMARK 4.3. (a) The resulting network-flow problem can be tackled with existing efficient solvers (Bertsekas (1992)) or commercial solvers like *CPLEX* (https://www.ibm.com/jm-en/marketplace/ibm-ilog-cplex) which exploit the network structure.

(b) For the distributional bound (23), one can also use the thresholded Wasserstein distance W_p^t instead of W_p to be computational more efficient. A large threshold t will result in a better approximation of the true Wasserstein distance, but will also require more computation time.

4.2. *Regular grids*. In this section, we are going to derive an explicit formula for the distributional bound from Section 3.2, when the support of $\boldsymbol{r}$ is a regular grid of L^D points in the unite hypercube $[0,1]^D$. Here, D is a positive integer and

L a power of two. In this case, a spanning tree can be constructed from a dyadic partition. The general case is analogous, but more cumbersome. For $0 \leq l \leq l_{\max}$ with

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

let P_l be the natural partition of $\operatorname{supp}(\boldsymbol{r})$ into 2^{Dl} squares of each $L^D/2^{Dl}$ points.

THEOREM 4.4. *Under the assumptions described above*, (19) *reads*

$$Z_{\mathcal{T},p}(\boldsymbol{u}) = \sum_{l=0}^{l_{\max}} D^{p/2} 2^{-p(l+1)} \sum_{F \in P_l} |S_F \boldsymbol{u}|. \tag{28}$$

This expression can be evaluated efficiently (*in* $L^D \log_2 L$ *operations*) *and used with Theorem* 3.4 *to obtain a stochastic bound of the limiting distribution on regular grids*.

The proof of this theorem can be found in Appendix A.4.

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

SMS microscopy is based on the principle that at any given time only a very small number of markers are in the on state. As the probability of switching from the off to the on state is small for each individual marker and they remain in the on state only for a very short time (1–100 ms). This allows SMS microscopy to resolve features below the diffraction barrier that limits conventional far-field microscopy (see Hell (2007) for a survey) because with overwhelming probability at most one marker within a diffraction limited spot is in the on state (Aspelmeier, Egner and Munk (2015)). At the same time, this requires quite long acquisition times (1 min–1 h) to guarantee sufficient sampling of the probe. As a consequence, if the probe moves during the acquisition, the final image will be blurred.

Correcting for this drift, and thus improving image quality is an area of active research (Geisler et al. (2012), Deschout et al. (2014), Hartmann et al. (2016)). In order to investigate the validity of such a drift correction method, we introduce a test of the Wasserstein distance between the image obtained from the first half of

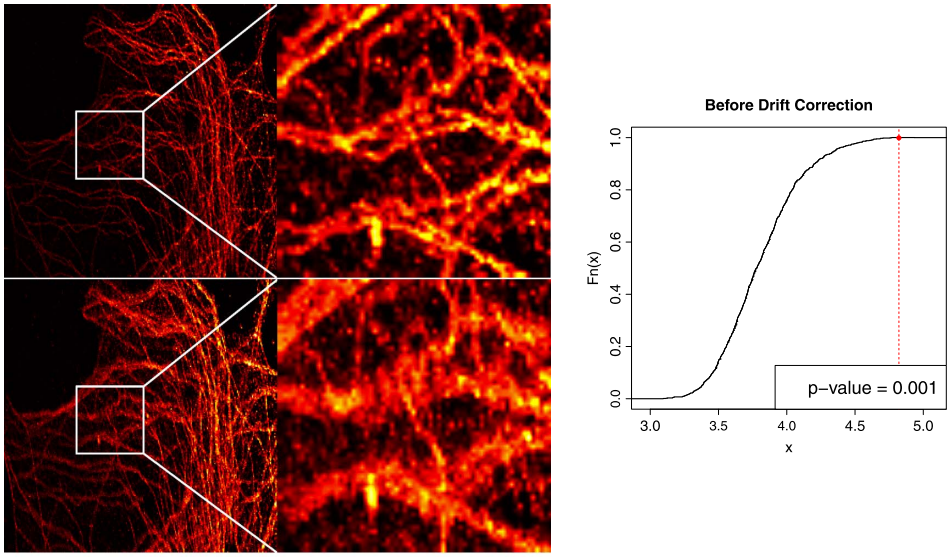
the recording time and the second half. This test is based on the distributional upper bound of the limiting distribution which was developed in Section 3.2 in combination with a lower bound of the Wasserstein distance (Pele and Werman (2009)). In fact, there is no standard method for problems of this kind and we argue that the (thresholded) Wasserstein distance is particular useful in such a situation as the specimen moves between the frames without loss of mass, hence the drift induces a transport structure between successive frames. In the following, we compare the distribution from the first half of frames with the distribution from the second half scaled with the sample sizes (as in (21)). We reject the hypothesis that the distributions from the first and the second half are the same, if our test statistic is larger than the $1-\alpha$ quantile of the distributional bound of the limiting distribution in (23). If we have statistical evidence that the thresholded Wasserstein distance is not zero, we can also conclude that there is a significant difference in the Wasserstein distance itself.

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

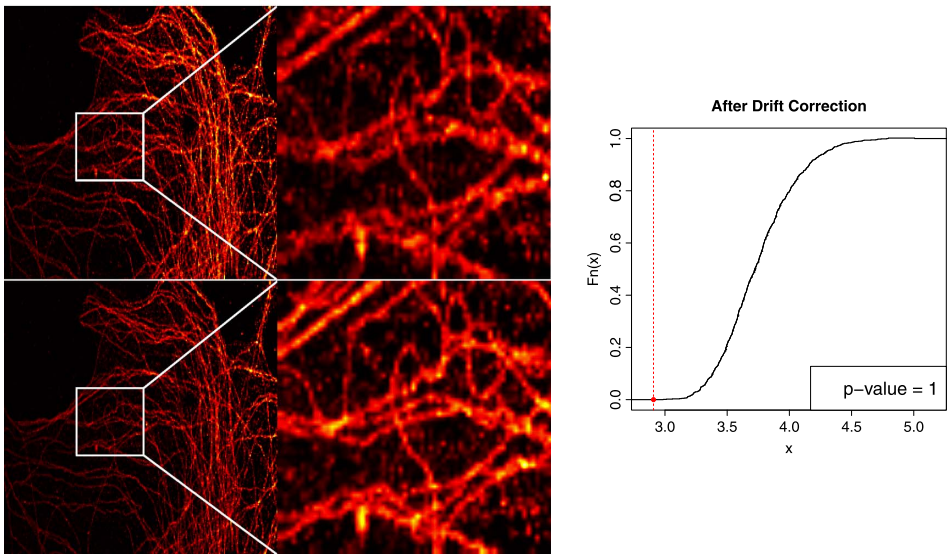
Data and results. We consider an SMS image of a tubulin structure presented in Hartmann et al. (2016) to assess their drift correction method. This image is recorded in 40.000 single frames over a total recording time of 10 minutes (i.e., 15 ms per frame). We compare the aggregated sample collected during the first 50% ($\hat{=}$ 20.000 frames) of the total observation time with the aggregated sample obtained in the last 50% on a 256×256 grid for both the original uncorrected values and for the values where the drift correction of Hartmann et al. (2016) was applied. Heat maps of these four samples are shown in the left-hand side of Figure 2 (no correction in (a) and corrected in (b)).

The question we will address is: "To what extend has the drift been properly removed by the drift correction?" In addition, from the application of the thresholded Wasserstein distance for different thresholds we expect to obtain detailed understanding for which scales the drift has been removed. As Hartmann et al. (2016) have corrected with a global drift function one might expect that on small spatial scales not all effects have been removed.

We compute the thresholded Wasserstein distance W_1^t between the two pairs of samples as described in Section 4.1 with different thresholds $t \in \{2, 3, \dots, 14\}/256$. We compare these values with a sample from the stochastic upper bound for the limiting distribution on regular grids obtained as described in Section 4.2. This allows us to obtain a test for the null hypothesis "no difference" based on Theorem 3.4. To visualize the outcomes of theses tests for different thresholds t, we



a) Before drift correction



b) After drift correction

FIG. 2. (a) *Left*: *Aggregated samples of the first* (*first row*) *and the last* (*second row*) 50% *of the observation time as heat maps of relative frequency without correction for the drift of the probe. Magnifications of a small area are shown to highlight the blurring of the picture. Right*: *Empirical distribution function of a sample from the upper bound* (*tree approximation*) *of the limiting distribution. The red dot* (*line*) *indicates the scaled thresholded Wasserstein distance for* $t = 6/256$. (b) *Same setup as in* (a) *after drift correction. Here, the difference between the fist and the second* 50% *is no longer significant.*

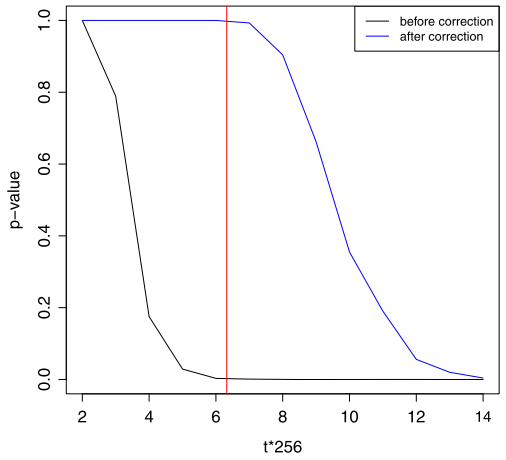


FIG. 3. *P-values for the null hypothesis "no difference" for different thresholds t before and after the drift correction. The red line indicates the magnitude of the total drift.*

have plotted the corresponding p-values in Figure 3. The red line indicates the magnitude of the drift over the total recording time. As the magnitude is approximately $6/256$, we plot in the right-hand side of Figure 2(a) and (b) the empirical distribution functions of the upper bound (23) and indicate the value of the test-statistic for $t = 6/256$ with a red dot without the drift correction and with the correction, respectively.

As shown in Figure 3, the differences caused by the drift of the probe are recognized as highly statistically significant ($p \leq 0.05$) for thresholds larger than $t = 4/256$. After the drift correction method is applied, the difference is no longer significant for thresholds smaller than $t = 14/256$. The estimated shift during the first and the second 50% of the observations is three pixels in x-direction and one pixel in y-direction. That shows that the significant difference that is detected when comparing the images without drift correction for $t \in \{5, 6, 7, 8, 9, 10\}/256$ is caused in fact by the drift. The fact that there is still a significant difference for large thresholds ($t \geq 14$) in the corrected pictures suggests further intrinsic and local inhomogeneous motion of the specimen or nonpolynomial drift that is not captured by the drift model used in Hartmann et al. (2016) and bleaching effects of fluorescent markers.

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

APPENDIX A: PROOFS

A.1. Hadamard directional differentiability. In this section, we follow mainly Shapiro (1991) and Römisch (2004). Let $\mathcal{U}$ and $\mathcal{Y}$ be normed spaces.

DEFINITION A.1 (cf. Römisch (2004), Shapiro (1991)). (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 the limit

$$f'_u(h) = \lim_{n\to\infty} \frac{f(u + t_n h_n) - f(u)}{t_n} \tag{29}$$

exist.

(b) *Hadamard directional differentiability tangentially to a set*. Let K be a subset of D_f, f is directionally differentiable tangentially to K in the sense of Hadamard at u if the limit (29) exists for all sequences h_n that converge to h of the form $h_n = t_n^{-1}(k_n - u)$ where $k_n \in K$ and $t_n \searrow 0$. This derivative is defined on the contingent (Bouligand) cone to K at u

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

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)).

The delta method for mappings that are directionally Hadamard differentiable tangentially to a set reads as follows.

THEOREM A.2 (Römisch (2004), Theorem 1). *Let K be a subset of $\mathcal{U}$, $f\colon K \to \mathcal{Y}$ 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 $f'_u(\cdot)\colon T_K(u) \to \mathcal{Y}$.*

(ii) *For each n, $X_n\colon \Omega_n \to K$ 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 that takes values in $T_K(u)$.*

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

Hadamard directional differentiability of the Wasserstein distance on countable metric spaces. For $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_p(\mathcal{X})$ the pth power of the pth Wasserstein distance is the optimal value of an infinite dimensional linear program. We use this fact to verify that the pth power of the Wasserstein distance (4) on the countable metric spaces $\mathcal{X}$ is directionally Hadamard differentiable with methods of sensitivity analysis of optimal values.

The pth power of the Wasserstein distance on countable metric spaces is the optimal value of the following infinite dimensional linear program:

$$
\begin{aligned}
&\min_{\boldsymbol{w}\in\ell^1_{d^p_{x_0}}(\mathcal{X}\times\mathcal{X})} \sum_{x,x'\in\mathcal{X}} d^p(x,x')w_{x,x'} \\
&\text{subject to } \sum_{x'\in\mathcal{X}} w_{x,x'} = r_x \quad \forall x\in\mathcal{X}, \\
&\qquad\qquad \sum_{x\in\mathcal{X}} w_{x,x'} = s_{x'} \quad \forall x'\in\mathcal{X}, \\
&\qquad\qquad w_{x,x'} \geq 0 \quad \forall x,x'\in\mathcal{X}.
\end{aligned} \tag{30}
$$

THEOREM A.3. *W_p^p as a map from $(\mathcal{P}_p(\mathcal{X})\times\mathcal{P}_p(\mathcal{X}), \|\cdot\|_{\ell^1_{d^p_{x_0}}})$ to $\mathbb{R}$, $(\boldsymbol{r},\boldsymbol{s})\mapsto W_p^p(\boldsymbol{r},\boldsymbol{s})$ is Hadamard directionally differentiable tangentially to $\mathcal{P}_p(\mathcal{X})\times\mathcal{P}_p(\mathcal{X})$. The contingent cone on which the derivative is defined is given by*

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

with

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

and the directional derivative is as follows:

$$(\boldsymbol{d}_1,\boldsymbol{d}_2)\mapsto \sup_{(\boldsymbol{\lambda},\boldsymbol{\mu})\in\mathcal{S}^*(\boldsymbol{r},\boldsymbol{s})} -\big(\langle\boldsymbol{\lambda},\boldsymbol{d}_1\rangle + \langle\boldsymbol{\mu},\boldsymbol{d}_2\rangle\big), \tag{31}$$

where $\mathcal{S}^*(\boldsymbol{r},\boldsymbol{s})$ is set of optimal solutions of the dual problem which is defined in (6).

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 pth power of the pth 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

$$C\big(\boldsymbol{w},(\boldsymbol{r},\boldsymbol{s})\big) = \begin{pmatrix} \boldsymbol{w} \\ \Sigma_1\boldsymbol{w}-\boldsymbol{r} \\ \Sigma_2\boldsymbol{w}-\boldsymbol{s}\end{pmatrix}. \tag{32}$$

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, that is, $\Sigma_1\boldsymbol{w} = \sum_{x'\in\mathcal{X}} w_{x,x'}$ and $\Sigma_2\boldsymbol{w} = \sum_{x\in\mathcal{X}} w_{x,x'}$.

Furthermore, we need the closed convex set $K = \ell^1_{d^p_{x_0}}(\mathcal{X}\times\mathcal{X})_+ \times \{\mathbf{0}\} \times \{\mathbf{0}\}$ were $\ell^1_{d^p_{x_0}}(\mathcal{X}\times\mathcal{X})_+$ are the elements in $\ell^1_{d^p_{x_0}}(\mathcal{X}\times\mathcal{X})$ that have only nonnegative entries. With these definitions the pth power of the pth Wasserstein distance is the optimal value of the following abstract parametrized optimization problem:

$$\min_{\boldsymbol{w}\in\ell^1_{d^p_{x_0}}(\mathcal{X}\times\mathcal{X})} f(\boldsymbol{w}) \quad \text{s.t. } C\big(\boldsymbol{w},(\boldsymbol{r},\boldsymbol{s})\big)\in K. \tag{33}$$

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 (30) is obviously convex as it is a linear program with linear constraints. Note that the definition of a convex problem (Definition 2.163) in Bonnans and Shapiro (2000) is slightly different from the usual definition of a convex program as they require convexity of the constraint function (32) *with respect to* $-K$. This condition can be shown by easy calculations for our problem.

The set of primal optimal solutions, $\mathcal{S}(\boldsymbol{r},\boldsymbol{s})$, is according to Theorem 4.1 in Villani (2009) nonempty.

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

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

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), Definition 4.8). According to Theorem 4.9 in Bonnans and Shapiro (2000) the following condition is necessary and sufficient for the directional regularity constraint to hold:

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

where $\mathbb{R}_+(\boldsymbol{d}_1,\boldsymbol{d}_2)=\{t(\boldsymbol{d}_1,\boldsymbol{d}_2), t\geq 0\}$. We are going to show that the directional regularity condition in a direction $(\boldsymbol{d}_1,\boldsymbol{d}_2)\in\mathcal{D}(\boldsymbol{r},\boldsymbol{s})$ holds for all primal optimal solutions $\boldsymbol{w}_0\in\mathcal{S}(\boldsymbol{r},\boldsymbol{s})$.

For a primal optimal solution $\boldsymbol{w}_0$, it is

$$C\big(\boldsymbol{w}_0,(\boldsymbol{r},\boldsymbol{s})\big)=(\boldsymbol{w}_0,\mathbf{0},\mathbf{0}).$$

Since $C(\boldsymbol{w},(\boldsymbol{r},\boldsymbol{s}))$ is linear in $(\boldsymbol{w},(\boldsymbol{r},\boldsymbol{s}))$ and bounded with respect to the product norm on the space $\ell^1_{d^p_{x_0}}(\mathcal{X}\times\mathcal{X})\times\ell^1_{d^p_{x_0}}(\mathcal{X})\times\ell^1_{d^p_{x_0}}(\mathcal{X})$, it holds that

$$DC\big(\boldsymbol{w}_0,(\boldsymbol{r},\boldsymbol{s})\big)\big(\ell^1_{d^p_{x_0}}(\mathcal{X}\times\mathcal{X}),\mathbb{R}_+(\boldsymbol{d}_1,\boldsymbol{d}_2)\big)=(\boldsymbol{w},\Sigma_1\boldsymbol{w}-t\boldsymbol{d}_1,\Sigma_2\boldsymbol{w}-t\boldsymbol{d}_2)$$

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

$$\mathbf{0} \in \operatorname{int}\{(\boldsymbol{w}_0, \mathbf{0}, \mathbf{0}) + (\boldsymbol{w}, \Sigma_1 \boldsymbol{w} - t\boldsymbol{d}_1, \Sigma_2 \boldsymbol{w} - t\boldsymbol{d}_2) - K\}.$$

This set is just $\ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X}) \times \ell^1_{d^p_{x_0}}(\mathcal{X})$ as $\boldsymbol{w} \in \ell^1_{d^p_{x_0}}(\mathcal{X} \times \mathcal{X})$ 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 $\boldsymbol{r}_n = \boldsymbol{r} + t_n \boldsymbol{d}_1 + o(t_n)$ and $\boldsymbol{s}_n = \boldsymbol{s} + t_n \boldsymbol{d}_2 + o(t_n)$ with $t_n \searrow 0$, $\boldsymbol{r}, \boldsymbol{s} \in \mathcal{P}_p(\mathcal{X})$, $\boldsymbol{d}_1 \in \mathcal{D}(\boldsymbol{r})$ and $\boldsymbol{d}_2 \in \mathcal{D}(s)$ there exist a sequence of primal optimal solutions $\boldsymbol{w}_n$ that converges to the primal optimal solution $\boldsymbol{w}_0$ of the unperturbed problem. For n large enough $t_n \leq 1$, hence we can assume without loss of generality that $t_n \leq 1$ for all n. In this case, $\boldsymbol{r}_n$ and $\boldsymbol{s}_n$ are probability measure with existing pth moment, that is, elements of $\mathcal{P}_p(\mathcal{X})$. This yields that Theorem 5.20 in Villani (2009) is applicable. This theorem gives us the stability of the optimal solution as $\mathcal{P}_p(\mathcal{X})$ is a closed subset of $\ell^1_{d^p_{x_0}}(\mathcal{X})$.

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 (31) from the result of that theorem.

The Lagrangian L of a parametrized optimization problem

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

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 and C the constraint function and $\langle \cdot, \cdot \rangle$ the dual pairing (see, e.g., Section 2.5.2 in Bonnans and Shapiro (2000)). We refer to λ as Lagrange multiplier. For the transport problem, this yields with $(\boldsymbol{r}, \boldsymbol{s})$ being the parameter and the definition of the constraint function in (32)

$$\begin{aligned} &L\big(\boldsymbol{w}, (\boldsymbol{v}, \boldsymbol{\lambda}, \boldsymbol{\mu}), (\boldsymbol{r}, \boldsymbol{s})\big) \\ &\quad = \sum_{x, x' \in \mathcal{X}} d^p(x, x') w_{x,x'} + \langle \boldsymbol{v}, \boldsymbol{w} \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{w}^T \mathbb{1} - \boldsymbol{r} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{w} \mathbb{1} - \boldsymbol{s} \rangle. \end{aligned}$$

Differentiating this in the Fréchet sense with respect to $(\boldsymbol{r}, \boldsymbol{s})$ and applying $(\boldsymbol{d}_1, \boldsymbol{d}_2)$ to this linear operator results in

$$D_{(r,s)} L\big(\boldsymbol{w}, (\boldsymbol{v}, \boldsymbol{\lambda}, \boldsymbol{\mu}), (\boldsymbol{r}, \boldsymbol{s})\big)(\boldsymbol{d}_1, \boldsymbol{d}_2) = -\big(\langle \boldsymbol{\lambda}, \boldsymbol{d}_1 \rangle + \langle \boldsymbol{\mu}, \boldsymbol{d}_2 \rangle\big)$$

as the Lagrangian is linear and bounded in $(\boldsymbol{r}, \boldsymbol{s})$. As this derivative is independent of $\boldsymbol{w}$ and the set of Lagrange multipliers $\Lambda(\boldsymbol{r}, \boldsymbol{s})$ equals the set of dual solutions $\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{s})$ 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

$$\begin{aligned}(\boldsymbol{d}_1, \boldsymbol{d}_2) \mapsto & \inf_{\boldsymbol{w}\in\mathcal{S}(\boldsymbol{r},\boldsymbol{s})} \sup_{(\boldsymbol{\lambda},\boldsymbol{\mu})\in\Lambda(\boldsymbol{r},\boldsymbol{s})} D_{(r,s)}L\big(\boldsymbol{w}, (\boldsymbol{v}, \boldsymbol{\lambda}, \boldsymbol{\mu}), (\boldsymbol{r}, \boldsymbol{s})\big)(\boldsymbol{d}_1, \boldsymbol{d}_2) \\ = & \inf_{\boldsymbol{w}\in\mathcal{S}(\boldsymbol{r},\boldsymbol{s})} \sup_{(\boldsymbol{\lambda},\boldsymbol{\mu})\in\Lambda(\boldsymbol{r},\boldsymbol{s})} -\big(\langle\boldsymbol{\lambda}, \boldsymbol{d}_1\rangle + \langle\boldsymbol{\mu}, \boldsymbol{d}_2\rangle\big) \\ = & \sup_{(\boldsymbol{\lambda},\boldsymbol{\mu})\in\mathcal{S}^*(\boldsymbol{r},\boldsymbol{s})} -\big(\langle\boldsymbol{\lambda}, \boldsymbol{d}_1\rangle + \langle\boldsymbol{\mu}, \boldsymbol{d}_2\rangle\big).\end{aligned}$$

□

A.2. The limit distribution under equality of measures. First, observe that for the case $\boldsymbol{r} = \boldsymbol{s}$ the set of dual solutions $\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{r})$ in (6) reduces to

$$\begin{aligned}\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{r}) = \big\{&(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) \times \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) : \langle\boldsymbol{r}, \boldsymbol{\lambda}\rangle + \langle\boldsymbol{r}, \boldsymbol{\mu}\rangle = 0, \\ &\lambda_x + \mu_{x'} \le d^p\big(x, x'\big)\ \forall x, x' \in \mathcal{X}\big\} \\ = \big\{&(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) \times \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) : \lambda_x = -\mu_x \text{ for } x \in \operatorname{supp}(\boldsymbol{r}), \\ &\lambda_x + \mu_{x'} \le d^p\big(x, x'\big)\ \forall x, x' \in \mathcal{X}\big\}.\end{aligned}$$

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 nonnegative. The conjunction of these two conditions yields $\lambda_x + \mu_x = 0$.

This set is a subset of the set given in (7), but changing $\mathcal{S}^*(\boldsymbol{r}, \boldsymbol{r})$ to $\mathcal{S}^*(\boldsymbol{r})$ does not change the optimal value of the linear programs in Theorems 2.1 and 2.4 as the Gaussian process $\mathbf{G}$ is zero at all $x \notin \operatorname{supp}(\boldsymbol{r})$.

In the case, that the support of $\boldsymbol{r}$, that is, $\{x \in \mathcal{X} : r_x > 0\}$, is the whole ground space $\mathcal{X}$, the set $\mathcal{S}^*(\boldsymbol{r})$ is independent of $\boldsymbol{r}$ and it reduces to

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

PROOF OF THEOREM 2.4(a). For the two sample case, the delta method together with the continuous mapping theorem and equation (15) gives

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

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

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

The independence of $\mathbf{G}$ and $\mathbf{G}'$ yield 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

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

□

PROOF OF DECOMPOSITION IN REMARK 2.3(c). For the alternative representation of the distributional limit we decompose the Gaussian process $\mathbf{G}$ with mean zero and covariance structure as defined in (8) into $\mathbf{G} = \mathbf{G}^+ - \mathbf{G}^-$ with $\mathbf{G}^+$, $\mathbf{G}^-$ nonnegative, then the limiting distribution in (9) can be rewritten as follows:

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

The Lagrangian for this problem is given by

$$\begin{aligned}L(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{w},\boldsymbol{z}) &= \sum_{x\in\mathcal{X}} G^+_x\lambda_x + \sum_{x'\in\mathcal{X}} G^-_{x'}\mu_{x'} \\ &\quad + \sum_{x\in\mathcal{X}} z_x(\lambda_x+\mu_x)\mathbb{1}_{\{r_x>0\}} + \sum_{x,x'\in\mathcal{X}} w_{x,x'}\big(\lambda_x+\mu_{x'}-d^p(x,x')\big).\end{aligned}$$

From this, we can derive the dual via

$$\min_{\boldsymbol{w}\ge 0\in\ell^1_{d^p_{x_0}}(\mathcal{X}\times\mathcal{X}),\boldsymbol{z}\in\ell^1_{d^p_{x_0}}(\mathcal{X})} \sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\in\ell^\infty_{d^{-p}_{x_0}}(\mathcal{X})} L(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{w},\boldsymbol{z}),$$

where $\boldsymbol{w}\ge 0$ to be understood componentwise. It yields

$$\begin{aligned}&\inf_{\boldsymbol{w}\ge 0,\boldsymbol{z}} \sum_{x,x'\in\mathcal{X}} d^p(x,x')w_{x,x'} \\ &\quad \text{s.t. } \sum_{x'\in\mathcal{X}} w_{x,x'} = G^+_x + z_x\mathbb{1}_{\{r_x>0\}} \\ &\quad\quad \sum_{x\in\mathcal{X}} w_{x,x'} = G^-_{x'} + z_{x'}\mathbb{1}_{\{r_x>0\}},\end{aligned}$$

where the minimum over $\boldsymbol{w}$ equals the pth power of the pth Wasserstein distance. More precisely, the linear program above is equivalent to

$$\inf_{\boldsymbol{z}(\boldsymbol{r})\in\ell^1_{d^p_{x_0}}(\mathcal{X})} W_p^p\big(\mathbf{G}^+ + \boldsymbol{z}(\boldsymbol{r}), \mathbf{G}^- + \boldsymbol{z}(\boldsymbol{r})\big),$$

where $\boldsymbol{z}(\boldsymbol{r})$ depends on $\boldsymbol{r}$ through the support of $\boldsymbol{r}$ in the following sense: $z_x = 0$ for $x\in\mathcal{X}$ such that $r_x = 0$. $\square$

A.3. Proof of Theorem 3.1.

Simplify the set of dual solutions $\mathcal{S}^*$. As a first step, we rewrite the set of dual solutions $\mathcal{S}^*$ given in definition (12) in our tree notation as

$$\mathcal{S}^* = \{\boldsymbol{\lambda} \in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) : \lambda_x - \lambda_{x'} \le d_{\mathcal{T}}(x, x')^p, x, x' \in \mathcal{X}\}. \tag{34}$$

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 $x = x_1, \dots, x_n = x'$ the sequence of vertices defining the unique path joining x and x', such that $(x_j, x_{j+1}) \in E$ for $j = 1, \dots, n-1$; that this path contains only a finite number of edges, was proven in Section 3. Then

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

such that (34) is satisfied for all $x, x' \in \mathcal{X}$. 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

$$\mathcal{S}^* = \{\boldsymbol{\lambda} \in \ell^\infty_{d^{-p}_{x_0}}(\mathcal{X}) : |\lambda_x - \lambda_{\text{parent}(x)}| \le d_{\mathcal{T}}(x, \text{parent}(x))^p, x \in \mathcal{X}\}. \tag{35}$$

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

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

Furthermore, we define for $\boldsymbol{\mu} \in \ell^1_{d^p_{x_0}}(\mathcal{X})$,

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

and

$$\boldsymbol{\mu}_n = \sum_{x \in A_{\le n} \setminus \text{root}(\mathcal{T})} \eta_x \tilde{\boldsymbol{e}}^{(x)} = \boldsymbol{\mu} \mathbb{1}_{A_{<n}} + \sum_{x \in A_{=n}} \frac{1}{d^p(x, x_0)} \eta_x \boldsymbol{e}(x).$$

Here,

$A_{\le n} = \{x \in \mathcal{X}$: level of $x \le n$, x is within the first n vertices of its level$\}$,

$A_{=n} = \{x \in \mathcal{X}$: level of $x = n$, x is within the first n vertices of its level$\}$,

$A_{>n} = \{x \in \mathcal{X}$: level of $x > n$ or x is not within the first n vertices of its level$\}$

and $\boldsymbol{e}(x)$ the sequence 1 at x and 0 everywhere else. For this sequence $\boldsymbol{\mu}_n$, it holds

$$\begin{aligned}\|\boldsymbol{\mu}-\boldsymbol{\mu}_n\|_{\ell^1_{d^p_{x_0}}} &= \sum_{x\in X} d^p(x,x_0)\Big|\boldsymbol{\mu}\mathbb{1}_{A_{>n}} - \sum_{\tilde{x}\in A_{=n}} \frac{1}{d^p(\tilde{x},x_0)}\eta_{\tilde{x}}\boldsymbol{e}^{(\tilde{x})}\Big|_x \\ &\le \|\boldsymbol{\mu}\mathbb{1}_{A_{>n}}\|_{\ell^1_{d^p_{x_0}}} + \Big|\sum_{x\in A_{=n}}\eta_x\Big|.\end{aligned}$$

As $n\to\infty$, the first part tends to zero as $\boldsymbol{\mu}\in\ell^1_{d^p_{x_0}}(\mathcal{X})$, and

$$\begin{aligned}\Big|\sum_{x\in A_{=n}}\eta_x\Big| &\le \sum_{x\in A_{=n}}\sum_{x'\in\text{children}(x)}|\mu_{x'}|d^p(x',x_0) \\ &\le \sum_{x\in A_{\ge n}}|\mu_x|d^p(x,x_0)\xrightarrow{n\to\infty}0.\end{aligned}$$

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

$$\begin{aligned}\langle\boldsymbol{\mu},\boldsymbol{\lambda}\rangle &= \lim_{n\to\infty}\langle\boldsymbol{\mu}_n,\boldsymbol{\lambda}\rangle \\ &= \lim_{n\to\infty}\sum_{x\in A_{\le n}}\eta_x\langle\tilde{\boldsymbol{e}}^{(x)},\boldsymbol{\lambda}\rangle \\ &= \lim_{n\to\infty}\sum_{x\in A_{\le n}}\sum_{x'\in\text{children}(x)}\mu_{x'}(\lambda_x-\lambda_{\text{parent}(x)}) \\ &\le \lim_{n\to\infty}\sum_{x\in A_{\le n}}\Big|\sum_{x'\in\text{children}(x)}\mu_{x'}\Big||\lambda_x-\lambda_{\text{parent}(x)}| \\ &= \lim_{n\to\infty}\sum_{x\in A_{\le n}}\big|(S_{\mathcal{T}}\boldsymbol{\mu})_x\big||\lambda_x-\lambda_{\text{parent}(x)}|.\end{aligned}\tag{36}$$

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

$$|\lambda_x-\lambda_{\text{parent}(x)}|\le d^P\big(x,\text{parent}(x)\big).\tag{37}$$

By condition (3) $\boldsymbol{G}\sim\mathcal{N}(0,\Sigma(\boldsymbol{r}))$ is an element of $\ell^1_{d^p_{x_0}}(\mathcal{X})$. For $\boldsymbol{\lambda}\in\mathcal{S}^*$, we get with (36) and (37) that

$$\langle\boldsymbol{G},\boldsymbol{\lambda}\rangle\le\lim_{n\to\infty}\sum_{x\in A_{\le n}}\big|(S_{\mathcal{T}}\mathbf{G})_x\big|d_{\mathcal{T}}\big(x,\text{parent}(x)\big)^p.\tag{38}$$

Therefore, $\max_{\boldsymbol{\lambda}\in\mathcal{S}^*}\langle\boldsymbol{G},\boldsymbol{\lambda}\rangle$ is bounded by $\lim_{n\to\infty}\sum_{x\in A_{\le n}}|(S_{\mathcal{T}}\boldsymbol{G})_x|d_{\mathcal{T}}(x,\text{parent}(x))^p$. We can define the sequence $\boldsymbol{\nu}\in\ell^\infty_{d^{-p}_{x_0}}(\mathcal{X})$ by

$$\begin{aligned}\nu_{\text{root}} &= 0, \\ \nu_x-\nu_{\text{parent}(x)} &= \text{sign}\big((S_{\mathcal{T}}\mathbf{G})_x\big)d_{\mathcal{T}}\big(x,\text{parent}(x)\big)^p.\end{aligned}\tag{39}$$

From (35) and the fact that $d^p(x, \text{parent}(x)) \le d^p(x, \text{root}(\mathcal{T}))$, we see that $\boldsymbol{v} \in \mathcal{S}^*$ and by plugging $\boldsymbol{v}$ into equation (38) we can conclude that $\langle \boldsymbol{G}, \boldsymbol{v} \rangle$ attains the upper bound in (38).

As the last step of our proof, we verify that the limit in (38) exists. Therefore, we rewrite condition (3) in terms of the edges and recall that $x_0 = \text{root}(\mathcal{T})$:

$$\sum_{x \in \mathcal{X}} d_{\mathcal{T}}(x, x_0)^p \sqrt{r_x} \ge \sum_{x \in \mathcal{X}} \sum_{x' \in \text{children}(x)} d_{\mathcal{T}}\big(x, \text{parent}(x)\big)^p \sqrt{r_{x'}}. \tag{40}$$

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

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{x \in \mathcal{X} \setminus \{\text{root}(\mathcal{T})\}} \big|(S_{\mathcal{T}} \boldsymbol{G})_x\big| d_{\mathcal{T}}\big(x, \text{parent}(x)\big)^p \Bigg] \\
&\quad \le \sum_{x \in \mathcal{X}} d_{\mathcal{T}}\big(x, \text{parent}(x)\big)^p \sqrt{(S_{\mathcal{T}} r)_x \big(1 - (S_{\mathcal{T}} r)_x\big)} \\
&\quad \le \sum_{x \in \mathcal{X}} \sum_{x' \in \text{children}(x)} d_{\mathcal{T}}\big(x, \text{parent}(x)\big)^p \sqrt{r_{x'}} \\
&\quad < \infty
\end{aligned}$$

due to Hölder's inequality and (40). This bound shows that the limit in (38) is almost surely finite, and hence, concludes the proof.

A.4. Proof of Theorem (4.4). Define $\text{supp}(\boldsymbol{r})'$ by adding to $\text{supp}(\boldsymbol{r})$ all center-points of sets in P_l for $0 \le l < l_{\max}$. We identify center points of $P_{l_{\max}}$ with the points in $\text{supp}(\boldsymbol{r})$. A tree with vertices $\text{supp}(\boldsymbol{r})'$ can now be built using the inclusion relation of the sets $\{P_l\}_{0 \le l \le l_{\max}}$ as an ancestry relation. More precisely, the leaves of the tree are the points of $\text{supp}(\boldsymbol{r})$ and the parent of the center point of $F \in P_l$ is the center point of the unique set in P_{l-1} that contains F. If we use the Euclidean metric to define the distance between neighboring vertices, we get

$$d_{\mathcal{T}}\big(x, \text{parent}(x)\big) = \frac{\sqrt{D} 2^{-l}}{2},$$

if $x \in P_l$. A measure $\boldsymbol{r}$ naturally extends to a measure on $\text{supp}(\boldsymbol{r})'$ if we give zero mass to all inner vertices. We also denote this measure by $\boldsymbol{r}$. Then, if $x \in \text{supp}(\boldsymbol{r})'$ is the center point of the set $F \in P_l$ for some $0 \le l \le l_{\max}$, we have that $(S_{\mathcal{T}} \boldsymbol{r})_x = S_F \boldsymbol{r}$ where $S_F \boldsymbol{r} = \sum_{x \in F} r_x$. Inserting these two formulas into (23) yields (28).

APPENDIX B: ADDITIONAL MATERIAL TO SECTION 2.3

PROOF OF THEOREM 2.10. For the SEF in (16), condition (3) reads

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} d^p(x_0,x)\sqrt{\exp\left(\sum_{i=1}^{s}\eta_i T_x^i - A(\boldsymbol{\eta})\right)h_x} \\
&\quad = \frac{1}{\sqrt{\lambda(\boldsymbol{\eta})}}\sum_{x\in\mathcal{X}} d^p(x_0,x)\exp\left(\frac{1}{2}\sum_{i=1}^{s}\eta_i T_x^i\right)\sqrt{h_x} \qquad (41)\\
&\quad \le \frac{\lambda(\frac{1}{2}\boldsymbol{\eta})}{\sqrt{\lambda(\boldsymbol{\eta})}}\sum_{x\in\mathcal{X}} d^p(x_0,x)\exp\left(\frac{1}{2}\sum_{i=1}^{s}\eta_i T_x^i\right)h_x < \infty,
\end{aligned}$$

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

The following examples show that all three conditions in Theorem 2.10 are necessary.

EXAMPLE B.1. Let $\mathcal{X}$ be the countable metric space $\mathcal{X} = \{\frac{1}{k}\}_{k\in\mathbb{N}}$ 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 a 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 $\mathcal{X}$ and set $x_0 = 1$. It holds

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

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 with $\frac{1}{2}$ and, therefore,

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

that is, condition (3) is not fulfilled.

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

EXAMPLE B.2. Consider $\mathcal{X} = \mathbb{N}$ with the metric $d(k,l) = \sqrt{|k! - l!|}$. The family of Poisson distributions constitute a SEF with natural parameter space $\mathcal{N} = (-\infty, \infty)$ which satisfies condition (2) in Theorem 2.10, that is, closed with respect to multiplication with $\frac{1}{2}$. The first moment with respect to this metric exists and $h_k < 1$ for all $k \geq 2$. Condition (3) 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 $\eta > 1$, that is, the summability condition (3) is not fulfilled.

If the pth moment does not exist, it is clear that condition (3) cannot be fulfilled as $\sqrt{x} \geq x$ for $x \in [0,1]$.

Acknowledgments. The authors would like to thank M. Klatt and Y. Zemel for careful reading of the manuscript. A. Munk is grateful to helpful comments of J. Wellner. Furthermore, the authors are grateful for the comments of the reviewers which improved this manuscript.

REFERENCES

AJTAI, M., KOMLÓS, J. and TUSNÁDY, G. (1984). On optimal matchings. *Combinatorica* **4** 259–264. MR0779885

ASPELMEIER, T., EGNER, A. and MUNK, A. (2015). Modern statistical challenges in high-resolution fluorescence microscopy. *Annu. Rev. Stat. Appl.* **2** 163–202.

BARBOUR, A. D. and BROWN, T. C. (1992). Stein's method and point process approximation. *Stochastic Process. Appl.* **43** 9–31. MR1190904

BERTSEKAS, D. P. (1981). A new algorithm for the assignment problem. *Math. Program.* **21** 152–171. MR0623835

BERTSEKAS, D. P. (1992). Auction algorithms for network flow problems: A tutorial introduction. *Comput. Optim. Appl.* **1** 7–66. MR1195629

BERTSEKAS, D. P. (2009). Auction algorithms. In *Encyclopedia of Optimization* 128–132. Springer, Berlin.

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** 1642–1645.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103

BOBKOV, S. and LEDOUX, M. (2014). One-dimensional empirical measures, order statistics and Kantorovich transport distances. Preprint.

BOISSARD, E. and LE GOUIC, T. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 539–563. MR3189084

BONNANS, J. F. and SHAPIRO, A. (2000). *Perturbation Analysis of Optimization Problems. Springer Series in Operations Research*. Springer, New York. MR1756264

BONNEEL, N., RABIN, J., PEYRÉ, G. and PFISTER, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imaging Vision* **51** 22–45. MR3300482

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

DE WET, T. and VENTER, J. H. (1972). Asymptotic distributions of certain test criteria of normality. *South African Statist. J.* **6** 135–149. MR0329116

DEDE, S. (2009). An empirical central limit theorem in $\mathbf{L}^1$ for stationary sequences. *Stochastic Process. Appl.* **119** 3494–3515. MR2568284

DEDECKER, J. and MERLEVÈDE, F. (2017). Behavior of the Wasserstein distance between the empirical and the marginal distributions of stationary α-dependent sequences. *Bernoulli* **23** 2083–2127. MR3624887

DEL BARRIO, E., CUESTA-ALBERTOS, J. A. and MATRÁN, C. (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *TEST* **9** 1–96. MR1790430

DEL BARRIO, E., GINÉ, E. and MATRÁN, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.* **27** 1009–1071. MR1698999

DEL BARRIO, E., GINÉ, E. and UTZET, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli* **11** 131–189. MR2121458

DEL BARRIO, E. and LOUBES, J.-M. (2017). Central limit theorems for empirical transportation cost in general dimension. Preprint. Available at arXiv:1705.01299v1.

DEL BARRIO, E., CUESTA-ALBERTOS, J. A., MATRÁN, C. and RODRÍGUEZ-RODRÍGUEZ, J. M. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *Ann. Statist.* **27** 1230–1239. MR1740113

DESCHOUT, H., ZANACCHI, F. C., MLODZIANOSKI, M., DIASPRO, A., BEWERSDORF, J., HESS, S. T. and BRAECKMANS, K. (2014). Precisely and accurately localizing single emitters in fluorescence microscopy. *Nat. Methods* **11** 253–266.

DÜMBGEN, L. (1993). On nondifferentiable functions and the bootstrap. *Probab. Theory Related Fields* **95** 125–140. MR1207311

EBERLE, A. (2014). Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *Ann. Appl. Probab.* **24** 337–377. MR3161650

EGNER, A., GEISLER, C., VON MIDDENDORFF, C., BOCK, H., WENZEL, D., MEDDA, R., ANDRESEN, M., STIEL, A. C., JAKOBS, S. et al. (2007). Fluorescence nanoscopy in whole cells by asynchronous localization of photoswitching emitters. *Biophys. J.* **93** 3285–3290.

EVANS, S. N. and MATSEN, F. A. (2012). The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 569–592. MR2925374

FÖLLING, J., BOSSI, M., BOCK, H., MEDDA, R., WURM, C. A., HEIN, B., JAKOBS, S., EGGELING, C. and HELL, S. W. (2008). Fluorescence nanoscopy by ground-state depletion and single-molecule return. *Nat. Methods* **5** 943–945.

FOURNIER, N. and GUILLIN, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields* **162** 707–738. MR3383341

FREITAG, G. and MUNK, A. (2005). On Hadamard differentiability in k-sample semiparametric models—With applications to the assessment of structural relationships. *J. Multivariate Anal.* **94** 123–158. MR2161214

GEISLER, C., HOTZ, T., SCHÖNLE, A., HELL, S. W., MUNK, A. and EGNER, A. (2012). Drift estimation for single marker switching based imaging schemes. *Opt. Express* **20** 7274–7289.

GOTTSCHLICH, C. and SCHUHMACHER, D. (2014). The shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems. *PLoS ONE* **9** e110214.

HARTMANN, A., HUCKEMANN, S., DANNEMANN, J., LAITENBERGER, O., GEISLER, C., EGNER, A. and MUNK, A. (2016). Drift estimation in sparse sequential dynamic imaging, with application to nanoscale fluoresence microscopy. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 563–587. MR3506793

HEILEMANN, M., VAN DE LINDE, S., SCHÜTTPELZ, M., KASPER, R., SEEFELDT, B., MUKHERJEE, A., TINNEFELD, P. and SAUER, M. (2008). Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. *Angew. Chem., Int. Ed. Engl.* **47** 6172–6176.

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

HOROWITZ, J. and KARANDIKAR, R. L. (1994). Mean rates of convergence of empirical measures in the Wasserstein metric. *J. Comput. Appl. Math.* **55** 261–273. MR1329874

HUNG, M. S., ROM, W. O. and WAREN, A. D. (1986). Degeneracy in transportation problems. *Discrete Appl. Math.* **13** 223–237. MR0837943

JAIN, N. C. (1977). Central limit theorem and related questions in Banach space. In *Probability* (*Proc. Sympos. Pure Math., Univ. Illinois, Urbana, Ill.*, 1976) **31** 55–65. Amer. Math. Soc., Providence, RI. MR0451328

JOHNSON, O. and SAMWORTH, R. (2005). Central limit theorem and convergence to stable laws in Mallows distance. *Bernoulli* **11** 829–845. MR2172843

KANTOROVIČ, L. V. and RUBINŠTEĬN, G. Š. (1958). On a space of completely additive functions. *Vestn. Leningr. Univ.* **13** 52–59. MR0102006

KLEE, V. and WITZGALL, C. (1968). Facets and vertices of transportation polytopes. In *Mathematics of the Decision Sciences, Part I* (*Seminar, Stanford, Calif.*, 1967) 257–282. Amer. Math. Soc., Providence, RI. MR0235832

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR1639875

LING, H. and OKADA, K. (2007). An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Trans. Pattern Anal. Mach. Intell.* **29** 840–853.

LUENBERGER, D. G. and YE, Y. (2008). *Linear and Nonlinear Programming*, 3rd ed. *International Series in Operations Research & Management Science* **116**. Springer, New York. MR2423726

MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Ann. Math. Stat.* **43** 508–515. MR0298812

MASON, D. M. (2016). A weighted approximation approach to the study of the empirical Wasserstein distance. In *High Dimensional Probability VII. Progress in Probability* **71** 137–154. Springer, Cham. MR3565262

MAUREY, B. (1973). Espaces de cotype p, $0 < p \leq 2$. In *Séminaire Maurey–Schwartz* (*année* 1972–1973), *Espaces L^p et applications radonifiantes*, *Exp. No.* 7 1–11. Centre de Math., École Polytech., Paris. MR0394093

MUNK, A. and CZADO, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 223–241. MR1625620

NI, K., BRESSON, X., CHAN, T. and ESEDOGLU, S. (2009). Local histogram based segmentation using the Wasserstein distance. *Int. J. Comput. Vis.* **84** 97–111.

ORLIN, J. B. (1993). A faster strongly polynomial minimum cost flow algorithm. *Oper. Res.* **41** 338–350. MR1214540

PANARETOS, V. M. and ZEMEL, Y. (2016). Amplitude and phase variation of point processes. *Ann. Statist.* **44** 771–812. MR3476617

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

RACHEV, S. T. and RÜSCHENDORF, L. (1994). On the rate of convergence in the CLT with respect to the Kantorovich metric. In *Probability in Banach Spaces*, 9 (*Sandjberg*, 1993). *Progress in Probability* **35** 193–207. Birkhäuser, Boston, MA. MR1308518

RACHEV, S. T. and RÜSCHENDORF, L. (1998). *Mass Transportation Problems, Vol. I: Theory. Probability and Its Applications* (*New York*). Springer, New York. MR1619170

RACHEV, S. T., STOYANOV, S. V. and FABOZZI, F. J. (2011). *A Probability Metrics Approach to Financial Risk Measures*. Wiley, New York.

RIPPL, T., MUNK, A. and STURM, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.* **151** 90–109. MR3545279

ROLET, A., CUTURI, M. and PEYRÉ, G. (2016). Fast dictionary learning with a smoothed Wasserstein loss. In *Proceedings of the* 19*th International Conference on Artificial Intelligence and Statistics* (A. Gretton and C. C. Robert, eds.). *Proceedings of Machine Learning Research* **51** 630–638. PMLR, Cadiz, Spain.

RÖMISCH, W. (2004). Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences* Wiley, New York.

RUBNER, Y., TOMASI, C. and GUIBAS, L. J. (2000). The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **40** 99–121.

RUDOLF, D. and SCHWEIZER, N. (2018). Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli* **24** 2610–2639. MR3779696

RUST, M. J., BATES, M. and ZHUANG, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3** 793–796.

SCHMITZER, B. (2016). A sparse multiscale algorithm for dense optimal transport. *J. Math. Imaging Vision* **56** 238–259. MR3535020

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** 550–568. MR2543874

SHAPIRO, A. (1990). On concepts of directional differentiability. *J. Optim. Theory Appl.* **66** 477–487. MR1080259

SHAPIRO, A. (1991). Asymptotic analysis of stochastic programs. *Ann. Oper. Res.* **30** 169–186. MR1118896

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

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0838963

SOLOMON, J., DE GOES, F., PEYRÉ, G., CUTURI, M., BUTSCHER, A., NGUYEN, A., DU, T. and GUIBAS, L. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.* **34** 66.

SOMMERFELD, M. and MUNK, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 219–238. MR3744719

TALAGRAND, M. (1992). Matching random samples in many dimensions. *Ann. Appl. Probab.* **2** 846–856. MR1189420

TALAGRAND, M. (1994). The transportation cost from the uniform measure to the empirical measure in dimension ≥ 3. *Ann. Probab.* **22** 919–959. MR1288137

TAMELING, C. and MUNK, A. (2018). Computational strategies for inference based on empirical optimal transport.

VASERSHTEIN, L. N. (1969). Markov processes over denumerable products of spaces describing large system of automata. *Problemy Peredachi Informatsii* **5** 64–72. MR0314115

VILLANI, C. (2003). *Topics in Optimal Transportation*. *Graduate Studies in Mathematics* **58**. Amer. Math. Soc., Providence, RI. MR1964483

VILLANI, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454

WEED, J. and BACH, F. (2017). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. Preprint. Available at arXiv:1707.00087.

C. TAMELING
M. SOMMERFELD
A. MUNK
INSTITUTE FOR MATHEMATICAL STATISTICS
UNIVERSITY OF GOETTINGEN
37075 GÖTTINGEN
GERMANY
E-MAIL: carla.tameling@mathematik.uni-goettingen.de
maxsommerfeld@gmail.com
munk@math.uni-goettingen.de