Inference for Empirical Wasserstein Distances on Finite Spaces: Supplementary Material

Max Sommerfeld* Axel Munk*†

Keywords: optimal transport, Wasserstein distance, central limit theorem, directional Hadamard derivative, bootstrap, hypothesis testing

AMS 2010 Subject Classification: Primary: 62G20, 62G10, 65C60 Secondary: 90C08, 90C31

A An alternative representation of the limiting distribution

We give a second representation of the limiting distribution under the alternative $\boldsymbol{r} \neq \boldsymbol{s}$. The random part of the limiting distribution (8) is the linear program

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

With the representation (3) of $\Phi_p^*(\boldsymbol{r}, \boldsymbol{s})$ we obtain the dual linear program

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

Note that the constraints can only be satisfied if both $\sqrt{\lambda}\boldsymbol{G} - z\boldsymbol{r}$ and $\sqrt{1-\lambda}\boldsymbol{H} - z\boldsymbol{s}$ have only non-negative entries and $z \leq 0$. In this case the

*Felix Bernstein Institute for Mathematical Statistics in the Biosciences and Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen

†Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen

second term in the objective function is clearly minimized by $-z\boldsymbol{w}^*$, with $\boldsymbol{w}^*$ an optimal transport plan between these two measures $\boldsymbol{r} - \sqrt{\lambda}\boldsymbol{G}/z$ and $\boldsymbol{s} - \sqrt{1-\lambda}\boldsymbol{H}/z$ and the second term of the objective function is equal to $-zW_p^p(\boldsymbol{r} - \sqrt{\lambda}\boldsymbol{G}/z, \boldsymbol{s} - \sqrt{1-\lambda}\boldsymbol{H}/z)$.

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

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

With this we can thus write the random variable in the limiting distribution (8) as the one-dimensional non-linear optimization problem

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

B Bootstrap

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

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

is consistently estimated by the law of

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

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

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

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

$$\sup_{f \in \mathrm{BL}_1(\mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}})} |E\left[f(\rho_{n,m}\left\{(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - (\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\}) | X_1, \ldots, X_n, Y_1, \ldots, Y_m\right] - E\left[f((\sqrt{\lambda}\boldsymbol{G}, \sqrt{1-\lambda}\boldsymbol{H}))\right]| \tag{2}$$

converges to zero in probability.

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

$$\rho_{n,m}\left\{W_p^p(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - W_p^p(\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\} \tag{3}$$

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

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

Recall from Section 2 that

$$\phi_p : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}, \quad \phi_p(\boldsymbol{h}_1, \boldsymbol{h}_2) = \max_{\boldsymbol{u} \in \Phi_p^*} \langle \boldsymbol{u}, \boldsymbol{h}_2 - \boldsymbol{h}_1 \rangle \tag{4}$$

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

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

1. *the plug-in bootstrap* (3) *is not consistent, that is,*

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

does not converge to zero in probability.

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

$$\phi_p(\rho_{n,m}\left\{(\hat{\boldsymbol{r}}_n^*, \hat{\boldsymbol{s}}_m^*) - (\hat{\boldsymbol{r}}_n, \hat{\boldsymbol{s}}_m)\right\}) \tag{5}$$

is consistent, that is

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

converges to zero in probability.

C Proofs

C.1 Proof of Theorem 4

By (Gal et al., 1997, Ch. 3, Thm. 3.1) the function $(\boldsymbol{r}, \boldsymbol{s}) \mapsto W_p^p(\boldsymbol{r}, \boldsymbol{s})$ is directionally differentiable with derivative (11) in the sense of Gâteaux, that is, the limit (9) exists for a fixed $\boldsymbol{h}$ and not a sequence $\boldsymbol{h}_n \to \boldsymbol{h}$ (see e.g. Shapiro (1990)). To see that this is also a directional derivative in the Hadamard sense (9) it suffices (Shapiro, 1990, Prop. 3.5) to show that $(\boldsymbol{r}, \boldsymbol{s}) \mapsto W_p^p(\boldsymbol{r}, \boldsymbol{s})$ is locally Lipschitz. That is, we need to show that for $\boldsymbol{r}, \boldsymbol{r}', \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{P}_{\mathcal{X}}$

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

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

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

for some constant $C > 0$ and some norm $||\cdot||$. To this end, we employ an argument similar to that used to prove the triangle inequality for the Wasserstein distance (see e.g. (Villani, 2008, p. 94)). Indeed, by the gluing Lemma (Villani, 2008, Ch. 1) there exist random variables X_1, X_2, X_3 with

marginal distributions $\boldsymbol{r}, \boldsymbol{s}$ and $\boldsymbol{s}'$, respectively, such that $E[d^p(X_1, X_3)] = W_p^p(\boldsymbol{r}, \boldsymbol{s}')$ and $E[d(X_2, X_3)] = W_1(\boldsymbol{s}, \boldsymbol{s}')$. Then, since (X_1, X_2) has marginals $\boldsymbol{r}$ and $\boldsymbol{s}$, we have

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

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

C.2 Proof of Theorem 5

Simplify the set of dual solutions Φ_p^* As a first step, we rewrite the set of dual solutions Φ_p^* given in (3) in our tree notation as

$$\Phi_p^* = \left\{\boldsymbol{u} \in \mathbb{R}^{\mathcal{X}} : u_x - u_{x'} \leq d_{\mathcal{T}}(x, x')^p, \quad x, x' \in \mathcal{X}\right\}. \tag{6}$$

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

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

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

$$\Phi_p^* = \left\{\boldsymbol{u} \in \mathbb{R}^{\mathcal{X}} : |u_x - u_{\text{parent}(x)}| \leq d_{\mathcal{T}}(x, \text{parent}(x))^p, \quad x \in \mathcal{X}\right\}. \tag{7}$$

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

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

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

Proof. We compute

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

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

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

and the claim follows. □

If $\boldsymbol{u} \in \Phi_p^*$, as given in (7), we have for $x \neq \mathrm{root}(\mathcal{T})$ that

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

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

$$\langle \boldsymbol{G}, \boldsymbol{u} \rangle = \langle S_{\mathcal{T}} \boldsymbol{G}, D_{\mathcal{T}} \boldsymbol{u} \rangle \leq \sum_{\mathrm{root}(\mathcal{T}) \neq x \in \mathcal{X}} |(S_{\mathcal{T}} \boldsymbol{G})_x| d_{\mathcal{T}}(x, \mathrm{parent}(x))^p. \tag{8}$$

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

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

From (7) we see that $\boldsymbol{v} \in \Phi_p^*$ and Lemma 1 shows that $\langle \boldsymbol{G}, \boldsymbol{v} \rangle$ attains the upper bound in (8). This concludes the proof.

C.3 Proof of Corollary 1

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

$$\text{cov}((S_{\mathcal{T}}\boldsymbol{G})_i, (S_{\mathcal{T}}\boldsymbol{G})_j) = \sum_{\substack{k \leq i \\ l \leq j}} E\left[G_k G_l\right] = \sum_{k \leq i} r_k(1 - r_k) - \sum_{\substack{k \leq i \\ l \leq j \\ k \neq l}} r_k r_l$$

$$= \bar{r}_i - \sum_{\substack{k \leq i \\ l \leq i}} r_k r_l - \sum_{\substack{k \leq i \\ i < l \leq j}} r_k r_l = \bar{r}_i - \bar{r}_i^2 - \bar{r}_i(\bar{r}_j - \bar{r}_i)) = \bar{r}_i - \bar{r}_i \bar{r}_j.$$

Therefore, we have that for a standard Brownian bridge B

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

Together with $d(x_j, \text{parent}(x_j)) = (x_{j+1} - x_j)^2$, and (15) this proves the Corollary.

References

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

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

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

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

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

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

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