SIAM J. MATH. DATA SCI.
Vol. 4, No. 3, pp. 957–978

k-Variance: A Clustered Notion of Variance*

Justin Solomon†, Kristjan Greenewald‡, and Haikady Nagaraja§

Abstract. We introduce k-variance, a generalization of variance built on the machinery of random bipartite matchings. k-variance measures the expected cost of matching two sets of k samples from a distribution to each other, capturing local rather than global information about a measure as k increases; it is easily approximated stochastically using sampling and linear programming. In addition to defining k-variance and proving its basic properties, we provide in-depth analysis of this quantity in several key cases, including one-dimensional measures, clustered measures, and measures concentrated on low-dimensional subsets of $\mathbb{R}^n$. We conclude with experiments and open problems motivated by this new way to summarize distributional shape.

Key words. variance, optimal transport, Wasserstein, clustering

AMS subject classifications. 49Q25, 62G30, 62H30

DOI. 10.1137/20M1385895

1. Introduction.

A key task in statistics and data science is to describe the *shape* of a dataset or distribution in a simple form. The most basic methods for summarizing distributions extract scalar measurements characterizing spread, normality, support, decay, and other aspects of distributional geometry. Among these measurements, the simplest and most popular choice is *variance*, which measures the squared deviation of a random variable from its mean.

A scalar is unlikely to capture all relevant or interesting information about a distribution, and indeed variance is not sensitive to skew, asymmetry, and other structural properties. A typical way to address this issue is to compute higher-order moments, which—if completely known—can often reconstruct a distribution. While this solution works mathematically, each (standarized) moment measures the allotment of mass in a distribution relative to its mean, which is hard to interpret in the multimodal or clustered cases.

In this paper, we introduce a generalization of variance we call k-variance, intended to address some of the issues above. The basic idea of k-variance is to draw $2k$ samples from a distribution and to evaluate the transport cost of matching the first k samples to the second

*Received by the editors December 14, 2020; accepted for publication (in revised form) January 18, 2022; published electronically July 7, 2022.
https://doi.org/10.1137/20M1385895
Funding: The first author acknowledges the generous support of the Army Research Office grants W911NF1710068 and W911NF2010168, the Air Force Office of Scientific Research award FA9550-19-1-031, the National Science Foundation grant IIS-1838071, the CSAIL Systems that Learn program, the MIT-IBM Watson AI Laboratory, the Toyota-CSAIL Joint Research Center, a gift from Adobe Systems, and the Skoltech-MIT Next Generation Program.

†Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (jsolomon@mit.edu).
‡MIT-IBM Watson AI Lab, Cambridge, MA 02142 USA (Kristjan.H.Greenewald@ibm.com).
§Division of Biostatistics, The Ohio State University, Columbus, OH 43210 USA (nagaraja.1@osu.edu).

k samples. k-variance coincides with variance in the $k = 1$ case. But, for larger values of k, samples get matched to closer counterparts in the distribution rather than between different modes, making k-variance a more localized measure of variance.

Our construction of k-variance seems to indicate that a tightly clustered distribution about a few means might have high (1-)variance if those means are far apart, but that k-variance of such a distribution will decay rapidly in k relative to that of a unimodal Gaussian. Indeed, we will prove that this is the case—but only for measures embedded in dimensions $\gtrapprox 5$. In lower dimensions, k-variance exhibits surprising—and somewhat counterintuitive—behavior, which we can capture in detail for one-dimensional k-variance using the theory of order statistics.

k-variance can be approximated using a simple randomized algorithm, wherein we draw $2k$ points and solve a $k \times k$ transportation problem; unsurprisingly, the accuracy of this easy-to-implement estimator can be improved by averaging over multiple draws. We provide variance bounds demonstrating that k-variance requires averaging over fewer such draws as k and/or the ambient dimension increases.

We conclude with some experiments demonstrating the behavior of k-variance as a measure of intramode variability as well as a number of open problems motivated by our work.

Contributions. We introduce a generalization of variance for probability measures on $\mathbb{R}^n$ that we call "k-variance," built on constructions from optimal transport. Beyond introducing k-variance and its basic properties (section 4), we

- give alternative expressions and bounds for k-variance of probability measures on $\mathbb{R}$ (section 5);
- use results in empirical optimal transport to characterize k-variance of probability measures concentrated on low-dimensional sets (section 6), higher-dimensional sets (section 7), and with cluster structure (section 8);
- bound the variance of empirical estimators for k-variance in terms of sample size and dimension (section 9); and
- provide numerical experiments to demonstrate behavior of k-variance and confirm our predicted theory (section 10).

2. Related work.

For the most part, we incorporate discussion of related work into the text below as it arises; our work principally uses results from the theory of optimal transport (cf. [24, 22, 19]) and—in one dimension—from the theory of order statistics (cf. [10]).

Before commencing our technical discussion, however, we note that our work is built on recent advances in the theory of *random Euclidean bipartite matchings*. This theory seeks to characterize the cost of matching two independently drawn k-samples of a measure to one another, where the cost of matching two points is proportional to the pth power of Euclidean distance. See [13, 4, 11, 12, 15] and references therein for relevant mathematical theory, and see [26, 7] for applications in other disciplines. While these works focus on bounding the transport cost in specific cases or connecting it to physical applications, here we show how the matching cost can be understood as a generalization of variance useful for characterizing the shape of a probability measure.

3. Preliminaries.

We begin with mathematical preliminaries to establish notation.

3.1. Variance. Our work focuses on generalizing the *variance* of a random variable X drawn from a probability measure $\mu \in \text{Prob}(\mathbb{R}^d)$, which is the expected squared deviation of that variable from its mean $\overline{X} := \mathbb{E}[X]$:

$$\text{Var}(X) := \mathbb{E}_{X\sim\mu}[\|X - \overline{X}\|_2^2]. \tag{3.1}$$

A simple argument reveals an alternative formula for variance:

$$\text{Var}(X) = \frac{1}{2}\mathbb{E}_{X,Y\sim\mu}[\|X - Y\|_2^2]. \tag{3.2}$$

3.2. Optimal transport. Take $\mu, \nu \in \text{Prob}(\mathbb{R}^d)$ to be two Radon probability measures. Then, we can define the (squared) 2-*Wasserstein distance* between μ and ν via

$$\mathcal{W}_2^2(\mu, \nu) := \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}[\|X - Y\|_2^2], \tag{3.3}$$

where $\Pi(\mu, \nu) \subseteq \text{Prob}(\mathbb{R}^d \times \mathbb{R}^d)$ denotes the set of *measure couplings* whose marginals are μ and ν, resp. The Wasserstein distance is a basic object of study in analysis, statistics, machine learning, and related disciplines. Intuitively, $\mathcal{W}_2(\mu, \nu)$ measures the amount of work it takes to displace μ onto ν as distributions of mass over $\mathbb{R}^d$, where the cost of moving a particle of mass from $x \in \mathbb{R}^d$ to $y \in \mathbb{R}^d$ is $\|x - y\|_2^2$; see [19] for a comprehensive introduction, applications, and related discussion.

Of particular importance to our development is the Wasserstein distance between empirical measures of the same size, which can be written as $\mu_k = \frac{1}{k}\sum_{i=1}^k \delta_{x_i}$ and $\nu_k = \frac{1}{k}\sum_{j=1}^k \delta_{y_j}$ for some $\{x_i\}_{i=1}^k, \{y_j\}_{j=1}^k \subset \mathbb{R}^d$. In this case, the transport problem (3.3) becomes a linear assignment problem with cost $C_{ij} := \|x_i - y_j\|_2^2$:

$$\mathcal{W}_2^2(\mu_k, \nu_k) = \left\{ \begin{array}{rl} \min_{T\in\mathbb{R}^{k\times k}} & \langle T, C\rangle \\ \text{s.t.} & T\mathbb{1} = \mathbb{1}/k \\ & T^\top\mathbb{1} = \mathbb{1}/k \\ & T \geq 0, \end{array} \right. \tag{3.4}$$

where $\mathbb{1}$ denotes the vector of all ones. The constraints of (3.4) form a scaled version of the Birkhoff polytope (set of doubly stochastic matrices), whose vertices define bijections between the x_i's and the y_j's.

There is a probabilistic link between (3.3) and (3.4). For general $\mu, \nu \in \text{Prob}(\mathbb{R}^d)$, we can define an *empirical (plug-in) estimator* of $\mathcal{W}_2^2(\mu, \nu)$ by drawing $x_1, \ldots, x_k \sim \mu$ and $y_1, \ldots, y_k \sim \nu$ and approximating $\mathcal{W}_2^2(\mu, \nu) \approx \mathcal{W}_2^2(\mu_k, \nu_k)$ as in (3.4). As derived in [8, Theorem 2], under straightforward assumptions this approximation converges with rate $k^{-2/d}$ for large k when $d > 4$.

4. k-variance. We can introduce optimal transport into the variance formula (3.2) using the $k = 1$ case of (3.4) by writing $\|x - y\|_2^2 = \mathcal{W}_2^2(\delta_x, \delta_y)$. That is, an equivalent formula to (3.2) is the following: $\text{Var}(X) = \mathbb{E}_{X,Y\sim\mu}[\mathcal{W}_2^2(\delta_X, \delta_Y)]$. This observation immediately suggests a generalization of variance using optimal transport:

Definition 4.1 (k-variance). *Given a probability measure $\mu \in \mathrm{Prob}(\mathbb{R}^d)$ and a parameter $k \in \mathbb{N}$, define* k-variance *as*

$$\mathrm{Var}_k(\mu) := \frac{1}{2} \cdot \rho(k,d) \cdot \mathbb{E}_{\substack{X_1,\dots,X_k \sim \mu \\ Y_1,\dots,Y_k \sim \mu}} \left[\mathcal{W}_2^2 \left(\frac{1}{k} \sum_{i=1}^k \delta_{X_i}, \frac{1}{k} \sum_{i=1}^k \delta_{Y_i} \right) \right], \tag{4.1}$$

where $\rho(k,d)$ is the ambient scaling rate *chosen to account for the rate at which the expectation approaches zero:*

$$\rho(k,d) := \begin{cases} k & \text{if } d = 1 \\ k/\log k & \text{if } d = 2 \\ k^{2/d} & \text{if } d > 2. \end{cases} \tag{4.2}$$

We define

$$\mathrm{Var}_\infty(\mu) := \lim_{k \to \infty} \mathrm{Var}_k(\mu) \tag{4.3}$$

when such a limit exists. For $X \sim \mu$, we will identify $\mathrm{Var}_k(X) := \mathrm{Var}_k(\mu)$.

See section 7 for formulas motivating our choice of $\rho(k,d)$.

Several simple properties of $\mathrm{Var}_k(\cdot)$ in analogy to variance follow from definitions and simple properties of $\mathcal{W}_2$.

Proposition 4.2 (basic properties of $\mathrm{Var}_k(\cdot)$). *We have the following properties for* Var_k*:*

(a) $\mathrm{Var}_1(\mu) = \mathrm{Var}(\mu)$ *for* $\mu \in \mathrm{Prob}(\mathbb{R}^d)$
(b) $\mathrm{Var}_k(\delta_a) = 0$ *for* $a \in \mathbb{R}^d$
(c) $\mathrm{Var}_k(X + a) = \mathrm{Var}_k(X)$ *for* $X \sim \mu \in \mathrm{Prob}(\mathbb{R}^d), a \in \mathbb{R}^d$
(d) $\mathrm{Var}_k(c \cdot X) = c^2 \cdot \mathrm{Var}_k(X)$ *for* $X \sim \mu \in \mathrm{Prob}(\mathbb{R}^d), c \in \mathbb{R}$
(e) $\mathrm{Var}_k(X + \tilde{X}) \geq \mathrm{Var}_k(X) + \mathrm{Var}_k(\tilde{X})$ *for independent* $X \sim \mu \in \mathrm{Prob}(\mathbb{R}^d), \tilde{X} \sim \nu \in \mathrm{Prob}(\mathbb{R}^d)$

Proof. Property (a) is argued above. Properties (b), (c), and (d) follow from simple properties of the cost matrix in (3.4) after substituting (4.1). To prove (e), we resort to the form (3.4). In this case, we can write

$$\mathrm{Var}_k(X + \tilde{X}) := \frac{1}{2} \cdot \rho(k,d) \cdot \mathbb{E}_{\substack{X_1,\dots,X_k \sim \mu; \tilde{X}_1,\dots,\tilde{X}_k \sim \nu \\ Y_1,\dots,Y_k \sim \mu; \tilde{Y}_1,\dots,\tilde{Y}_k \sim \mu}} \left[\mathcal{W}_2^2 \left(\frac{1}{k} \sum_{i=1}^k \delta_{X_i + \tilde{X}_i}, \frac{1}{k} \sum_{i=1}^k \delta_{Y_i + \tilde{Y}_i} \right) \right].$$

The cost matrix of the linear program (3.4) in this expectation has entries

$$C_{ij} = \|X_i + \tilde{X}_i - Y_i - \tilde{Y}_i\|_2^2 = \|X_i - Y_i\|_2^2 + \|\tilde{X}_i - \tilde{Y}_i\|_2^2 + 2(X_i - Y_i) \cdot (\tilde{X}_i - \tilde{Y}_i).$$

Splitting the minimization in (3.4) into three minimizations corresponding to the terms in our expression for C_{ij} above shows

$$\mathrm{Var}_k(X + \tilde{X}) \geq \mathrm{Var}_k(X) + \mathrm{Var}_k(\tilde{X}) + 2\rho(k,d)\mathbb{E}\left[\min_{T \in \mathcal{B}_k} \sum_{ij} [(X_i - Y_i) \cdot (\tilde{X}_i - \tilde{Y}_i)] T_{ij} \right],$$

where $\mathcal{B}_k$ indicates the constraint set in (3.4). By Jensen's inequality,

$$\begin{aligned}\mathrm{Var}_k(X+\tilde{X}) &\geq \mathrm{Var}_k(X)+\mathrm{Var}_k(\tilde{X})+2\rho(k,d)\left[\min_{T\in\mathcal{B}_k}\sum_{ij}\mathbb{E}[(X_i-Y_i)\cdot(\tilde{X}_i-\tilde{Y}_i)]T_{ij}\right]\\ &= \mathrm{Var}_k(X)+\mathrm{Var}_k(\tilde{X}) \text{ by independence, yielding (e).}\end{aligned}$$

∎

In the following sections, we seek to provide intuition for $\mathrm{Var}_k(\cdot)$ in various settings. We organize our discussion around *dimensionality*, starting with one-dimensional measures, proceeding to measures with low-dimensional structures, and then considering the high-dimensional case. We conclude our theoretical discussion with another structured class of measures, those containing clusters of high probability.

5. One-dimensional k-variance.

The k-variance Var_k admits a particularly clean formulation for probability measures over the real numbers $\mathbb{R}$. Here, we derive this alternative interpretation of Var_k, show how it can be used to derive bounds and estimates describing the behavior of one-dimensional k-variance, and give a limiting formula as $k\to\infty$.

5.1. Alternative formula.

In one dimension, the 2-Wasserstein distance $\mathcal{W}_2$ between empirical measures consisting of the same number of points is given by the L^2 distance between the vectors of data points [24]. That is,

$$\mathcal{W}_2\left(\frac{1}{k}\sum_{i=1}^k\delta_{x_i},\frac{1}{k}\sum_{i=1}^k\delta_{y_i}\right)=\sqrt{\sum_{i=1}^k(x_i-y_i)^2}, \tag{5.1}$$

where, without loss of generality, $x_1\leq x_2\leq\cdots\leq x_k$ and $y_1\leq y_2\leq\cdots\leq y_k$.

To incorporate this formula into (4.1), take $X_{(i)}$ to be the ith *order statistic* of $X_1,\ldots,X_k\sim\mu\in\mathrm{Prob}(\mathbb{R})$, a random variable obtained by sorting $\{X_i\}_{i=1}^k$ and taking the ith element of the sorted list; similarly define order statics $Y_{(i)}$ for the samples $\{Y_i\}_{i=1}^k$. Then, for $d=1$ we can write

$$\mathrm{Var}_k(\mu)=\frac{1}{2}\cdot\mathbb{E}_{\substack{X_1,\ldots,X_k\sim\mu\\ Y_1,\ldots,Y_k\sim\mu}}\left[\sum_{i=1}^k(X_{(i)}-Y_{(i)})^2\right]=\sum_{i=1}^k\mathrm{Var}(X_{(i)}) \tag{5.2}$$

by linearity of expectation and by applying (5.1) and (3.2). Hence, in one dimension, the k-variance is exactly the sum of the variances of the order statistics from a random sample of size k. It also represents the variance of the sum of a *ranked-set sample* of size k in a perfect ranking setup (cf. [10, page 263]).

Example 5.1 (uniform distribution). Suppose μ is the uniform distribution on the unit interval. Then, $X_{(i)}\sim\mathrm{Beta}(i,k+1-i)$. Hence,

$$\mathbb{E}(X_{(i)}) = \frac{i}{k+1} \tag{5.3}$$

$$\mathrm{Var}(X_{(i)}) = \frac{i(k+1-i)}{(k+1)^2(k+2)} = \frac{p_i(1-p_i)}{k+2} \tag{5.4}$$

$$\begin{aligned}\mathbb{E}((X_{(i)} - p_i)^4) &= \frac{3i(k-i+1)[2(k+1)^2 + i(k-i+1)(k+5)]}{(k+1)^4(k+2)(k+3)(k+4)} \\ &= \frac{3p_i(1-p_i)[2 + p_i(1-p_i)(k+5)]}{(k+2)(k+3)(k+4)},\end{aligned} \tag{5.5}$$

where $p_i = i/k+1$; we include some of the expressions above to assist in our proof of Proposition 5.5. Substituting (5.4) into our expression for one-dimensional k-variance,

$$\mathrm{Var}_k(\mathrm{Unif}([0,1])) = \sum_{i=1}^{k} \mathrm{Var}(X_{(i)}) = \frac{1}{(k+1)^2(k+2)} \sum_{i=1}^{k} i(k+1-i) = \frac{k}{6(k+1)}. \tag{5.6}$$

This sequence is increasing, and taking a limit as $k \to \infty$ shows $\mathrm{Var}_\infty(\mathrm{Unif}([0,1])) = 1/6$.

Example 5.2 (exponential distribution). Suppose μ is an exponential distribution with parameter λ. Then, we can sample from the order statistics of μ by drawing independent and identically distributed (i.i.d.) exponential variables Z_j with rate 1 and computing the following [21]:

$$X_{(i)} = \frac{1}{\lambda} \sum_{j=1}^{i} \frac{Z_j}{k-j+1}.$$

Substituting the variance of an exponential random variable,

$$\mathrm{Var}(X_{(i)}) = \sum_{j=1}^{i} \left(\frac{1}{\lambda(k-j+1)} \right)^2.$$

This gives the following expression for k-variance:

$$\mathrm{Var}_k(\mathrm{Exp}(\lambda)) = \frac{1}{\lambda^2} \sum_{i=1}^{k} \sum_{j=1}^{i} \frac{1}{(k-j+1)^2} = \frac{H_k}{\lambda^2} \approx \log(k) + \gamma,$$

where H_k is the kth harmonic number and γ is the Euler's constant. Taking $k \to \infty$ shows $\mathrm{Var}_\infty(\mathrm{Exp}(\lambda)) = \infty$.

5.2. Properties of k-variance in 1D. We can immediately derive alternative expressions/bounds for $\mathrm{Var}_k(\cdot)$ in 1D by applying properties of order statistics.

Proposition 5.3 (bounding $\mathrm{Var}_k(\cdot)$ in 1D). *When $d = 1$, we can write*

$$\mathrm{Var}_k(\mu) = k\sigma^2 - \sum_{i=1}^{k} \left(\overline{X}_{(i)} - \overline{X}\right)^2 \leq k\sigma^2. \tag{5.7}$$

Moreover, we can bound

$$\mathrm{Var}_k(\mu) \geq k\sigma^2 - 2\sum_{i<j} \sigma_{(i)}\sigma_{(j)} \cdot \frac{i(k+1-j)}{j(k+1-i)} \tag{5.8}$$

with equality for uniform distributions. In these expressions, $X \sim \mu$, $\sigma^2 = \mathrm{Var}(X)$, *and* $\sigma^2_{(i)} = \mathrm{Var}(X_{(i)})$ *for* $X_1, \ldots, X_k \sim \mu$.

Proof. We can obtain (5.7) by rearranging a sum:

$$\begin{aligned} k\sigma^2 = \sum_{i=1}^{k} \mathbb{E}[(X_i - \overline{X})^2] = \sum_{i=1}^{k} \mathbb{E}[(X_{(i)} - \overline{X})^2] &= \sum_{i=1}^{k} \mathbb{E}[(X_{(i)} - \overline{X}_{(i)} + \overline{X}_{(i)} - \overline{X})^2] \\ &= \mathrm{Var}_k(\mu) + \sum_{i=1}^{k} (\overline{X}_{(i)} - \overline{X})^2. \end{aligned} \tag{5.9}$$

Removing the final term provides inequality (5.7).

To derive (5.8), we rely on a bound on the correlation of order statistics stated in [10, page 74] and references therein. In particular, for $i < j$ they show

$$\mathrm{Corr}(X_{(i)}, X_{(j)}) \leq \frac{i(k+1-j)}{j(k+1-i)}, \tag{5.10}$$

where $\mathrm{Corr}(\cdot,\cdot)$ denotes the correlation of random variables with equality when the parent distribution is uniform. We know $\sum_i X_{(i)} = \sum_i X_i$ given the X_i's are i.i.d. variables with variance σ^2; computing the variance of both sides shows

$$k\sigma^2 = \mathrm{Var}_k(\mu) + 2\sum_{i<j} \mathrm{Cov}(X_{(i)}, X_{(j)}).$$

Substituting (5.10), by definition of correlation we have

$$\mathrm{Var}_k(\mu) = k\sigma^2 - 2\sum_{i<j} \mathrm{Cov}(X_{(i)}, X_{(j)}) \geq k\sigma^2 - 2\sum_{i<j} \sigma_{(i)}\sigma_{(j)} \cdot \frac{i(k+1-j)}{j(k+1-i)}$$

as needed. ∎

Remark 5.4 (approximating Var_k). The expression (5.7) suggests the following means of approximating $\mathrm{Var}_k(\mu)$ for large k:

$$\mathrm{Var}_k(\mu) \approx k\sigma^2 - \sum_{i=1}^{k} (F^{-1}(p_i) - \overline{X})^2, \tag{5.11}$$

where $p_i = {}^{i}/_{k+1}$ and F^{-1} is the quantile function associated to μ. Intuitively, this expression indicates that our index of total local variability is approximately a global variability index minus an index of between-local-group variability.

Another standard approach to working with order statistics involves Taylor series expansions about quantiles of the sampled probability measure. Following this strategy yields a useful approximation to $\mathrm{Var}_k(\cdot)$ as well as a limiting formula under certain assumptions about the distribution function.

Proposition 5.5. *Using the notation of Proposition 5.3, suppose that σ^2 is finite and that μ has a differentiable distribution function $f(x)$ with CDF $F(x)$. Moreover, suppose* (i) $f(x) > 0$, *and* (ii) $f'(x)/[f(x)]^3$ *is bounded on* $F^{-1}((0,1))$. *Then, as $k \to \infty$ we have*

$$\mathrm{Var}_k(\mu) \approx \frac{1}{k+2}\sum_{i=1}^{k} \frac{p_i(1-p_i)}{[f(F^{-1}(p_i))]^2}, \tag{5.12}$$

where $p_i = i/k+1$. As $k \to \infty$, under the assumptions above we have

$$\mathrm{Var}_k(\mu) \to \int_0^1 \frac{u(1-u)}{[f(F^{-1}(u))]^2}\,du = \int_{F^{-1}((0,1))} \frac{F(x)(1-F(x))}{f(x)}\,dx. \tag{5.13}$$

The rate of convergence of $\mathrm{Var}_k(\mu)$ to the limiting integral $\mathrm{Var}_\infty(\mu)$ is of $O(1/\sqrt{k})$.

Proof. Note that $X_{(i)} \overset{d}{=} F^{-1}(U_{(i)})$, where $U_{(i)}$ is the ith order statistic from the standard uniform parent. We begin with a Taylor expansion for $F^{-1}(U_{(i)})$ given in [3]. With $p_i = i/k+1$,

$$F^{-1}(U_{(i)}) = F^{-1}(p_i) + (U_{(i)} - p_i)(F^{-1}(p_i))' + \frac{1}{2}(U_{(i)} - p_i)^2(F^{-1}(V_i))'' \tag{5.14}$$

for some random variable $V_i \in (p_i, U_{(i)})$. Differentiating inverse functions shows

$$(F^{-1}(u))' = \frac{1}{f(F^{-1}(u))} \qquad \text{and} \qquad (F^{-1}(u))'' = -\frac{f'(F^{-1}(u))}{[f(F^{-1}(u))]^3}. \tag{5.15}$$

Substituting into (5.14) and taking variance of both sides shows

$$\begin{aligned}\sigma_{(i)}^2 = \;& \mathrm{Var}(U_{(i)} - p_i)[f(F^{-1}(p_i))]^{-2} + \frac{1}{4}\mathrm{Var}((U_{(i)} - p_i)^2 \cdot (F^{-1}(V_i))'') \\ & + \frac{1}{2}f(F^{-1}(p_i))^{-1}\mathrm{Cov}(U_{(i)} - p_i, (U_{(i)} - p_i)^2(F^{-1}(V_i))''),\end{aligned} \tag{5.16}$$

where $\mathrm{Var}_k(\mu) = \sum_{i=1}^k \sigma_{(i)}^2$.

Applying the identity $\mathrm{Var}[Y] \le \mathbb{E}[Y^2]$, the variance factor in the second term of (5.16) is bounded above by $E((U_{(i)} - p_i)^4[(F^{-1}(V_i))'']^2)$, which in turn is bounded by $M^2 \cdot \mathbb{E}((U_{(i)} - p_i)^4)$, where M is an upper bound for $(F^{-1}(u))''$ for $u \in (0,1)$. From (5.5), we obtain

$$\begin{aligned}\sum_{i=1}^{k} \mathbb{E}((U_{(i)} - p_i)^4) &= \sum_{i=1}^{k} \frac{3p_i(1-p_i)[2 + p_i(1-p_i)(k+5)]}{(k+2)(k+3)(k+4)} \\ &= \frac{6}{(k+3)(k+4)}\sum_{i=1}^{k}\frac{p_i(1-p_i)}{(k+2)} + \frac{3(k+5)}{(k+3)(k+4)}\sum_{i=1}^{k}\frac{p_i^2(1-p_i)^2}{(k+2)} \\ &\approx \frac{6}{k^2}\int_0^1 u(1-u)\,du + \frac{3}{k}\int_0^1 u^2(1-u)^2\,du = \frac{1}{k^2} + \frac{1}{10k}.\end{aligned}$$

Here, the ratios of the first and second terms on the right and left side of the approximation $\approx$ each approach 1 for large k. Thus, we conclude that when summed over i, the second term in (5.16) contributes an amount of size $O(1/k)$.

The covariance term in (5.16) can be bounded as follows:

$$\begin{aligned}
\mathrm{Cov}(\cdots) &\le [\mathrm{Var}(U_{(i)})\mathrm{Var}((U_{(i)}-p_i)^2F^{-1}(V_i))'')]^{1/2}\\
&\le [\mathrm{Var}(U_{(i)})]^{1/2}[\mathbb{E}((U_{(i)}-p_i)^4M^2]^{1/2}\\
&= M\cdot\left[\frac{p_i(1-p_i)}{k+2}\cdot\left[\frac{6p_i(1-p_i)}{(k+2)(k+3)(k+4)}+\frac{3(k+5)p_i^2(1-p_i)^2}{(k+2)(k+3)(k+4)}\right]\right]^{1/2}\\
&= M\cdot\frac{p_i(1-p_i)}{k+2}\cdot\left[\frac{6}{(k+3)(k+4)}+\frac{3(k+5)p_i(1-p_i)}{(k+3)(k+4)}\right]^{1/2}\\
&< M\cdot\frac{p_i(1-p_i)}{k+2}\cdot\frac{C}{\sqrt{k+3}}
\end{aligned}$$

for some constant C ($C=3$ suffices). Summing over i,

$$\begin{aligned}
&\sum_{i=1}^{k}\frac{1}{2}f(F^{-1}(p_i))^{-1}\mathrm{Cov}(U_{(i)}-p_i,(U_{(i)}-p_i)^2(F^{-1}(V_i))'')\\
&\le\frac{MC}{2\sqrt{k+3}}\frac{1}{k+2}\sum_{i=1}^{k}\frac{p_i(1-p_i)}{f(F^{-1}(p_i))}\approx\frac{MC}{2\sqrt{k}}\int_0^1\frac{u(1-u)}{f(F^{-1}(u))}\,du=\frac{MC}{2\sqrt{k}}\int_{F^{-1}((0,1))}F(x)(1-F(x))\,dx,
\end{aligned}$$

where the equality follows upon using the transformation $u=F(x)$. If the support of F is bounded, the integral above is always finite. Even when the support is infinite, the integral is finite whenever the variance or the second moment of F is finite by the comparison test: Finiteness of the variance implies that as $x\to\infty$, $x^2(1-F(x))\to 0$, and as $x\to-\infty$, $x^2F(x)\to 0$. Consequently, the covariance sum is of $O(1/\sqrt{k})$.

Summing the first term in (5.16) over i, using (5.3) we find

$$\begin{aligned}
\sum_{i=1}^{k}[f(F^{-1}(p_i))]^{-2}\mathrm{Var}(U_{(i)}-p_i)&=\frac{1}{k+2}\sum_{i=1}^{k}\frac{p_i(1-p_i)}{[f(F^{-1}(p_i))]^2},\ \text{validating (5.12)}\\
&\approx\int_0^1\frac{u(1-u)}{[f(F^{-1}(u))]^2}\,du.
\end{aligned}$$

The transformation $u=F(x)$ shows that

$$\int_0^1\frac{u(1-u)}{[f(F^{-1}(u))]^2}\,du=\int_{F^{-1}((0,1))}\frac{F(x)(1-F(x))}{f(x)}\,dx$$

as desired. ∎

Remark 5.6 (relationship to [6]). In [6], Bobkov and Ledoux provide a comprehensive discussion of one-dimensional optimal transport from samples in an attempt to understand convergence of empirical approximations to a measure in the Wasserstein metric. Their analysis

focuses on the "one-sided" convergence of an empirical approximation to a true measure, while k-variance is based on the Wasserstein distance between two different empirical approximations.

That said, along the way their discussion does make some similar observations to our discussion above. For instance, their Theorem 4.3 shows the same link to order statistics as our (5.2). The "J_2 functional" defined in their (5.3) is the right-hand side of (5.13); in our notation, their Theorem 5.1 (and, in particular, their Corollary B.6) implies a bound

$$\mathrm{Var}_k(\mu) \leq \frac{k}{k+1} J_2(\mu). \tag{5.17}$$

This establishes half of our equality in (5.13). Their results show $\limsup_{k\to\infty} \mathrm{Var}_k(\mu) \leq J_2(\mu)$, while we are able to show under stronger assumptions that $\lim_{k\to\infty} \mathrm{Var}_k(\mu) = J_2(\mu)$.

Example 5.7 (uniform distribution, continued). Continuing Example 5.1, we can apply (5.13) to compute

$$\mathrm{Var}_\infty(\mathrm{Unif}([0,1])) = \int_0^1 \frac{x(1-x)}{1}\,dx = \frac{1}{6}. \tag{5.18}$$

As expected, this expression agrees with (5.6) as $k \to \infty$.

Example 5.8 (Weibull distribution with shape parameter α). For this distribution, $F(x) = 1 - \exp\{-x^\alpha\}$ and $f(x) = \alpha x^{\alpha-1}\exp\{-x^\alpha\}$ for $x > 0$, with shape parameter $\alpha > 0$. As $x \to 0^+$,

$$\frac{F(x)(1-F(x))}{f(x)} = \frac{1-\exp\{-x^\alpha\}}{\alpha x^{\alpha-1}} \approx \frac{x^\alpha}{\alpha x^{\alpha-1}} = \frac{x}{\alpha},$$

and consequently the integral (5.13) is always convergent at the lower limit of integration. As $x \to \infty$,

$$\frac{F(x)(1-F(x))}{f(x)} = \frac{1-\exp\{-x^\alpha\}}{\alpha x^{\alpha-1}} \approx \frac{1}{\alpha x^{\alpha-1}},$$

and hence the integral (5.13) is convergent at the upper limit if and only if $\alpha > 2$.

Now, for $\alpha > 2$, (5.13) implies

$$\begin{aligned}\mathrm{Var}_\infty(\mathrm{Weib}(\alpha)) &= \int_0^\infty \frac{1-\exp\{-x^\alpha\}}{\alpha x^{\alpha-1}}\,dx = \frac{1}{\alpha^2}\int_0^\infty (1-e^{-y})y^{2/\alpha-2}\,dy \\ &= \frac{1}{\alpha^2(2/\alpha-1)}\int_0^\infty (1-e^{-y})\,d(y^{2/\alpha-1}).\end{aligned}$$

Upon integration by parts we see that

$$\int_0^\infty (1-e^{-y})\,d(y^{2/\alpha-1}) = \Big[(1-e^{-y})y^{2/\alpha-1}\Big]_0^\infty - \int_0^\infty e^{-y}y^{2/\alpha-1}\,dy.$$

For $\alpha > 2$, the first term yields 0 at both upper and lower limits, and the second term equals $\Gamma(2/\alpha)$. Thus,

$$\mathrm{Var}_\infty(\mathrm{Weib}(\alpha)) = \begin{cases} \frac{\Gamma(2/\alpha)}{\alpha(\alpha-2)} & \text{when } \alpha > 2 \\ \infty & \text{when } \alpha \leq 2. \end{cases}$$

Example 5.9 (Tukey's symmetric λ distribution). This distribution is defined by its quantile function $F^{-1}(u)$, given by

$$F^{-1}(u) = \begin{cases} \frac{1}{\lambda}(u^\lambda - (1-u)^\lambda) & \text{when } \lambda \neq 0 \\ \log(u/(1-u)) & \text{when } \lambda = 0 \end{cases} \tag{5.19}$$

for $u \in [0,1]$ and $\lambda \in \mathbb{R}$. When $\lambda = 0$, we obtain the standard logistic distribution.

When $\lambda \neq 0$, the quantile density function $(F^{-1}(u))'$ is given by $u^{\lambda-1} + (1-u)^{\lambda-1}$, and

$$(F^{-1}(u))'' = (\lambda - 1)[u^{\lambda-2} - (1-u)^{\lambda-2}]$$

is bounded if and only if $\lambda \geq 2$. Hence, we satisfy the sufficient conditions needed for Proposition 5.5. Thus for $\lambda \geq 2$,

$$\begin{aligned} \mathrm{Var}_\infty(\mathrm{Tukey}(\lambda)) &= \int_0^1 u(1-u)(F^{-1}(u))'\,du = \int_0^1 u(1-u)[u^{\lambda-1} + (1-u)^{\lambda-1}]\,du \\ &= 2 \cdot \mathrm{Beta}(\lambda+1, 2) = \frac{2}{(\lambda+1)(\lambda+2)}. \end{aligned} \tag{5.20}$$

The integral on the right is finite whenever $\lambda > -1$, and the expression holds for $\lambda = 0$. For $\lambda \in (-1, 2)$, we can only say that $\limsup_{k\to\infty} \mathrm{Var}_k(\mu)$ is bounded above by the right-hand side.

6. Low-dimensional measures. Having worked out the case of one-dimensional measures, we now consider measures that have low-dimensional structure but are embedded in a higher-dimensional space. Specifically, consider the following definition.

Definition 6.1 (ε-fattening and ε-covering number, [23, 25]). *For any compact $X \subset \mathbb{R}^d$ and $S \subseteq X$, the ε-fattening of S is $S_\varepsilon := \{y : D(y, S) \leq \varepsilon\}$, where D denotes the Euclidean distance. The ε-covering number $\mathcal{N}_\varepsilon(S)$ of S is the minimum m such that there exist m points $x_1, \ldots, x_m \in \mathbb{R}^d$ with $S \subseteq \bigcup_i B_\varepsilon(X_i)$.*

We borrow a recent bound on empirical transport, specialized to $\mathcal{W}_2$.

Proposition 6.2 ([25, Proposition 15]). *Suppose $\mathrm{supp}(\mu) \subseteq S_\varepsilon$ for some $\varepsilon > 0$, where S satisfies $\mathcal{N}_{\varepsilon'}(S) \leq (3\varepsilon')^{-d'}$ for all $\varepsilon' \leq 1/27$ and some $d' > 4$. Then, for all $k \leq (3\varepsilon)^{-d'}$, we have $\mathbb{E}[\mathcal{W}_2^2(\mu, \hat{\mu}_k)] \leq C_1 \cdot k^{-2/d'}$, where $C_1 = 27^2(2 + 1/(3^{d'/2-2} - 1))$ and μ_k denotes the k-point empirical measure.*

Translating this to our setting using the triangle inequality, we get the following.

Proposition 6.3 ($\mathrm{Var}_k(\cdot)$ for low-dimensional distributions). *Suppose $\mathrm{supp}(\mu) \subseteq S_\varepsilon$ for some $\varepsilon > 0$, where S satisfies $\mathcal{N}_{\varepsilon'}(S) \leq (3\varepsilon')^{-d'}$ for all $\varepsilon' \leq 1/27$ and some $d' > 4$. Then, for all $k \leq (3\varepsilon)^{-d'}$, we have $\mathrm{Var}_k(\mu) \leq C_1 \cdot k^{2/d - 2/d'}$.*

Unsurprisingly, the proposition above shows that if we measure the d-dimensional k-variance of an intrinsically d'-dimensional measure, at least when $4 < d' < d$, we have $\mathrm{Var}_k(\mu) \to 0$ as $k \to \infty$. As a special case, we see that empirical measures have k-variance tending to zero for higher-dimensional measures. Interestingly, this is not the case in low dimensions, as we can see in the following example.

Example 6.4 (two-point empirical measures). Take $\mu = (\delta_{-0.5e_1}+\delta_{0.5e_1})/2$ where $e_1 = (1,0,\dots,0) \in \mathbb{R}^d$. In this case, $\mathcal{W}_2$ between two k-samples from μ counts the imbalance in the number of -0.5 versus 0.5 samples between the two draws. Hence, $\mathrm{Var}_k(\mu)$ is the expected absolute difference $|A-B|$ of two binomial variables $A, B \sim B(k, 1/2)$, scaled by $\rho(k,d)/2k$. From [20, equation (2.9)], for binomially distributed variables $X_1, X_2 \sim B(k,p)$ we have

$$\mathbb{E}(|X_1 - X_2|) = 2kp(1-p) \cdot {}_2F_1\left(1-k, \frac{1}{2}; 2; 4p(1-p)\right),$$

where ${}_2F_1$ is Gauss' hypergeometric function. Substituting $p = \frac{1}{2}$ shows

$$\mathbb{E}(|X_1 - X_2|) = \frac{k\Gamma(2)}{2\Gamma(\frac{1}{2})\Gamma(\frac{3}{2})}\int_0^1 t^{-\frac{1}{2}}(1-t)^{k-1}\,dt = \frac{k\Gamma(k+\frac{1}{2})}{2\Gamma(\frac{3}{2})\Gamma(k+1)} = \binom{2k}{k}\cdot\frac{k}{2^{2k}}.$$

Hence,

$$\mathrm{Var}_k(\mu) = \frac{\rho(k,d)}{2^{2k+1}}\binom{2k}{k} \approx \frac{\rho(k,d)}{2\sqrt{\pi k}} = \frac{1}{2\sqrt{\pi}}\cdot\begin{cases}\sqrt{k} & \text{if } d=1\\ (\log k)^{-1}\cdot\sqrt{k} & \text{if } d=2\\ k^{2/d-1/2} & \text{if } d\geq 3\end{cases}$$

by Stirling's approximation. So, $\mathrm{Var}_k(\mu)$ diverges for $d \leq 3$, converges to $1/2\sqrt{\pi}$ for $d=4$, and converges to 0 for $d \geq 5$.

While Proposition 6.3 provides an upper bound on the k-variance, a lower bound can also be obtained that yields the same dependence on intrinsic dimension. Using [25, Theorem 1], we have the following.

Proposition 6.5. *Let the* (ε,τ) *covering number be* $\mathcal{N}_\varepsilon(\mu,\tau) = \inf\{\mathcal{N}_\varepsilon(S) : \mu(S) \geq 1-\tau\}$. *Define* $d_* = \lim_{\tau\to 0}\liminf_{\varepsilon\to 0}\frac{\log \mathcal{N}_\varepsilon(\mu,\tau)}{-\log\varepsilon}$, *called the* lower Wasserstein dimension *by* [25]. *Then,*

$$\mathrm{Var}_k(\mu) \gtrsim k^{-2/t}$$

for any $t < d_*$.

This proposition establishes that the rate of change of the k-variance as k varies is directly related to the intrinsic dimensionality of μ.

7. Higher-dimensional measures.

A surprising result of our experiments detailed in section 10 is that one-dimensional k-variance seems to have totally different behavior than k-variance for measures on $\mathbb{R}^d$ for large d. While we cannot provide as complete a story as section 5 for the one-dimensional case, some results in the theory of random Euclidean matching are directly relevant to our construction and can provide some insight into the behavior of $\mathrm{Var}_k(\cdot)$.

Example 7.1 (unit cube). Suppose $\mu = \mathrm{Unif}([0,1]^d)$. Then, for large k we have the following formula [15, equation (1.1)]:

$$\mathbb{E}_{\substack{X_1,\dots,X_k\sim\mu\\ Y_1,\dots,Y_k\sim\mu}}\left[\mathcal{W}_2^2\left(\frac{1}{k}\sum_{i=1}^k\delta_{X_i}, \frac{1}{k}\sum_{i=1}^k\delta_{Y_i}\right)\right] \approx \begin{cases}k^{-1} & \text{if } d=1\\ (\log k)/k & \text{if } d=2\\ k^{-2/d} & \text{if } d\geq 3.\end{cases} \tag{7.1}$$

These formulas motivate our choice of scaling factors $\rho(k,d)$ in Definition 4.1. [8, Theorem 2] observes similar rates for $d > 4$ for general measures with support in the unit ball, but their upper bound decays more slowly in k than (7.1) for $d \leq 4$.

Example 7.2 (unit square). [5] predicts a similar $(\log k)/k$ rate for measures with positive density on the unit square $[0,1]^2$. Specifically for $\mu = \mathrm{Unif}([0,1]^2)$, we can obtain the following limit [1, Theorem 1.1]:

$$\lim_{k\to\infty} \frac{k}{\log k} \mathbb{E}_{\substack{X_1,\dots,X_k\sim\mu \\ Y_1,\dots,Y_k\sim\mu}} \left[\mathcal{W}_2^2 \left(\frac{1}{k}\sum_{i=1}^k \delta_{X_i}, \frac{1}{k}\sum_{i=1}^k \delta_{Y_i} \right) \right] = \frac{1}{2\pi}.$$

Hence, we have $\mathrm{Var}_\infty([0,1]^2) = 1/2\pi$.

8. Clustered measures. In section 6, we analyzed the behavior of k-variance in the presence of an approximately low-dimensional structure of dimension d'. In this section, we provide an alternative formulation for approximately zero-dimensional structures ($d' = 0$), i.e., measures that are concentrated around finite sets of points or cluster centers. In particular, we analyze approximately clustered measures, where the definition of "approximate" is motivated from a clustering perspective.

We consider the following definitions, again from [25] similar to our discussion in section 6, which provide two ways of identifying clusterable structure in probability measures.

Definition 8.1 ((m,σ^2)-Gaussian mixture). *A distribution μ is an* (m,σ^2)-Gaussian mixture *if it is a mixture of m Gaussian distributions in $\mathbb{R}^d$, and the trace of the covariance matrix of each mixture component is bounded above by σ^2.*

Definition 8.2 (clusterable measure). *A distribution μ is* (m,Δ)-clusterable *if* $\mathrm{supp}(\mu)$ *lies in the union of m balls of radius at most Δ.*

The following proposition from [25] directly suggests a k-variance bound.

Proposition 8.3 ([25, Propositions 13 and 14]). *If μ is a (m,σ^2)-Gaussian mixture and $\log 1/\sigma \geq 25/8$, then for all $k \leq m(32\sigma^2 \log 1/\sigma)^{-2}$,*

$$\mathbb{E}[\mathcal{W}_2^2(\mu, \hat{\mu}_k)] \leq 84\sqrt{m/k}, \tag{8.1}$$

where $\hat{\mu}_k$ is the empirical measure obtained by drawing k samples. The same rate holds for (m,Δ)-clusterable distributions for all $k \leq m(2\Delta)^{-4}$.

The $k^{-1/2}$ rate is tight as it corresponds to the parametric rate.

Application of the triangle inequality to Proposition 8.3 immediately yields the following.

Proposition 8.4 (Var_k for clustered distributions). *Suppose $d > 4$. For the (m,σ^2)-Gaussian mixture case with $k \leq m(32\sigma^2 \log 1/\sigma)^{-2}$,*

$$\mathrm{Var}_k(\mu) \leq \frac{168 m^{1/2}}{k^{1/2-2/d}}.$$

For the (m,Δ)-clusterable case with $k \leq m(2\Delta)^{-4}$,

$$\mathrm{Var}_k(\mu) \leq \frac{168 m^{1/2}}{k^{1/2-2/d}}.$$

Roughly, this proposition shows that as d increases and k satisfies the inequality, clustered distributions have increasingly small $\mathrm{Var}_k(\cdot)$, though the rate of increase slows rapidly once d gets beyond ~ 10.

9. Variance of empirical k-variance. We conclude our mathematical discussion by considering the problem of how to *compute* k-variance in practice. There exists an extremely simple *empirical estimator* directly motivated by the expectation in (4.1): simply draw $2k$ samples, solve the linear program (3.4), and use the resulting value. A simple implementation of this algorithm takes roughly $O(k^2d + k^3)$ time, accounting for the time taken to compute the pairwise cost matrix as well as solving the transport linear program (our implementation uses [14]). Here we bound the variance of this estimator, roughly showing that fewer trials need to be averaged to compute k in large dimension.

In detail, we consider the empirical estimator built from n trials:

$$\widehat{\mathrm{Var}}_k(\mu) := \frac{\rho(k,d)}{2n}\sum_{j=1}^{n}\mathcal{W}_2^2(\hat{\mu}_k^j, \hat{\mu}_k'^j),$$

where $\hat{\mu}_k^j$, $\hat{\mu}_k'^j$ are independent empirical measures formed from k i.i.d. samples as in (4.1).

The following theorem helps characterize the variance of our estimator above.

Theorem 9.1 (empirical variance). *Suppose $\mu \in \mathrm{Prob}(\mathbb{R}^d)$ has support in a set of radius R. For each $j \in \{1,\dots,n\}$, take $\hat{\mu}_k^j, \hat{\mu}_k'^j$ to be independent empirical measures each constructed from k i.i.d. samples from μ ($X^{k,j} := (X_1^j,\dots,X_k^j)$ for $\hat{\mu}_k$ and $Y^{k,j} := (Y_1^j,\dots,Y_k^j)$ for $\hat{\mu}_k'$). Then,*

$$\mathbb{P}\left(\left|\widehat{\mathrm{Var}}_k(\mu) - \mathrm{Var}_k(\mu)\right| \geq \rho(k,d)R^2\sqrt{\frac{\log(kn)}{kn}}\right) \leq \frac{2}{k^2n^2}. \tag{9.1}$$

Proof. We use McDiarmid's inequality.

Lemma 9.2 (McDiarmid's inequality [17]). *Let $X^m := (X_1,\dots,X_m)$ be an m-tuple of $\mathcal{X}$-valued independent random variables. Suppose $g : \mathcal{X}^m \to \mathbb{R}$ is a map that for any $i = 1,\dots,m$ and $x_1,\dots,x_m,x_i' \in \mathcal{X}$ satisfies*

$$\left|g(x_1,\dots,x_m) - g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_m)\right| \leq c_i \tag{9.2}$$

for some nonnegative $\{c_i\}_{i=1}^m$. Then for any $t > 0$

$$\mathbb{P}\Big(g(X_1,\dots,X_m) - \mathbb{E}g(X_1,\dots,X_m) \geq t\Big) \leq e^{-\frac{2t^2}{\sum_{i=1}^m c_i^2}}, \tag{9.3a}$$

$$\mathbb{P}\Big(\left|g(X_1,\dots,X_m) - \mathbb{E}g(X_1\dots,X_m)\right| \geq t\Big) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^m c_i^2}}. \tag{9.3b}$$

Consider $\frac{1}{n}\sum_{j=1}^n \mathcal{W}_2^2(\hat{\mu}_k^j, \hat{\mu}_k'^j)$ as a function of the nk independent samples from which it is computed, each sample being a pair (x_i^j, y_i^j). Using Kantorovich–Rubinstein duality, we have the general formula

$$\mathcal{W}_2^2(P,Q) = \sup_{(f,g)\in\Phi} \mathbb{E}_P[f] + \mathbb{E}_Q[g],$$

where $\Phi = \{(f,g) \in L^1(P) \times L^1(Q) : f(x) + g(y) \le \|x-y\|^2\}$. In our case, separately for each j, we can write

$$\mathcal{W}_2^2(\hat{\mu}_k^j, \hat{\mu}_k'^j) = \mathcal{W}_2^2\left(\frac{1}{k}\sum_{\ell=1}^k \delta_{x_\ell^j}, \frac{1}{k}\sum_{\ell=1}^k \delta_{y_\ell^j}\right) = \sup_{(f,g)\in\Phi} \frac{1}{k}\sum_{\ell=1}^k (f(x_\ell^j) + g(y_\ell^j)).$$

Recall that the (x_ℓ^j, y_ℓ^j) are independent across ℓ and j. Consider replacing one of the elements (x_i^j, y_i^j) with some $(x_i'^j, y_i'^j)$, forming $\bar{\mu}_k^j$ and $\bar{\mu}_k'^j$. Since the (x_ℓ^j, y_ℓ^j) are identically distributed, by symmetry we can set $i = 1$. We thus bound

$$\begin{aligned}\mathcal{W}_2^2(\hat{\mu}_k^j, \hat{\mu}_k'^j) - \mathcal{W}_2^2(\bar{\mu}_k^j, \bar{\mu}_k'^j) = &\sup_{(f,g)\in\Phi} \frac{1}{k}\left((f(x_1^j) + g(y_1^j)) + \sum_{\ell=2}^k (f(x_\ell^j) + g(y_\ell^j))\right) \\ &- \sup_{(f,g)\in\Phi} \frac{1}{k}\left((f(x_1'^j) + g(y_1'^j)) + \sum_{\ell=2}^k (f(x_\ell^j) + g(y_\ell^j))\right) \le \frac{2R^2}{k},\end{aligned}$$

where we have assumed the space is bounded with radius R and used the definition of Φ and [24, Remark 1.13].

Hence by symmetry and scaling by $\frac{\rho(k,d)}{2n}$ as in the expression in the theorem we have

$$\left|\frac{\rho(k,d)}{2n}\mathcal{W}_2^2(\hat{\mu}_k^j, \hat{\mu}_k'^j) - \frac{\rho(k,d)}{2n}\mathcal{W}_2^2(\bar{\mu}_k^j, \bar{\mu}_k'^j)\right| \le \frac{\rho(k,d)R^2}{kn},$$

satisfying the condition (9.2) for McDiarmid's inequality for each of the nk random variables (x_i^j, y_i^j). Therefore, for any $t > 0$, by (9.3b) we have

$$\mathbb{P}\left(\left|\frac{\rho(k,d)}{2n}\sum_{j=1}^n (\mathcal{W}_2^2(\hat{\mu}_k^j, \hat{\mu}_k'^j) - \mathbb{E}\mathcal{W}_2^2(\hat{\mu}_k, \hat{\mu}_k'))\right| \ge t\right) \le 2e^{-\frac{2knt^2}{R^4\rho(k,d)^2}}.$$

Setting $t = \rho(k,d)R^2\sqrt{\frac{\log(kn)}{kn}}$ then yields

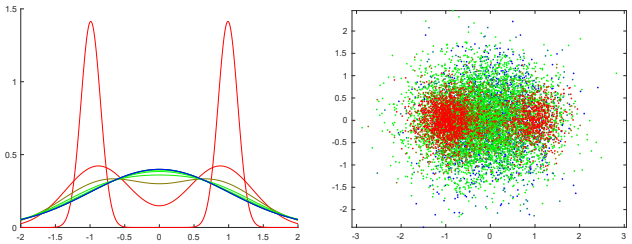
$$\mathbb{P}\left(\left|\frac{\rho(k,d)}{2n}\sum_{j=1}^n \left(\mathcal{W}_2^2(\hat{\mu}_k^j, \hat{\mu}_k'^j) - \mathbb{E}\mathcal{W}_2^2(\hat{\mu}_k^j, \hat{\mu}_k'^j)\right)\right| \ge \rho(k,d)R^2\sqrt{\frac{\log(kn)}{kn}}\right) \le \frac{2}{k^2n^2}.$$

Recalling the definition of $\mathrm{Var}_k(\mu)$, the theorem results. ∎

Remark 9.3 (interpretation of Theorem 9.1). In words, as dimension d and size k increase, we need a smaller number n of independent trials n of k-samples to estimate k-variance accurately. Eventually, even choosing $n = 1$ suffices.

Remark 9.4 (alternative forms for Theorem 9.1). Theorem 9.1 is written in terms of the number n of sets of k replicates. We can rewrite it in terms of k and $m = kn$, i.e., when a total of m samples are available and one is choosing a k to partition them. We have for $d > 2$

$$\mathbb{P}\left(\left|\widehat{\mathrm{Var}}_k(\mu) - \mathrm{Var}_k(\mu)\right| \ge \rho(k,d)R^2\sqrt{\frac{\log m}{m}}\right) \le \frac{2}{m^2}.$$



(a) $d = 1$ *Distribution functions* (b) $d = 2$ *Samples*

Figure 1. *Example distributions for the experiments in subsection 10.1. For $d = 1$ we show the density functions corresponding to the different colors, and for $d = 2$ we show samples drawn from the various measures.*

This in a sense reverses the trade-off, with finer divisions of the m available samples (smaller k) reducing the overall variance (only slightly for large d, however).

10. Experiments.

In this section, we provide some simple experiments demonstrating the behavior of $\mathrm{Var}_k(\cdot)$ and suggesting how it might be used to understand properties of distributions and datasets that are not well captured by variance alone.

10.1. Gaussian mixtures.

We begin with a synthetic experiment illustrating the behavior of k-variance in different dimensions and in the presence of multimodality. In our experiments, we consider mixtures $\mathcal{G}_x := 0.5\mathcal{N}(-x\cdot e_1, \sigma I_{d\times d}) + 0.5\mathcal{N}(x\cdot e_1, \sigma I_{d\times d})$ of two isotropic Gaussians, where $e_1 \in \mathbb{R}^d$ is the first standard basis vector in $\mathbb{R}^d$. We choose $\sigma(x)$ so that $\mathrm{Var}_1(\mathcal{G}_x) = 1$; note $\sigma(x)$ decreases as $|x|$ increases, leading to bimodal/approximately clustered distributions. See Figure 1 for examples in dimensions 1 and 2.

Figure 2 shows k-variance of $\mathcal{G}_x$ as a function of k (horizontal axis) and x (color) in different ambient dimensions d. We use the empirical estimator of k-variance averaged over 10,000 trials for each point in the plot. We can make a number of observations based on this experiment:

- For $d \geq 3$, the k-variance is smaller for clustered distributions (red) than unimodal Gaussians (blue) with identical (1-)variance.
- The $d \in \{1, 2\}$ cases exhibit unique, nonmonotonic behavior. For instance, when $d = 1$, k-variance is highest for the sharply bimodal distributions (red), then decreases for wide-and-flat distributions (dark red/green), and then increases again for Gaussians (blue).
- For larger dimension d, the curves look smoother. This is a byproduct of the results in section 9, which predict that the empirical estimator of $\mathrm{Var}_k(\cdot)$ has lower variance in high dimension given a fixed number of samples.

10.2. Low-dimensional measures.

Now, we consider the case explored in section 6, in which our probability measure is embedded in a low-dimensional slice of the ambient space $\mathbb{R}^d$. When d is sufficiently large, Proposition 6.3 predicts that the k-variance for such a measure will decay to zero at a rate determined by the intrinsic dimensionality of the measure.

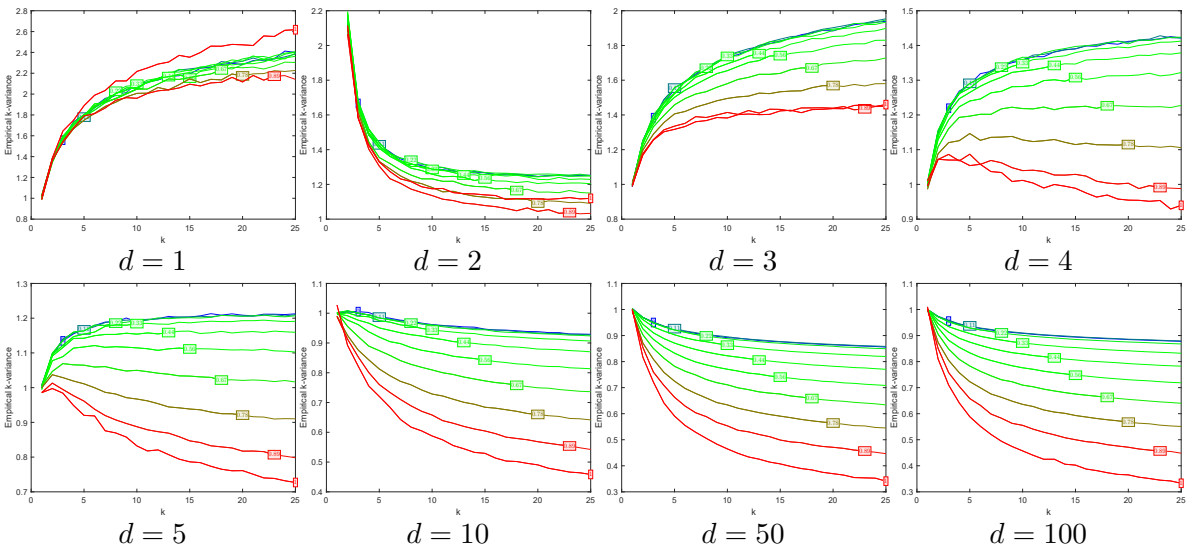


Figure 2. *k-variance experiments with Gaussian mixture models (see subsection 10.1) in increasing dimension. As predicted, k-variance follows similar patterns in $d \geq 3$ and is lower for clustered distributions, but for $d \in \{1, 2\}$ the behavior is different. Colors range from bimodal mixtures of low-variance Gaussians (red) to unimodal Gaussian measures (blue); each curve is labeled with the value of x, the distance of the mixture component means to the origin.*

As an initial experiment, we consider Gaussian measures in dimension $d = 1{,}000$ supported on a d'-dimensional hyperplane, where d' varies from 1 to d. Here, we create the d'-dimensional measure by creating a Gaussian with covariance

$$\Sigma_{d'} = \operatorname{diag}(\underbrace{1/d', \ldots, 1/d'}_{d' \text{ slots}}, \underbrace{0, \ldots, 0}_{d-d' \text{ slots}}).$$

Here, the $1/d'$ entries ensure that the measure has variance 1. Figure 3 plots the k-variance of the d'-dimensional measures on a logarithmic scale; we use the empirical estimator of k-variance averaged over 1,000 trials.

As predicted by Proposition 6.3, the slopes of the fit lines in Figure 3 cleanly correlated with intrinsic dimensionality. As d' increases, the lines also become smoother, again a byproduct of the variance bounds in section 9. This is a happy coincidence: We are able to distinguish the slopes of the different lines for large d'—even though they are close in value—because we can estimate $\operatorname{Var}_k(\cdot)$ more accurately in this regime.

Figure 4 shows a similar experiment to Figure 3, but now the data lie on the sphere $S^{d'-1}$ embedded in $\mathbb{R}^d$; we sample uniformly from $S^{d'-1}$ by normalizing the samples from the previous experiment to unit length. Once again the trendlines strongly fit the power law we expect to see, but the slopes are now less negative compared to Figure 3 since the intrinsic dimensionality has decreased by 1. In particular, $d' = 1$ corresponds to a zero-dimensional sphere S^0, i.e., two points on the real line. Since S^0 is thus a discrete dataset, this explains the approximately $-1/2$ slope in the log-log plot, corresponding to the $k^{-1/2}$ decay indicated in section 8.

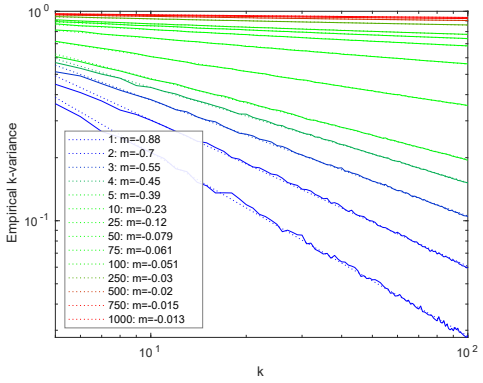


Figure 3. *k-variance of measures supported on low-dimensional hyperplanes in $\mathbb{R}^{1000}$. Each curve corresponds to a different intrinsic dimensionality d' marked in the legend; m is the slope of the best-fit curve in the log-log plot. Note the correlation between m and the intrinsic dimensionality of the measure.*

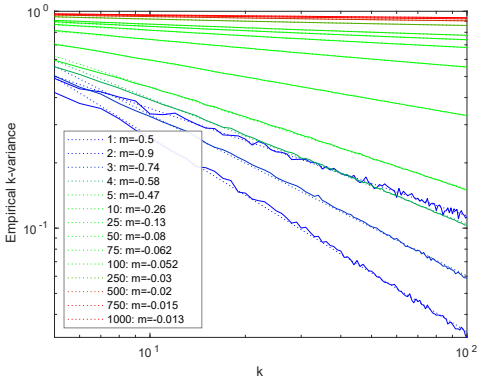


Figure 4. *Similar experiment to Figure 3, now with points in the unit sphere $S^{d'-1}$ embedded in $\mathbb{R}^d$.*

10.3. Maximizing k-variance to recover cluster subspace. In this section, we consider finding linear projections that maximize k-covariance. While for 1-variance, this problem reduces to principal component analysis (PCA), for high k, the theoretically expected behavior is that maximizing k-variance will increasingly prefer the directions of large within-cluster variation to directions of variation between cluster centers. This type of analysis could be useful for tasks like domain adaptation, where it is important to find representations such that the data distribution is invariant to changing environments [2], and fairness, where it is important to treat clusters like protected groups identically [27].

To optimize the k-variance over the set of low-dimensional projections, we use the subspace robust Wasserstein method in [18]; experimental results are shown in Figure 5. The scatter plots show the first two dimensions of 20-dimensional synthetic datasets consisting of 5 Gaussian clusters with centers sampled from a random (Wishart-distributed) five-dimensional Gaussian distribution. We consider the following cases:

- homogeneous covariance: Wishart-sampled five-dimensional cluster covariance shared across clusters, and
- heterogeneous covariance: Each cluster's covariance is sampled independently, and all cluster covariances are restricted to a random seven-dimensional subspace.

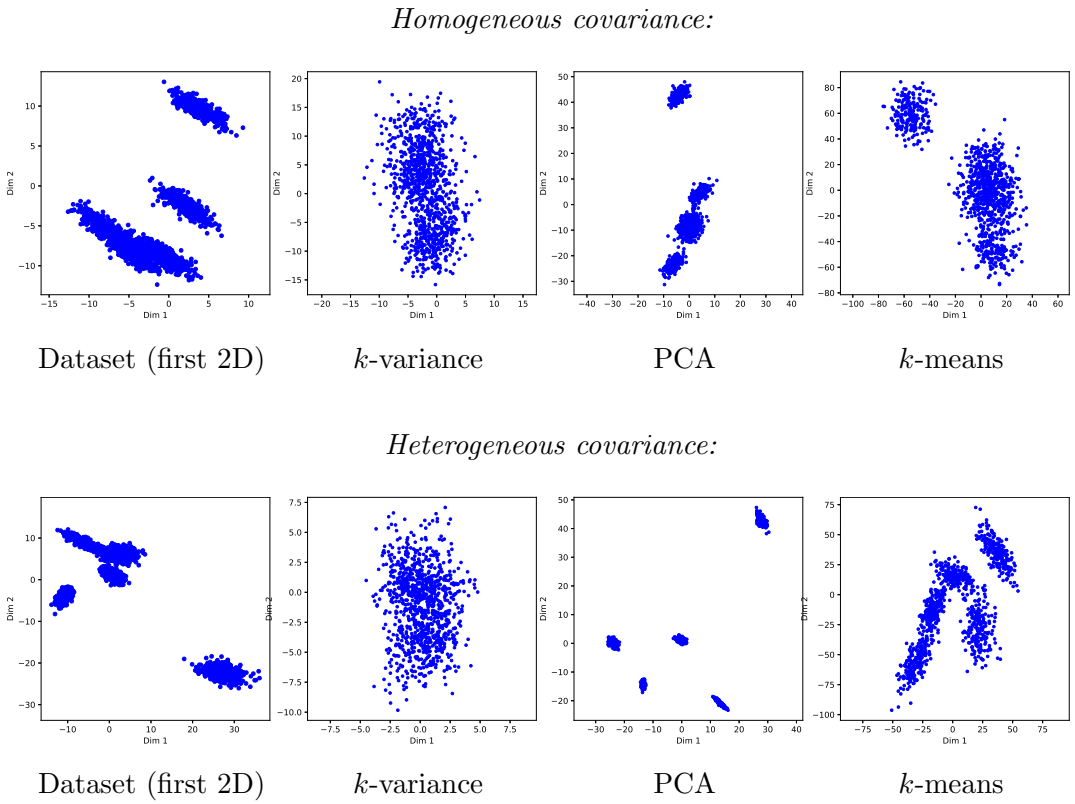
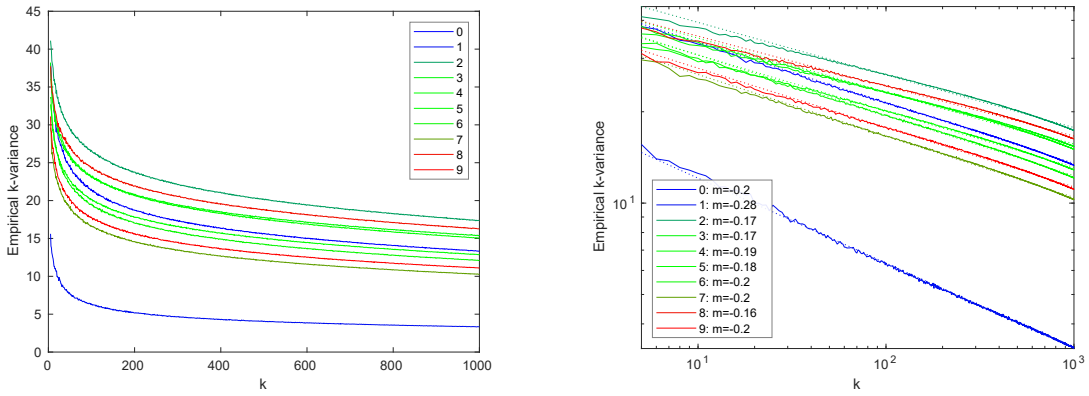


Figure 5. *Maximizing the k-variance of projections of clustered datasets. Two 20-dimensional datasets are considered ("Homogeneous covariance" and "Heterogeneous covariance") with scatter plots shown of the first two dimensions. The projected k-variance ($k = 500$), 1-variance (via PCA), and within-cluster 1-variance (via PCA on the output of k-means), respectively, are all maximized over the set of five-dimensional projections, and the first 2 dimensions of the resulting projection shown in a scatter plot. As expected, PCA is drawn to the larger-variance cluster center distribution, whereas k-variance successfully recovers a projection that maximizes the within-cluster variation, in this case projecting away the cluster center distribution and leaving a single cluster. k-means is an alternative approach to achieve the same effect, but it fails due to the eccentricities of the cluster covariances.*

Since the ambient dimension is 20, the cluster center distribution subspace will not be similar to—but not fully orthogonal to—the cluster covariance subspace. We compare five-dimensional projections found by maximizing k-variance ($k = 500$), 1-variance (via PCA), and within-cluster 1-variance via PCA on the output of k-means. As expected, PCA is drawn to the larger-variance cluster center distribution, whereas k-variance successfully recovers a projection that maximizes the within-cluster variation, in this case projecting away the cluster center distribution and leaving a single cluster. k-means is an alternative approach to achieve the same effect but fails due to the eccentricity of the cluster covariances.

This experiment (a) confirms that k-variance is dominated by the within-cluster variation as predicted and (b) demonstrates as a proof-of-concept the possibility of using the k-variance as an objective function in machine learning and data analysis. While here we *maximized* the k-variance to extract the shapes of the clusters, in alternative instances a useful approach



(a) Linear (b) Log-log

Figure 6. *k-variance of each MNIST digit from 0 to 9 in* (a) *linear and* (b) *log-log scale. Each digit is considered as a 28^2-dimensional vector ($d = 784$); there are approximately $6,000$ images per digit.*

may also be to *minimize* the k-variance subject to a 1-variance constraint, which could create a *clustered* representation for machine learning classifiers; this intuition is confirmed by the follow-up work [9] showing that small k-variance improves classifier generalization.

10.4. Digits. Figure 6 plots approximate k-variance for the MNIST dataset of handwritten digits [16], separated by digits. We use the stochastic estimator for k-variance from section 9, where sampling from the distribution of handwritten digits is simulated by a bootstrapped strategy of sampling from the dataset with replacement. Our distributions in this case are over $\mathbb{R}^{784}$, representing 28×28 images. Given the high ambient dimension and the well-documented observation from past work that the MNIST digits roughly lie on low-dimensional submanifolds of $\mathbb{R}^{784}$, we expect k-variance to diminish to zero in this experiment. Hence the relevant measurement is the rate at which this decay occurs.

Beyond varying amounts of variance between different digits ($k = 1$), our experiments also reveal that the digit "1" has k-variance decaying in k roughly $1.5\times$ faster than the other digits. This provides a quantitative indicator of the observation that there are fewer variations in the way "1" is written relative to other digits.

Less importantly, on the far right of the plots we see decay of the k-variance begin to accelerate. This downward turn occurs roughly at the size of the dataset, because at this scale the bootstrapped estimator becomes less effective: For extremely large k the dataset looks like a collection of discrete points rather than a smooth distribution over $\mathbb{R}^{784}$.

11. Conclusion and future work. We can compute k-variance easily using a few lines of code, revealing potentially interesting structure hidden in a dataset or probability distribution. Hence, it is a straightforward addition to the data analysis toolkit. While its properties in four or fewer dimensions are somewhat unexpected, beyond this point k-variance provides an intuitive means of measuring intracluster variance. Somewhat surprisingly given the "curse of dimensionality" associated to optimal transport [25], we can use fewer data points to estimate k-variance of high-dimensional measures, as shown in section 9.

Beyond its immediate relevance as an analytical tool, k-variance motivates a wide variety of challenging research problems moving forward:

- Are there nontrivial pairs of measures $\mu, \nu \in \text{Prob}(\mathbb{R}^d)$ with $\text{Var}_k(\mu) = \text{Var}_k(\nu)$ for all $k \geq 1$? Under what conditions can a measure be reconstructed from its mean and sequence of k-variance values?
- Beyond the empirical estimator proposed in this paper, are there more efficient or unbiased stochastic estimators for k-variance?
- Is it possible to generalize k-variance to a notion of "k-covariance" for $d > 1$?
- Are there analogs of k-variance for higher-order moments of a measure?
- How do gradient flows of k-variance behave?

Further exploring k-variance in the context of data analysis and machine learning will be fruitful as well. For example, we can leverage the connection between the rates of change of k-variance with increasing k and intrinsic dimension to expose different-dimensioned structures, which create phase transitions in these curves. Moreover, we can use k-variance as an objective function to either emphasize or decrease within-cluster variation.

Acknowledgments. The authors thank Philippe Rigollet for early feedback and in particular noticing the connection of our work to random bipartite matching and to [6]; Lawrence Stewart for early discussion and experiments; David Wu for early discussions and help deriving combinatorial identities; Mikhail Yurochkin for discussion and feedback; and David Palmer and Paul Zhang for assistance running some experiments.

REFERENCES

[1] L. Ambrosio and F. Glaudo, *Finer estimates on the 2-dimensional matching problem*, J. Ec. Polytech. Math., 6 (2019), pp. 737–765.

[2] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, *Invariant Risk Minimization*, preprint, arXiv:1907.02893, 2019.

[3] B. C. Arnold and N. Balakrishnan, *Approximations to Moments of Order Statistics*, in Relations, Bounds and Approximations for Order Statistics, Springer, New York, 1989, pp. 73–107.

[4] F. Barthe and C. Bordenave, *Combinatorial Optimization over Two Random Point Sets*, in Séminaire de Probabilités XLV, C. Donati-Martin, A. Lejay and A. Rouault, eds., Springer, New York, 2013, pp. 483–535.

[5] D. Benedetto and E. Caglioti, *Euclidean random matching in 2D for non-constant densities*, J. Stat. Phys., 181 (2020), pp. 854–869.

[6] S. Bobkov and M. Ledoux, *One-dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances*, vol. 261, Amer. Math. Soc., Providence, RI, 2019.

[7] S. Caracciolo, C. Lucibello, G. Parisi, and G. Sicuro, *Scaling hypothesis for the Euclidean bipartite matching problem*, Phys. Rev. E, 90 (2014), 012118.

[8] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré, *Faster Wasserstein distance estimation with the Sinkhorn divergence*, in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, H.-T. Lin, eds., Conference on Neural Information Processing Systems, 2020.

[9] C.-Y. Chuang, Y. Mroueh, K. Greenewald, A. Torralba, and S. Jegelka, *Measuring generalization with optimal transport*, in Advances in Neural Information Processing Systems 34 pre-proceedings, M. Ranzato, A. Beygelzimer, K. Nguyen, P.S. Liang, J.W. Vaughan, Y. Dauphin, eds., 2021, preprint.

[10] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed., Wiley, Hoboken, NJ, 2003.

[11] J. B. de Monvel and O. Martin, *Almost sure convergence of the minimum bipartite matching functional in Euclidean space*, Combinatorica, 22 (2002), pp. 523–530.
[12] S. Dereich, M. Scheutzow, and R. Schottstedt, *Constructive quantization: Approximation by empirical measures*, Ann. Inst. Henri Poincaré Probab. Stat., (49) 2013, pp. 1183–1203.
[13] V. Dobrić and J. E. Yukich, *Asymptotics for transportation cost in high dimensions*, J. Theoret. Probab., 8 (1995), pp. 97–118.
[14] I. S. Duff and J. Koster, *On algorithms for permuting large entries to the diagonal of a sparse matrix*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 973–996.
[15] M. Goldman and D. Trevisan, *Convergence of asymptotic costs for random Euclidean matching problems*, Probab. Math. Phys., 2 (2021), pp. 341–362.
[16] Y. LeCun, C. Cortes, and C. J. Burges, *The MNIST Database of Handwritten Digits*, 1998, http://yann.lecun.com/exdb/mnist.
[17] C. McDiarmid, *On the Method of Bounded Differences*, in Surveys in Combinatorics, London Math. Soc. Lecture Note Ser. 141, J. Siemons, ed. Cambridge University Press, Cambridge, UK, 1989, pp. 148–188.
[18] F.-P. Paty and M. Cuturi, *Subspace Robust Wasserstein Distances*, in Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp. 5072–5081.
[19] G. Peyré and M. Cuturi, *Computational optimal transport: With applications to data science*, Found. Trends Mach. Learn., 11 (2019), pp. 355–607.
[20] T. Ramasubban, *The mean difference and the mean deviation of some discontinuous distributions*, Biometrika, 45 (1958), pp. 549–556.
[21] A. Rényi, *On the theory of order statistics*, Acta Math. Acad. Sci. Hung., 4 (1953), pp. 191–231.
[22] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Birkäuser, New York, 55 (2015).
[23] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Publ. Math. Inst. Hautes Études Sci., 81 (1995), pp. 73–205.
[24] C. Villani, *Topics in Optimal Transportation*, vol: 58, Amer. Math. Soc., Providence, RI, 2003.
[25] J. Weed and F. Bach, *Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance*, Bernoulli, 25 (2019), pp. 2620–2648.
[26] J. E. Yukich, *Probability Theory of Classical Euclidean Optimization Problems*, Springer, New York, 2006.
[27] M. Yurochkin and Y. Sun, *SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness*, in International Conference on Learning Representations, 2020.