Bernoulli **25**(4A), 2019, 2620–2648
https://doi.org/10.3150/18-BEJ1065

Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance

JONATHAN WEED[1] and FRANCIS BACH[2]

[1]*Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. E-mail: jweed@mit.edu*
[2]*INRIA, Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University, 75005 Paris, France. E-mail: francis.bach@inria.fr*

The Wasserstein distance between two probability measures on a metric space is a measure of closeness with applications in statistics, probability, and machine learning. In this work, we consider the fundamental question of how quickly the empirical measure obtained from n independent samples from μ approaches μ in the Wasserstein distance of any order. We prove sharp asymptotic and finite-sample results for this rate of convergence for general measures on general compact metric spaces. Our finite-sample results show the existence of multi-scale behavior, where measures can exhibit radically different rates of convergence as n grows.

Keywords: optimal transport; quantization; Wasserstein metrics

1. Introduction

The Wasserstein distance[1] is a measure of the closeness of probability distributions on metric spaces which has proven extremely useful in data science and machine learning, particularly in the analysis of images [43,46,49] and text [29,59]. This distance is especially useful in tasks such as classification and clustering, since it captures geometric features of the underlying data. Moreover, unlike other measures of distance between distributions, such as the Kullback–Leibler divergence or χ^2-divergence, the Wasserstein distance between two compactly support measures is finite even when neither measure is absolutely continuous with respect to the other, a situation that often arises when considering empirical distributions arising in practice.

Concretely, the Wasserstein distance measures how closely two measures can be coupled, where closeness is measured with respect to the underlying metric. For $p \in [1, \infty)$, the Wasserstein distance of order p between two distributions μ and ν on a metric space (X, D) is defined

[1]Calling this quantity the "Wasserstein distance" is, to borrow the verdict of Villani [55], "very questionable," since the attribution to Leonid Vasershtein is dubious at best. However, since this terminology is now by far the most common, we follow Villani's lead and adopt it in this work as well. Vershik [54] provides a review of the historical issues surrounding Kantorovich's role in defining this quantity.

1350-7265

as

$$W_p(\mu,\nu) := \inf_{\gamma\in\mathcal{C}(\mu,\nu)} \left(\int D(x,y)^p \,\mathrm{d}\gamma(x,y) \right)^{1/p},$$

where the infimum is taken over all *couplings* γ of μ and ν, that is, distributions on $X \times X$ whose first and second marginals agree with μ and ν, respectively [27]. It can be shown that W_p is a metric on the space of probability measures on X with finite pth moment [55], Chapter 6.

In statistical contexts, direct access to a distribution of interest μ is generally not available; instead, the statistician has access to i.i.d. samples from μ, or, equivalently, to an empirical distribution $\hat{\mu}_n$. For $\hat{\mu}_n$ to serve as a reasonable proxy to μ, we should insist that $\hat{\mu}_n$ and μ are close in the Wasserstein sense. In the large-n limit, this is indeed the case: if X is compact and separable and μ is a Borel measure, then for any $p \in [1,\infty)$,

$$W_p(\mu,\hat{\mu}_n) \to 0 \qquad \mu\text{-a.s.}$$

This result follows from the fact that Wasserstein distances metrize weak convergence [55], Corollary 6.13, and the fact that empirical measure $\hat{\mu}_n$ converges weakly to μ almost surely [53].

This raises the question of quantifying the rate of convergence of $\hat{\mu}_n$ to μ in W_p distance either in expectation or with high probability. This question is closely related to the *optimal quantization* problem [22], which asks how well a given distribution μ can be approximated by a discrete distribution with finite support, such as the empirical measure $\hat{\mu}_n$. This problem has wide applications in information theory, under the name rate distortion [13,48]; machine learning [10, 38]; and numerical methods [12]. Unfortunately, like many statistics and optimization problems involving measures on $\mathbb{R}^d$, the convergence of $\hat{\mu}_n$ to μ exhibits the so-called "curse of dimensionality" [3]. In the high-dimensional regime, the empirical distribution $\hat{\mu}_n$ becomes less and less representative as d becomes large [23], so that the convergence of $\hat{\mu}_n$ to μ in Wasserstein distance is slow.

This curse of dimensionality seems unavoidable. It was noted by Dudley [18] that any measure μ that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ satisfies

$$\mathbb{E}\big[W_1(\mu,\hat{\mu}_n)\big] \gtrsim n^{-1/d}.$$

This lower bound is asymptotically tight: Dudley showed that, when $d > 2$, a compactly supported measure on $\mathbb{R}^d$ satisfies

$$\mathbb{E}\big[W_1(\mu,\hat{\mu}_n)\big] \lesssim n^{-1/d}.$$

These results have been sharpened over the years, culminating in a tight almost sure limit theorem due to Dobrić and Yukich [17].

In short, these arguments establish that a d-dimensional measure yields a convergence rate in the W_1 distance of exactly $n^{-1/d}$. These results are in a sense disappointing, since they show that slow convergence is a necessary price to pay for high-dimensional data. However, they raise several questions about the behavior of the Wasserstein metric in practice:

- When can faster rates be achieved for measures that are not absolutely continuous with respect to the Lebesgue measure?

- Under what conditions can sharper finite-sample (i.e., non-asymptotic) rates be obtained?

Our goal in this work is to answer the above questions in a very general sense. We consider a bounded metric space X subject to mild technical conditions and prove upper and lower bounds on the rate of convergence for $W_p(\mu, \hat{\mu}_n)$ for all $p \in [1, \infty)$. Inspired by the bounds of [18], we show essentially tight asymptotic convergence rates for a large class of measures. In particular, our upper and lower bounds improve on many existing results in the literature [7,16,18,21], either in the generality with which they are applicable or the rates which are obtained. These results show that the rate of convergence of $W_p(\mu, \hat{\mu}_n)$ depends on a notion of the intrinsic dimension of the measure μ, which can be significantly smaller than the dimension of the metric space on which μ is defined.

Our second goal is to obtain finite-sample results which hold outside the asymptotic regime. A common phenomenon in practice is for a measure to exhibit different dimensional structure at different scales; this so-called multi-scale behavior arises in a range of applications [33,50,56]. We show that the convergence of $\hat{\mu}_n$ to μ in W_p for such measures can exhibit wildly different rates as n increases. In particular, they can enjoy a much faster convergence rate when n is small than they do in the large-n limit. We illustrate this phenomenon via a number of examples inspired by measures that arise in practice.

In both of the above regimes, we consider exclusively the question of how the expectation $\mathbb{E}[W_p(\mu, \hat{\mu}_n)]$ behaves. Controlling this quantity suffices to understand the behavior of $W_p(\mu, \hat{\mu}_n)$ because the Wasserstein distance concentrates very well around its expectation, a fact which we prove in Section 6. Combining this observation with the bounds we prove on $\mathbb{E}[W_p(\mu, \hat{\mu}_n)]$ yields sharp high-probability bounds.

We end by giving applications of our work to machine learning and statistics and sketch directions for future work.

2. Preliminaries

In this section, we present the mild assumptions on X under which our results hold. We also give background on Wasserstein distances and compare our results to prior work.

2.1. Assumptions

We are concerned with measures on a compact metric space (X, D). The first assumption is entirely standard and allows us to avoid many measure-theoretic difficulties:

Assumption 1. The metric space X is Polish, and all measures are Borel.

Since we limit ourselves to the compact case, $\operatorname{diam}(X)$ is necessarily finite, and for normalization purposes we assume the following.

Assumption 2. $\operatorname{diam}(X) \leq 1$.

Assumption 2 can always be made to hold by a simple rescaling of the metric, at the price of a multiplicative factor of $\operatorname{diam}(X)$. Indeed, if we denote by $\tilde{W}_p$ the Wasserstein distance with respect to the rescaled metric $\tilde{D} := \frac{1}{\operatorname{diam}(X)} D$, then the metric space $(X, \tilde{D})$ has diameter 1 and $W_p(\mu, \nu) = \operatorname{diam}(X) \tilde{W}_p(\mu, \nu)$ for all $p \in [1, \infty)$ and all measures μ and ν on X.

2.2. Background on Wasserstein distances

Above, we defined the Wasserstein p distance between two distributions μ and ν on (X, D) as

$$W_p(\mu, \nu) := \inf_{\gamma \in \mathcal{C}(\mu, \nu)} \left(\int D(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}.$$

A second definition, due to Monge [37,47], reads as follows:

$$W_p(\mu, \nu) := \inf_{T : \mu \circ T^{-1} = \nu} \left(\int D\big(x, T(x)\big)^p \, \mathrm{d}\mu(x) \right)^{1/p},$$

where the infimum is taken over all *transports* $T : X \to X$ such that the pushforward measure $\mu \circ T^{-1}$ equals ν. In general, this infimum in the Monge definition is not attained. Nevertheless, this formulation has an easy geometric interpretation: the Wasserstein distance measures the cost of moving mass from the measure μ to the measure ν with respect to the metric of X.

The special case W_1, which is also known as the Kantorovich-Rubinstein distance [55] or earth mover distance [43], has a particularly simple dual representation:

$$W_1(\mu, \nu) = \sup_{f \in \operatorname{Lip}(X)} \left| \int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu \right|, \tag{1}$$

where the supremum is taken over all 1-Lipschitz functions on X [26]. This dual representation makes W_1 significantly easier to bound [55], Remark 6.6. A more general dual formulation is also available for W_p for $p \neq 1$, but it is less simple to manipulate; more details appear in Section 6.

2.3. Related work

Our work generalizes several strands of work on the convergence rates of the empirical measure in Wasserstein distances. The first strand, inaugurated by Dudley [18], focuses on obtaining rates of convergence of $\hat{\mu}_n$ to μ based on the inherent dimension of the measure μ. In that paper, he analyzed the Fortet–Mourier metric [20] – which is equivalent to W_1 on bounded metric spaces – and obtained results matching the ones we present in Section 4 for the convergence of $\hat{\mu}_n$ to μ in W_1 distance, with a rate depending on the covering number of the support of μ. Dudley's argument relied extensively on the dual characterization of W_1 as a supremum over Lipschitz test functions, as in (1). As a result, his technique does not extend to W_p for $p \neq 1$.

An extension of Dudley's techniques to other values of p appears in [7]. Their approach is similar to ours, but our analysis is tighter: in the language of Section 4.2, they prove an upper

bound based on the quantity d_M whereas we obtain an upper bound based on the smaller quantity d_p^*.

A second strand [16,21] focuses on measures on $\mathbb{R}^d$ and obtains upper bounds on the rate of convergence of $\hat{\mu}_n$ to μ in W_p for all $p \in (0, \infty)$. The upper bounds arise from the construction of explicit couplings between $\hat{\mu}_n$ and μ. The construction of these couplings depends on the fact that μ is a measure on $\mathbb{R}^d$ and does not extend easily to general metric spaces. Moreover, the rates obtained, while tight for measures which are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, are not tight in general, as we show below. Nevertheless, the techniques employed in [16,21] are very similar to those employed in [7], and we follow the same approach.

We also note several other recent works [6,8] which have focused on obtaining tail bounds for the quantity $W_p(\mu, \hat{\mu}_n)$. The arguments of [8] rely on transportation inequalities such as the celebrated Bobkov-Götze inequality [5]. These arguments were simplified in [6], but, as noted in [7], the analysis becomes much easier if the development of tail bounds is divided into two steps: an estimate of the expectation $\mathbb{E}[W_p(\mu, \hat{\mu}_n)]$ and a concentration bound showing how well $W_p(\mu, \hat{\mu}_n)$ concentrates near that expectation. This is the approach we adopt: bounds on the expected value appear in Sections 4 and 5, and concentration bounds are obtained in Section 6.

2.4. Notation

The metric on X will always be denoted $D(\cdot, \cdot)$. Given a point $x \in X$ and $r > 0$, denote by $B(x, r)$ the open ball of radius r around x. The symbol log denotes the natural logarithm. The notation $f(n) \lesssim g(n)$ indicates that there exists a constant C, depending on f and g but not n, such that $f(n) \leq Cg(n)$ for all n.

3. Dyadic transport

In order to prove upper bounds for the Wasserstein distance, we show how to construct an efficient transport between two measures based on a recursive partitioning of the underlying space. By analogy with the dyadic intervals in $\mathbb{R}$, we seek a sequence of partitions of a set such that each partition is a refinement of the last, and such that the elements of the kth partition have diameter of order δ^k for some δ.

We formalize these requirements in the following definition [15], Section A. Denote by $\mathcal{B}(X)$ the Borel subsets of X.

Definition 1. A *dyadic partition* of a set $S \subseteq X$ with parameter $\delta < 1$ is a sequence $\{\mathcal{Q}^k\}_{1 \leq k \leq k^*}$ with $\mathcal{Q}^k \subseteq \mathcal{B}(X)$ possessing the following properties:

- The sets in $\mathcal{Q}^k$ form a partition of S.
- If $Q \in \mathcal{Q}^k$, then $\mathrm{diam}(Q) \leq \delta^k$.
- If $Q^{k+1} \in \mathcal{Q}^{k+1}$ and $Q^k \in \mathcal{Q}^k$, then either $Q^{k+1} \subseteq Q^k$ or $Q^{k+1} \cap Q^k = \varnothing$. That is, the $(k+1)$th partition is a refinement of the kth partition.

The following proposition bounds $W_p^p(\mu, \nu)$ in terms of the mass μ and ν assign to elements of a dyadic partition.

Proposition 1. *Let μ and ν be two Borel probability measures on X, and let S be a set such that $\mu(S) = \nu(S) = 1$. If $\{\mathcal{Q}^k\}_{1 \le k \le k^*}$ is a dyadic partition of S with parameter δ, then*

$$W_p^p(\mu, \nu) \le \delta^{k^* p} + \sum_{k=1}^{k^*} \delta^{(k-1)p} \sum_{Q_i^k \in \mathcal{Q}^k} \left|\mu\left(Q_i^k\right) - \nu\left(Q_i^k\right)\right|.$$

The upper bound in Proposition 1 arises from the explicit construction of a coupling between μ and ν. Proposition 1 is not new and appears to have been rediscovered many times. In particular, it is implicit in the proof of [7], Proposition 1.1, and similar results have appeared before in papers bounding the convergence of W_p^p when $X = \mathbb{R}^d$ [16,21]. An analogous bound has also been used in the computer science community [2,24]. The idea of bounding the quantity $W_p^p(\mu, \nu)$ by considering the mass each measure assigns to elements of a sequence of partitions is present also in [1], where it is used to obtain sharp results for the case $X = [0, 1]^2$. We include a proof in the supplement [57] for clarity and because we could not find a suitably general version explicitly stated in the literature.

Proposition 1 is stated for W_p^p, but can easily adapted to optimal transport with a general cost $c(\cdot, \cdot)$ by replacing the requirement that $\operatorname{diam}(Q) \le \delta^k$ in Definition 1 by the requirement that $\sup_{x,y \in Q} c(x, y) \le \delta^k$.

Boissard and Le Gouic [7] used a version of Proposition 1 to prove a bound on $W_p^p(\mu, \hat{\mu}_n)$ based on the *covering number* of the set S, a definition of which appears in Section 4, below. However, their results are not sharp, and they do not recover the rates obtained in [18] for the case $p = 1$. In Section 4, we show how to improve their argument to obtain sharper results, which extend the rates from [18] to all $p \in [1, \infty)$.

4. Asymptotic upper and lower bounds

In this section, we show asymptotic upper and lower bounds for W_p that hold for all $p \in [1, \infty)$. These bounds extend results of [18] to the case $p \neq 1$ and improve the bounds of [7] by focusing on a set S to which μ assigns mass of *almost* 1 rather than on the larger set $\operatorname{supp}(\mu)$. We will also show a broad class of measures for which our bounds are asymptotically tight.

4.1. Definitions

To state our bounds, we will define several notions of dimension of a measure.

Definition 2. Given a set $S \subseteq X$, the *ε-covering number of S*, denoted $\mathcal{N}_\varepsilon(S)$, is the minimum m such that there exists m closed balls $B_1, \ldots, B_m$ of diameter ε such that $S \subseteq \bigcup_{1 \le i \le m} B_i$. The *$\varepsilon$-dimension of S* is the quantity

$$d_\varepsilon(S) := \frac{\log \mathcal{N}_\varepsilon(S)}{-\log \varepsilon}.$$

When working with measures instead of sets, it is convenient to be able to ignore a small fraction of the mass. The following definition appears in [18], which notes a connection to the $\varepsilon; \delta$ entropy introduced by [41].

Definition 3. Given a measure μ on X, the *(ε, τ)-covering number* is

$$\mathcal{N}_\varepsilon(\mu, \tau) := \inf\{\mathcal{N}_\varepsilon(S) : \mu(S) \geq 1 - \tau\}$$

and the *(ε, τ)-dimension* is

$$d_\varepsilon(\mu, \tau) := \frac{\log \mathcal{N}_\varepsilon(\mu, \tau)}{-\log \varepsilon}.$$

For convenience, let

$$\mathcal{N}_\varepsilon(\mu) := \mathcal{N}_\varepsilon(\mu, 0),$$
$$d_\varepsilon(\mu) := d_\varepsilon(\mu, 0).$$

Note that $\mathcal{N}_\varepsilon(\mu) = \mathcal{N}_\varepsilon(\text{supp}(\mu))$, and that $\mathcal{N}_\varepsilon(\mu, \tau)$ and $d_\varepsilon(\mu, \tau)$ increase as τ decreases. We now define our main notions of dimension of a measure.

Definition 4. The *upper and lower Wasserstein dimensions* are respectively,

$$d_p^*(\mu) := \inf\left\{s \in (2p, \infty) : \limsup_{\varepsilon \to 0} d_\varepsilon\left(\mu, \varepsilon^{\frac{sp}{s-2p}}\right) \leq s\right\},$$
$$d_*(\mu) := \lim_{\tau \to 0} \liminf_{\varepsilon \to 0} d_\varepsilon(\mu, \tau).$$

Note that the monotonicity of $d_\varepsilon(\mu, \tau)$ in τ implies that the limit in the definition of $d_*(\mu)$ exists. The definition of d_p^* is complicated by the fact that the behavior of the Wasserstein distance is very different when the dimension is small. For convenience, we only treat the case where the dimension is larger than $2p$. We note that the monotonicity of $d_\varepsilon(\mu, \tau)$ in τ also implies that $d_* \leq d_p^*$ for all p.

Our definition of the upper Wasserstein dimension is new. Dudley [18] considered measures satisfying a bound of the form $\mathcal{N}_\varepsilon(\mu, \varepsilon^{\frac{s}{s-2}}) \leq C\varepsilon^{-s}$ for all sufficiently small ε; the definition of $d_p^*(\mu)$ is the correct generalization to the $p \neq 1$ case. The lower Wasserstein dimension was introduced by Young [58], who credits the idea to Ledrappier [32], in the context of dynamical systems. The term Wasserstein dimension is ours, and is justified by Theorem 1 below.

4.2. Comparison with other notions of dimension

To make it easier to interpret the quantities $d_p^*(\mu)$ and $d_*(\mu)$, we sketch here their relationship with two other well-known notions of dimensions for the measure μ, the Minkowski dimension (also known as the Minkowski–Bouligand or box-counting dimension) and the Hausdorff dimension. Both quantities have long been studied in fractal and metric geometry [19].

Definition 5. The *Minkowski dimension* of a set S is the quantity

$$\dim_M(S) := \limsup_{\varepsilon \to 0} d_\varepsilon(S).$$

The *d-Hausdorff measure* of S is

$$\mathcal{H}^d(S) := \liminf_{\varepsilon \to 0} \left\{ \sum_{k=1}^{\infty} r_k^d : S \subseteq \bigcup_{k=1}^{\infty} B(x_k, r_k); r_k \leq \varepsilon \ \forall k \right\},$$

and its *Hausdorff dimension* is

$$\dim_H(S) := \inf\{d : \mathcal{H}^d(S) = 0\}.$$

Given a measure μ, the *Minkowski and Hausdorff dimensions* of μ are respectively,

$$d_M(\mu) := \inf\{\dim_M(S) : \mu(S) = 1\},$$
$$d_H(\mu) := \inf\{\dim_H(S) : \mu(S) = 1\}.$$

We note that the quantities $d_M(\mu)$ and $d_H(\mu)$ are upper and lower bounds on the Wasserstein dimensions.

Proposition 2.

$$d_H(\mu) \leq d_*(\mu) \leq d_p^*(\mu).$$

If $d_M(\mu) \geq 2p$, then

$$d_p^*(\mu) \leq d_M(\mu).$$

A proof appears in the supplement [57].

None of the inequalities in Proposition 2 can be replaced by equalities. Examples of measures μ for which $d_H(\mu) < d_*(\mu)$ are complicated; one appears in [25], Remark 7.8. It is much easier to find examples in which $d_*(\mu)$, $d_p^*(\mu)$, and $d_M(\mu)$ do not agree. For instance, it is easy to see that $d_*(\mu) = 0$ for any discrete measure, but the countable set $S := (\{k^{-1}\}_{k=1}^{\infty})^d \subset [0,1]^d$ has Minkowski dimension $d/2$. By choosing $d > 4p$ and choosing a measure μ supported on S with appropriately slow decay, one can ensure that $d_p^*(\mu)$ is strictly less than $d/2$, and hence strictly between $d_*(\mu)$ and $d_M(\mu)$.

4.3. Main result

With these definitions in place, we can state our main asymptotic bound.

Theorem 1. *Let $p \in [1, \infty)$. If $s > d_p^*(\mu)$, then*

$$\mathbb{E}\big[W_p(\mu, \hat{\mu}_n)\big] \lesssim n^{-1/s}.$$

If $t < d_*(\mu)$, *then*

$$W_p(\mu, \hat{\mu}_n) \gtrsim n^{-1/t}.$$

The upper and lower bounds are proved below and are corollaries of more precise results with explicit constants (Propositions 5 and 6). Note that the lower bound does not merely hold in expectation. Indeed, such a lower bound holds for *any* discrete measure supported on at most n points.

Theorem 1 improves on several existing results. For the upper bound, Dudley [18] showed that, if $s > d_1^*(\mu)$, then

$$\mathbb{E}\big[W_1(\mu, \hat{\mu}_n)\big] \lesssim n^{-1/s},$$

but his proof technique applied only to $p = 1$. Boissard and Le Gouic [7] extended this bound to all p, but only if $s > d_M(\mu) \geq 2p$. Since $d_p^*(\mu) \leq d_M(\mu)$ with some measures exhibiting strict inequality, our result is sharper.

Dudley [18] proved a lower bound for W_1 – and hence, by monotonicity of W_p in p, for W_p for all $p \in [1, \infty)$ – based on the quantity

$$d_{1/2}(\mu) := \liminf_{\varepsilon \to 0} \frac{\log \mathcal{N}_\varepsilon(\mu, 1/2)}{-\log \varepsilon},$$

which is easily seen to be smaller than $d_*(\mu)$, with strict inequality possible. Our argument is a simple extension of his.

4.4. Proof of upper bound

The upper bound of Theorem 1 follows from Proposition 5, below.

To apply the bound of Proposition 1, we need to show the existence of a suitable dyadic partition. The following proposition is an extension of [7], Lemma 2.1, and shows that we can choose a dyadic partition which provides an almost optimal covering of subsets of S.

Proposition 3. *Fix* $S \in \mathcal{B}(X)$. *Let* k^* *be any positive integer for which the covering number* $\mathcal{N}_{3^{-(k^*+1)}}(S)$ *is finite, and let* $\{S_k\}_{1 \leq k \leq k^*}$ *be a sequence of Borel subsets of* S. *There exists a dyadic partition* $\{\mathcal{Q}^k\}_{1 \leq k \leq k^*}$ *of* S *with parameter* $\delta = 1/3$ *such that for* $1 \leq k \leq k^*$, *the number of sets in* $\mathcal{Q}^k$ *intersecting* S_k *is at most* $\mathcal{N}_{3^{-(k+1)}}(S_k)$.

A proof appears in the supplement [57].

All the upper bounds we prove rely on the following fundamental estimate, which was used in [21] to provide bounds in the case where $X = \mathbb{R}^d$.

Proposition 4. *If* S *is any Borel set, then*

$$\mathbb{E}\left[\sum_{Q_i^k \in \mathcal{Q}^k} \left|\mu\big(Q_i^k\big) - \hat{\mu}_n\big(Q_i^k\big)\right|\right] \leq 2\big(1 - \mu(S)\big) + \sqrt{\big|\big\{i : Q_i^k \cap S \neq \varnothing\big\}\big|/n}.$$

Proof. Let $Q(S) = \{i : Q_i^k \cap S \neq \varnothing\}$, and write

$$S' = \bigcup_{i \in Q(S)} Q_i^k.$$

Since $n\hat{\mu}_n(Q_i^k)$ is a Binomial random variable with parameters $(n, \mu(Q_i^k))$, we have the simple bound (see, e.g., [4]):

$$\mathbb{E}\big|\mu\big(Q_i^k\big) - \hat{\mu}_n\big(Q_i^k\big)\big| \leq \sqrt{\mu\big(Q_i^k\big)/n} \wedge 2\mu\big(Q_i^k\big).$$

Applying the first bound on S' and the second bound on $X \setminus S'$ yields

$$\mathbb{E}\bigg[\sum_{Q_i^k \in \mathcal{Q}^k} \big|\mu\big(Q_i^k\big) - \hat{\mu}_n\big(Q_i^k\big)\big|\bigg] \leq 2\mu\big(X \setminus S'\big) + \sum_{i \in Q(S)} \sqrt{\mu\big(Q_i^k\big)/n}.$$

Since the second sum contains $|Q(S)|$ terms and $\sum_{i \in Q(S)} \mu(Q_i^k) = \mu(S') \leq 1$, the final bound follows from Cauchy–Schwarz. □

The key step in proving the upper bound of Theorem 1 is giving a bound for $\mathbb{E}[W_p^p(\mu, \hat{\mu}_n)]$ in terms of the quantity $d_\varepsilon(\mu, \varepsilon^{\frac{sp}{s-2p}})$, which appears in the definition of d_p^*.

Proposition 5. *Let $p \in [1, \infty)$. Suppose there exists an $\varepsilon' \leq 1$ and $s > 2p$ such that*

$$d_\varepsilon\big(\mu, \varepsilon^{\frac{sp}{s-2p}}\big) \leq s$$

for all $\varepsilon \leq \varepsilon'$. Then

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \leq C_1 n^{-p/s} + C_2 n^{-1/2},$$

where

$$C_1 := 3^{\frac{3sp}{s-2p}+1}\bigg(\frac{1}{3^{\frac{s}{2}-p} - 1} + 3\bigg) \quad \text{and} \quad C_2 := \big(27/\varepsilon'\big)^{\frac{s}{2}}.$$

In particular,

$$\mathbb{E}\big[W_p(\mu, \hat{\mu}_n)\big] \lesssim n^{-1/s}.$$

The assumption that $s > 2p$ implies that the first term in the above bound is asymptotically larger than the second term. Note also that C_1 decreases as s increases, so that as long as s is bounded away from $2p$, the constant C_1 has no dependence on the dimension s. On the other hand, C_2 does depend exponentially on s, even though the term $C_2 n^{-1/2}$ is asymptotically negligible.

The presence of two terms in the upper bound of Proposition 5 is a consequence of the weakness of the assumption that the bound on d_ε holds only for ε sufficiently small rather than for all ε. In Proposition 10, below, we remove the $n^{-1/2}$ term by adopting a stronger assumption on d_ε.

Proof. If $n < (27/\varepsilon')^s$, then the second term is larger than 1, so the bound holds from the trivial fact that $W_p^p(\mu,\nu) \le \mathrm{diam}(X) \le 1$ for any measures μ,ν supported on X. We therefore assume that $n \ge (27/\varepsilon')^s$.

For convenience, write $\alpha := sp/(s-2p)$ and $\ell := \lceil \frac{-\log \varepsilon'}{\log 3} \rceil$. Let $k^* := \lfloor \frac{\log n}{s \log 3} \rfloor - 2$. Let k' be the largest integer in the range $[\ell, k^*]$ satisfying $k' \le \frac{p}{\alpha} \cdot \frac{\log n}{s \log 3}$, or ℓ if no such integer exists.

Our assumptions imply that for all $k \ge \ell$,

$$\mathcal{N}_{3^{-k}}\left(\mu, 3^{-\alpha k}\right) \le 3^{ks}.$$

Hence for $k \ge k'$, there exists a set T_k of mass at least $1 - 3^{-\alpha k'}$ such that

$$\mathcal{N}_{3^{-k}}(T_k) \le 3^{ks}.$$

Applying Proposition 3 with $S_k = T_{k'}$ for $k < k'$ and $S_k = T_{k+1}$ for $k \ge k'$ implies the existence of a dyadic partition $\{\mathcal{Q}^k\}_{1 \le k \le k^*}$ of X such that the number of sets of $\mathcal{Q}^k$ intersecting S_k is at most $\mathcal{N}_{3^{-(k+1)}}(S_k)$.

Using this dyadic partition in Proposition 1 and applying Proposition 4 yields

$$\begin{aligned}\mathbb{E}\left[W_p^p(\mu,\hat{\mu}_n)\right] &\le 3^{-k^* p} + \sum_{k=1}^{k'-1} 3^{-(k-1)p} \sqrt{\frac{\mathcal{N}_{3^{-(k+1)}}(T_{k'})}{n}} \\ &\quad + \sum_{k=k'}^{k^*} 3^{-(k-1)p} \sqrt{\frac{\mathcal{N}_{3^{-(k+1)}}(T_{k+1})}{n}} \\ &\quad + 2 \cdot 3^{-\alpha k'} \sum_{k=1}^{k^*} 3^{-(k-1)p}.\end{aligned}$$

Since $\mathcal{N}_\varepsilon(T)$ increases as ε decreases, for $k \le k' - 1$ we have the bound

$$\mathcal{N}_{3^{-(k+1)}}(T_{k'}) \le \mathcal{N}_{3^{-k'}}(T_{k'}) \le 3^{k's}.$$

By construction, the sets T_k also satisfy for $k \ge k'$

$$\mathcal{N}_{3^{-(k+1)}}(T_{k+1}) \le 3^{(k+1)s}.$$

Combining these bounds with the bound $\sum_{k=1}^{k^*} 3^{-(k-1)p} \le 3/2$ for $p \ge 1$ yields

$$\begin{aligned}\mathbb{E}\left[W_p^p(\mu,\hat{\mu}_n)\right] &\le 3^{-k^* p} + \frac{3}{2}\left(\frac{3^{k's/2}}{\sqrt{n}} + 2 \cdot 3^{-\alpha k'}\right) + 3^{2p} \sum_{k=k'}^{k^*} \frac{3^{(k+1)(\frac{s}{2}-p)}}{\sqrt{n}} \\ &\le 3^{-k^* p} + \frac{3}{2}\left(\frac{3^{k's/2}}{\sqrt{n}} + 2 \cdot 3^{-\alpha k'}\right) + \frac{3^{-k^* p}}{3^{\frac{s}{2}-p} - 1} \sqrt{\frac{3^{(k^*+2)s}}{n}}.\end{aligned}$$

The choice of k^* implies that $3^{(k^*+2)s} \le n$ and that $3^{-k^*p} \le 3^{3p} n^{-p/s}$, and the choice of k' implies that $\alpha k' > p\frac{\log n}{s \log 3} - 3\alpha$, so that $3^{-\alpha k'} < 3^{3\alpha} n^{-p/s}$. Combining these estimates yields

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \le \left(3^{3p} + \frac{3^{3p}}{3^{\frac{s}{2}-p} - 1} + 3^{3\alpha+1}\right) n^{-p/s} + \frac{3 \cdot 3^{k's/2}}{2\sqrt{n}}.$$

The definition of k' implies that $sk' \le \max\{s\ell, (\frac{p}{\alpha} \cdot \frac{\log n}{\log 3})\}$, so

$$3^{k's/2} \le 3^{\ell s/2} + n^{p/2\alpha} = 3^{\ell s/2} + n^{1/2} n^{-p/s}.$$

Plugging in the definitions of C_1 and C_2 then yields the claim. □

Corollary 1 (Theorem 1, upper bound). *If $s > d_p^*(\mu)$, then*

$$\mathbb{E}\big[W_p(\mu, \hat{\mu}_n)\big] \lesssim n^{-1/s}.$$

Proof. If $s > d_p^*(\mu)$, then there exists an ε' such that $d_\varepsilon(\mu, \varepsilon^{\frac{sp}{s-2p}k}) \le s$ for all $\varepsilon \le \varepsilon'$. Apply Proposition 5. □

4.5. Proof of lower bound

Our asymptotic lower bounds involving $d_*(\mu)$ follow from a much simpler argument. One striking feature of this lower bound is that it actually holds not merely for the empirical measure $\hat{\mu}_n$ but indeed for *any* measure ν supported on at most n atoms. That such lower bounds are often tight for empirical measures is a rather surprising fact, which has been noted several times, including in Dudley's original paper [7,10,16,18,28].

The following proposition is adapted from [18] and forms the core of the lower bound.

Proposition 6. *Suppose that there exist positive constants ε', τ, and t such that*

$$\mathcal{N}_\varepsilon(\mu, \tau) \ge \varepsilon^{-t}$$

for all $\varepsilon \le \varepsilon'$. If $n > \varepsilon'^{-t}$ and ν is any measure supported on at most n points, then

$$W_p^p(\mu, \nu) \ge \tau 4^{-p} n^{-p/t}.$$

Proof. Choose $\varepsilon := n^{-1/t}/2$, and let $S := \bigcup_{x \in \mathrm{supp}(\nu)} B(x, \varepsilon/2)$. Since $\mathcal{N}_\varepsilon(\mu, \tau) \ge \varepsilon^{-t} > n$, we must have $\mu(S) < 1 - \tau$. Therefore, if $X \sim \mu$, then $D(X, \mathrm{supp}(\nu)) \ge \varepsilon/2$ with probability at least τ. Hence if (X, Y) is any coupling of μ and ν,

$$\mathbb{E}\big[D(X, Y)^p\big] \ge \mathbb{E}\big[D\big(X, \mathrm{supp}(\nu)\big)^p\big] \ge \tau(\varepsilon/2)^p = \tau 4^{-p} n^{-p/t}.$$ □

Corollary 2 (Theorem 1, lower bound). *If $t < d_*(\mu)$ and ν is any measure supported on at most n points, then*

$$W_p(\mu, \nu) \gtrsim n^{-1/t}.$$

Proof. By the definition of $d_*(\mu)$, for any $t < d_*(\mu)$, there exist constants ε' and τ as in the statement of Proposition 6. The claim follows. □

4.6. Regular spaces

The remark after Proposition 2 establishes that $d_*(\mu)$ and $d_p^*(\mu)$ do not agree in general. However, these dimensions do agree whenever the measure is sufficiently well behaved. In this section, we give several broad classes of examples for which they do match, and for which our bounds are therefore sharp.

The following proposition gives a simple condition under which this agreement occurs.

Proposition 7. *Let $\mathcal{H}^d$ be the d-dimensional Hausdorff measure on a closed set S. If $\mu \ll \mathcal{H}^d$ and* $\operatorname{supp}(\mu) \subseteq S$, *then for any $p \in [1, d/2]$,*

$$d \le d_*(\mu) \le d_p^*(\mu) \le d_M(S).$$

In particular, if $d = d_M(S)$, then $d_*(\mu) = d_p^*(\mu) = d$.

A proof appears in the supplement [57].

Proposition 7 immediately implies the result quoted in the introduction (up to subpolynomial factors): since the set $[0, 1]^d$ satisfies $d = d_M([0, 1]^d)$, Theorem 1 implies that any measure μ absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ (or, equivalently, to $\mathcal{H}^d$) must satisfy

$$n^{-1/t} \lesssim \mathbb{E}\big[W_p(\mu, \hat{\mu}_n)\big] \lesssim n^{-1/s}$$

for any $t < d < s$ and $p \in [1, d/2]$.

Limiting our attention to sets for which the Hausdorff measure is well behaved motivates the following definition, which appears in [22].

Definition 6. A set S is *regular of dimension d* if it is compact and there exists constants c and r_0 such that the d-dimensional Hausdorff measure $\mathcal{H}^d$ on S satisfies

$$\frac{1}{c} r^d \le \mathcal{H}^d\big(B(x, r)\big) \le c r^d,$$

for all $x \in S$ and $r \in [0, r_0]$.

It is well known (see, e.g., [34], Theorem 5.7) that $d_M(S) = d$ if S is regular of dimension d. We therefore obtain the following simple characterization.

Proposition 8. *If the support of μ is a regular set of dimension d and $\mu \ll \mathcal{H}^d$, then for any $p \in [1, d/2]$,*

$$d_*(\mu) = d_p^*(\mu) = d.$$

The following proposition, which appears in [22], shows that many well behaved sets are regular, and so implies the existence of many examples for which our results are tight.

Proposition 9 ([22]). *The following sets are regular of dimension d:*

- *Nonempty, compact convex sets spanned by an affine space of dimension d,*
- *Relative boundaries of nonempty, compact convex sets of dimension $d + 1$,*
- *Compact d-dimensional differentiable manifolds,*
- *Self-similar sets with similarity dimension d.*

Moreover, regularity is preserved under finite unions and bi-Lipschitz maps.

5. Finite-sample bounds and multiscale behavior

The results of Section 4 imply that for any sufficiently regular d-dimensional measure μ, the empirical measure $\hat{\mu}_n$ approaches μ in W_p at a rate of approximately $n^{-1/d}$. For example, if μ is absolutely continuous with respect to the Lebesgue measure on $[0, 1]^d$, Dudley showed that the slow $n^{-1/d}$ rate is unavoidable [18]. Faster rates can be obtained if μ is singular: for instance, if μ is a sum of a finite number of Dirac masses, then Proposition 1 can be used to show that $\hat{\mu}_n$ approaches μ at a much faster $n^{-1/2p}$ rate, independent of the ambient dimension.

However, what should one expect if μ is *approximately* a sum of Dirac masses (or, in general, approximately low dimensional)? Suppose for instance, that μ is the convolution of a sum of Dirac masses with an isotropic Gaussian of small variance. Since μ has a density, $W_p(\mu, \hat{\mu}_n)$ must scale like $n^{-1/d}$ eventually, but it is possible that the convergence of $\hat{\mu}_n$ to μ should improve due to the fact that μ is almost singular.

It turns out that this is indeed the case, as we show in this section. We begin by proving a sharper version of Proposition 5 better suited to non-asymptotic results. In the second half of this section, we show how this non-asymptotic bound can be used to prove faster convergence rates in the finite-sample regime for situations like the one described above.

5.1. Finite-sample behavior

The statement of Proposition 5 only assumes a bound on the quantity $d_\varepsilon(\mu, \tau)$ for sufficiently small ε. It is therefore well suited to establishing results of an asymptotic nature. On the other hand, the resulting bound did not give any indication of the behavior in the small-n regime, since the bound was vacuous for $n \lesssim (\varepsilon')^{-d_p^*}$.

If we have stronger control over $d_\varepsilon(\mu, \tau)$, then the proof of Proposition 5 can be modified to yield a finite-sample result. In particular, if we can control $d_\varepsilon(\mu, \tau)$ for all ε larger than a certain threshold, we can prove an upper bound without the $n^{-1/2}$ term present in Proposition 5.

Proposition 10. *Fix* $p \in [1, \infty)$. *Write* $d_{\geq \varepsilon}(\mu, \tau) := \sup_{\varepsilon' \in [\varepsilon, 1/9]} d_{\varepsilon}(\mu, \tau)$, *and let* $d_n := \inf_{\varepsilon > 0} \max\{d_{\geq \varepsilon}(\mu, \varepsilon^p), \frac{\log n}{-\log \varepsilon}\}$. *If* $d_n > 2p$, *then*

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \leq C_1 n^{-p/d_n},$$

where

$$C_1 := 27^p \left(2 + \frac{1}{3^{\frac{d_n}{2} - p} - 1}\right).$$

As in Proposition 5, the constant C_1 is independent of the dimension as long as d_n is bounded away from $2p$. Loosely speaking, the definition of d_n implies that $\mathbb{E} W_p^p(\mu, \hat{\mu}_n) \lesssim \varepsilon_n^p$ where ε_n satisfies $d_{\geq \varepsilon_n}(\mu, \varepsilon_n^p) \approx \frac{\log n}{-\log \varepsilon_n}$.

Proof. Fix an arbitrary ε, and let $d := \max\{d_{\geq \varepsilon}(\mu, \varepsilon^p), \frac{\log n}{-\log \varepsilon}\}$, where $d > 2p$. If $n^{-1/d} \geq 1/27$, then the bound $W_p^p(\mu, \hat{\mu}_n) \leq C_1 n^{-p/d}$ is trivial, so assume that $n > 3^{3d}$.

Let $k^* := \lfloor \frac{\log n}{d \log 3} \rfloor - 2$. As in the proof of Proposition 5, we can choose sets $S_1, \ldots, S_{k^*}$ such that $\mu(S_k) \geq 1 - \varepsilon^p$ and $\mathcal{N}_{3^{-(k+1)}}(S_k) = \mathcal{N}_{3^{-(k+1)}}(\mu, \varepsilon^p)$ for $1 \leq k \leq k^*$. Applying Proposition 3 to construct an appropriate dyadic partition and using Propositions 1 and 4 yields

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \leq 3^{-k^* p} + \sum_{k=1}^{k^*} 3^{-(k-1)p} \sqrt{\frac{\mathcal{N}_{3^{-(k+1)}}(\mu, \varepsilon^p)}{n}} + 2\varepsilon^p \sum_{k=1}^{k^*} 3^{-(k-1)p}.$$

By the definition of k^*, for $1 \leq k \leq k^*$,

$$3^{(k+1)d} \leq n,$$

so $3^{-(k+1)} \geq \varepsilon$. Hence $3\varepsilon^p \leq 3^{-k^* p}$ and $\mathcal{N}_{3^{-(k+1)}}(\mu, \varepsilon^p) \leq 3^{(k+1)d}$ for $1 \leq k \leq k^*$, and applying the bound $\sum_{k=1}^{k^*} 3^{-(k-1)p} \leq 3/2$ for $p \geq 1$ yields

$$\begin{aligned}
\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] &\leq 3^{-k^* p} + \frac{3^{-k^* p}}{3^{\frac{d}{2} - p} - 1} \sqrt{\frac{3^{(k^*+2)d}}{n}} + 3\varepsilon^p \\
&\leq \left(2 + \frac{1}{3^{\frac{d_n}{2} - p} - 1}\right) 3^{-k^* p} \\
&\leq C_1 n^{-p/d},
\end{aligned}$$

where in the last step we have used the fact that $d \geq d_n$ and $\frac{1}{3^{\frac{d}{2} - p} - 1}$ is decreasing in d.

Taking the infimum over all possible choices of ε yields the bound. □

Note in the proof of Proposition 10 that we in fact only needed control over $d_{\varepsilon'}(\mu, \tau)$ for ε' of the form 3^{-k} for k a positive integer, though for simplicity we have assumed that we can bound $d_{\varepsilon'}(\mu, \tau)$ for all $\varepsilon' \in [\varepsilon, 1/9]$.

The upper bound of Proposition 10 suggests that measures can have truly different rates of convergence at different scales. The following proposition shows that Proposition 10 is essentially tight and that this multiscale behavior indeed can occur, in the sense that for *any* decreasing sequence δ_n satisfying mild conditions, there exists a measure μ such that $n^{-1/d_n} \approx \delta_n$ and $\mathbb{E}[W_p(\mu, \hat{\mu}_n)] \geq Cn^{-1/d_n}$ for all n. In other words, for any desired rate of decrease, there exists a measure such that $\mathbb{E}[W_p(\mu, \hat{\mu}_n)]$ converges to 0 at precisely that rate. Such measures can even be found when the underlying metric is induced by the ℓ_∞ norm on real space. As with the lower bound proved in Proposition 6, above, the following bound in fact holds for *any* measure ν supported on at most n points.

A proof appears in the supplement [57].

Proposition 11. *Let δ_n be a nonincreasing sequence in $(0, 1)$ with the following properties*:

- *the bound $\delta_n > n^{-1}$ holds for all $n \geq 2$ (i.e., δ_n does not decrease too quickly),*
- *the sequence $\frac{\log n}{-\log \delta_n}$ is nondecreasing (i.e., the rate of decrease of δ_n slows), and*
- *there exist constants $c > 1$ and $\alpha \in [-1, 0)$ such that $\frac{1}{c} n^\alpha \leq \delta_n \leq cn^\alpha$ for all n sufficiently large (i.e., δ_n eventually decreases polynomially in n).*

There exists a measure μ on the metric space $([0, 1]^m, \ell_\infty)$ for some m such that, if d_n is defined as in Proposition 10, then $\frac{1}{16}\delta_n \leq n^{-1/d_n} \leq 4\delta_n$ and

$$\mathbb{E}\big[W_p(\mu, \nu)\big] \geq 2^{-6} n^{-1/d_n}$$

for all $p \in [1, \infty)$, all $n \geq 1$, and any measure ν supported on at most n points.

Proposition 10 only holds when $d_n > 2p$, so for completeness we conclude this section by providing a second bound that can be used when Proposition 10 does not apply. The following bound is always valid and is sharper when the asymptotic dimension of μ is small.

Proposition 12. *Let $m_n := \inf_{\varepsilon > 0} \max\{\mathcal{N}_\varepsilon(\mu, \varepsilon^p), n\varepsilon^{2p}\}$. Then*

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \leq C_1 \sqrt{\frac{m_n}{n}},$$

where $C_1 := 9^p + 3$.

Proof. Fix an arbitrary ε, and let $m := \max\{\mathcal{N}_\varepsilon(\mu, \varepsilon^p), n\varepsilon^{2p}\}$. If $\varepsilon \geq 1/9$, then the bound $W_p^p(\mu, \hat{\mu}_n) \leq C_1 \sqrt{\frac{m}{n}}$ is trivial, so assume $\varepsilon < 1/9$.

Let $k^* := \lfloor \frac{-\log \varepsilon}{\log 3} \rfloor - 1$. Following the proof of Proposition 10, we have

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \leq 3^{-k^* p} + \sum_{k=1}^{k^*} 3^{-(k-1)p} \sqrt{\frac{\mathcal{N}_{3^{-(k+1)}}(\mu, \varepsilon^p)}{n}} + 2\varepsilon^p \sum_{k=1}^{k^*} 3^{-(k-1)p}.$$

The monotonicity of $\mathcal{N}_\varepsilon(\mu, \tau)$ implies that $\mathcal{N}_{3^{-(k+1)}}(\mu, \varepsilon^p) \leq \mathcal{N}_\varepsilon(\mu, \varepsilon^p) \leq m$ for all $k \leq k^*$. Plugging in this estimate and applying the bound $\sum_{k=1}^{k^*} 3^{-(k-1)p} \leq 3/2$ for $p \geq 1$ yields

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \leq 3^{-k^* p} + \frac{3}{2}\sqrt{\frac{m}{n}} + \frac{3}{2}\varepsilon^p \leq 3^{-k^* p} + 3\sqrt{\frac{m}{n}}.$$

On the other hand, $3^{-k^* p} = 9^p 3^{-(k^*+2)p} < 9^p \varepsilon^p \leq 9^p \sqrt{\frac{m}{n}}$, so

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \leq C_1 \sqrt{\frac{m}{n}}.$$

Taking the infimum over all possible choices of ε yields the bound. □

5.2. Clusterable distributions

We now return to the situation described in the introduction to this section and analyze the case where μ is like a sum of Dirac masses. This is the simplest example of where multiscale behavior can occur. We validate the intuition presented above: when μ is approximately discrete, in the sense that it is supported on balls of small radius, then the convergence of $\hat{\mu}_n$ to μ enjoys the fast $n^{-1/2p}$ rate until n is large even if μ is absolutely continuous with respect to the Lebesgue measure. We show that a similar phenomenon occurs when μ is the convolution of a discrete distribution with a small Gaussian, where we show that it is enough that most of the mass of μ is near a discrete distribution, even though the support is unbounded.

Definition 7. A distribution μ is (m, Δ)-clusterable if $\mathrm{supp}(\mu)$ lies in the union of m balls of radius at most Δ.

Intuitively, the measure μ looks like a sum of m Dirac measures at "large scales," with high-dimensional information arriving only when we consider scales smaller than Δ.

Proposition 13. *If μ is (m, Δ) clusterable, then for all $n \leq m(2\Delta)^{-2p}$,*

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \leq \big(9^p + 3\big)\sqrt{\frac{m}{n}}.$$

Proof. Since $\mathrm{supp}(\mu)$ lies in the union of m balls of radius at most Δ, we have $\mathcal{N}_{2\Delta}(\mu) \leq m$. Therefore if $n \leq m(2\Delta)^{-2p}$, then

$$m_n = \inf_{\varepsilon > 0} \max\big\{\mathcal{N}_\varepsilon\big(\mu, \varepsilon^p\big), n\varepsilon^{2p}\big\} \leq \max\big\{\mathcal{N}_{2\Delta}(\mu), n(2\Delta)^{2p}\big\} \leq m,$$

and the claim follows from Proposition 12. □

Proposition 13 does not directly apply to the "Diracs plus Gaussian" case described in the introduction to this section because Gaussians are not compactly supported. Nevertheless, the following simple lemma allows us to reduce to the compactly supported setting.

Lemma 1. *For any probability measure ν, let $\mu = \nu * \mathcal{N}(0, \sigma^2 I_d)$ If $\hat{\mu}_n$ and $\hat{\nu}_n$ are empirical distributions consisting of n i.i.d. samples from μ and ν, respectively, then*

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \le 3^{p-1}\mathbb{E}\big[W_p^p(\nu, \hat{\nu}_n)\big] + 2 \cdot 3^{p-1}\sigma^p(d+2p)^{p/2}.$$

A proof of this lemma appears in the supplement [57]. Proposition 13 and Lemma 1 yield the following bound for the convolution of clusterable distributions with small Gaussians. In particular, since sums of Dirac measures are trivially clusterable, this proposition implies a bound for the "Diracs plus Gaussian" case.

Proposition 14. *Let ν be (m, Δ) clusterable with support lying in a ball of diameter 1 in $(\mathbb{R}^d, \ell_2)$, and let $\mu := \nu * \mathcal{N}(0, \sigma^2 I_d)$ be the convolution of ν with an isotropic Gaussian of variance σ^2. For all $n \le m[\sigma^{-2p}(d+2p)^{-p} \wedge (2\Delta)^{-2p}]$,*

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \le \big(27^p + 3^{p+1}\big)\sqrt{\frac{m}{n}}.$$

In short, this proposition establishes that the $n^{-1/2}$ rate holds in the small-n regime. Since the measure μ is absolutely continuous with respect to the Lebesgue measure, if $d > 2p$, then asymptotically we have

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \gtrsim n^{-p/d},$$

which can be substantially slower than the $n^{-1/2}$ rate.

Proof. Proposition 13 implies that, for all $n \le m(2\Delta)^{-2p}$,

$$\mathbb{E}\big[W_p^p(\nu, \hat{\nu}_n)\big] \le \big(9^p + 3\big)\sqrt{\frac{m}{n}}.$$

Since $n \le m(\sigma\sqrt{d+2p})^{-2p}$ as well, applying Lemma 1 yields

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \le \big(27^p + 3^p\big)\sqrt{\frac{m}{n}} + 2 \cdot 3^p\sigma^p(d+2p)^{p/2} \le \big(27^p + 3^{p+1}\big)\sqrt{\frac{m}{n}}. \qquad \square$$

5.3. Approximately low-dimensional sets

We now broaden considerably to the general case where μ is supported on an approximately low-dimensional set.

Definition 8 (See [52]). For any $S \subseteq X$, the *ε-fattening* of S is

$$S_\varepsilon := \big\{y : D(y, S) \le \varepsilon\big\}.$$

If $S' \subset S_\varepsilon$ for some S, then S' is close to S in the sense that every point of S' is within ε of some point in S. In particular, $S' \subset S_\varepsilon$ if the Hausdorff distance between S' and S is at most ε.

Measures supported on S_ε are "close" to measures supported on S, and if S is low-dimensional, then we obtain correspondingly better finite-sample rates.

Proposition 15. *Suppose* $\operatorname{supp}(\mu) \subseteq S_\varepsilon$ *for some* $\varepsilon > 0$ *and set* S *satisfying*

$$\mathcal{N}_{\varepsilon'}(S) \leq \left(3\varepsilon'\right)^{-d}$$

for all $\varepsilon' \leq 1/27$ *and for some* $d > 2p$. *Then for all* $n \leq (3\varepsilon)^{-d}$,

$$\mathbb{E}\left[W_p^p(\mu, \hat{\mu}_n)\right] \leq C_1 n^{-p/d},$$

where

$$C_1 := 27^p \left(2 + \frac{1}{3^{\frac{d}{2}-p} - 1}\right).$$

In other words, $\hat{\mu}_n$ converges to μ at the $n^{-p/d}$ rate until n is exponentially large in d. In particular, if μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^s$ where $s \gg d$, then $\hat{\mu}_n$ converges to μ much faster initially (at the rate $n^{-p/d}$) than it does in the limit (at the rate $n^{-p/s}$).

Proof. Given any covering of S by balls $B_1, \ldots, B_m$ of diameter ε', the ε-fattenings $(B_1)_\varepsilon, \ldots, (B_m)_\varepsilon$ provide a covering of S_ε by balls of diameter $\varepsilon' + 2\varepsilon$. This implies for all $\varepsilon' \geq \varepsilon$ that

$$\mathcal{N}_{3\varepsilon'}(\mu) \leq \mathcal{N}_{3\varepsilon'}(S_\varepsilon) \leq \mathcal{N}_{\varepsilon'+2\varepsilon}(S_\varepsilon) \leq \mathcal{N}_{\varepsilon'}(S) \leq \left(3\varepsilon'\right)^{-s},$$

and hence that

$$d_{\geq 3\varepsilon}(S_\varepsilon) \leq s.$$

Therefore, if $n \leq (3\varepsilon)^{-d}$, then

$$d_n \leq \max\left\{d_{\geq 3\varepsilon}(\mu), \frac{\log n}{-\log 3\varepsilon}\right\} \leq d.$$

The claim follows from Proposition 10. □

We can also relax the requirement that $\operatorname{supp}(\mu) \subseteq S_\varepsilon$ to the statement that μ is concentrated near S.

Proposition 16. *Let* S *be a set satisfying*

$$\mathcal{N}_\varepsilon(S) \leq (3\varepsilon)^{-d}$$

for all $\varepsilon \le 1/27$, for some $d > 2p$. Suppose there exists a positive constant σ such that μ satisfies

$$\mu(S_\varepsilon) \ge 1 - e^{-\varepsilon^2/2\sigma^2}$$

for all $\varepsilon > 0$. If $p \log \frac{1}{\sigma} \ge \frac{1}{18}$, then for all $n \le (18p\sigma^2 \log \frac{1}{\sigma})^{-d/2}$,

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \le C_1 n^{-p/d},$$

where

$$C_1 := 27^p \left(2 + \frac{1}{3^{\frac{d}{2}-p} - 1}\right).$$

In other words, we again get the fast $n^{-p/d}$ rate until n is of order approximately σ^{-d}.

Proof. Let $\varepsilon = \sqrt{2p\sigma^2 \log \frac{1}{\sigma}}$. As in the proof of Proposition 15, the assumptions imply that

$$d_{\ge 3\varepsilon}(S_\varepsilon) \le d.$$

Since

$$\mu(S_\varepsilon) \ge 1 - e^{-\varepsilon^2/2\sigma^2} = 1 - \sigma^p \ge 1 - (3\varepsilon)^p,$$

we conclude that as long as $n \le (18p\sigma^2 \log \frac{1}{\sigma})^{-d/2}$, then

$$d_n \le \max\left\{ d_{\ge 3\varepsilon}\big(\mu, (3\varepsilon)^p\big), \frac{\log n}{-\log 3\varepsilon} \right\} \le d,$$

and the claim follows from Proposition 10. □

The condition appearing in Proposition 16 is notable because it resembles a well known concentration property which holds for the Gaussian measure [31]. Establishing similar concentration results for other measures is closely connected to many topics in modern geometric probability, such as isoperimetry [36], logarithmic Sobolev inequalities [30], and so-called transportation inequalities, which link Wasserstein distances and the Kullback–Leibler divergence [5]. We note that, since we consider only bounded metric spaces, Proposition 16 cannot be applied directly to Gaussian measures; however, a truncation argument such as the one employed in the proof of Proposition 14 often suffices to reduce to the compactly supported case.

As an example, we now sketch one connection between Proposition 16 and the concentration of measure phenomenon. We require the following fundamental fact.

Proposition 17 (See [31], Proposition 1.3). *Given a function $f : X \to \mathbb{R}$, say that m_f is a* median *of f if*

$$\mathbb{P}\big[f(X) \ge m_f\big] \ge 1/2 \quad \textit{and} \quad \mathbb{P}\big[f(X) \le m_f\big] \ge 1/2.$$

If for all 1*-Lipschitz functions* $f : X \to \mathbb{R}$ *and medians* m_f,

$$\mathbb{P}\big[f(X) \geq m_f + t\big] \leq e^{-t^2/2\sigma^2}, \tag{2}$$

then for any set A *with* $\mu(A) \geq 1/2$,

$$\mu(A_\varepsilon) \geq 1 - e^{-\varepsilon^2/2\sigma^2}. \tag{3}$$

Conversely, if (3) *holds for all sets* A *with* $\mu(A) \geq 1/2$, *then* (2) *holds for any* 1*-Lipschitz function* f *with median* m_f.

Proposition 16 allows us to show that if μ possesses the Lipschitz concentration property described in Proposition 17, then $\hat{\mu}_n$ enjoys a fast rate of convergence to μ for any p, as long as μ assigns a constant fraction of mass to a low-dimensional set.

Proposition 18. *Suppose that* μ *satisfies* (2) *for all* 1*-Lipschitz functions with some* σ *satisfying* $p \log \frac{1}{\sigma} \geq \frac{1}{18}$. *If* S *is a set satisfying*

$$\mathcal{N}_\varepsilon(S) \leq (3\varepsilon)^{-d}$$

for all ε, *for some* $d > 2p$ *and* $\mu(S) \geq 1/2$, *then for all* $n \leq (18p\sigma^2 \log \frac{1}{\sigma})^{-d/2}$,

$$\mathbb{E}\big[W_p^p(\mu, \hat{\mu}_n)\big] \leq C_1 n^{-p/d},$$

where

$$C_1 = 27^p \left(2 + \frac{1}{3^{\frac{d}{2}-p} - 1}\right).$$

Proof. Combine Propositions 16 and 17. □

6. Concentration

In addition to proving bounds on the expected value of the quantity $W_p^p(\mu, \hat{\mu}_n)$, we also show that it concentrates well around its expectation. Previous work [6,8] has sought to obtain tail bounds of the form

$$\mathbb{P}\big[W_p^p(\mu, \hat{\mu}_n) \geq t\big] \leq \psi_n(t),$$

where $\psi_n(t)$ is some function exhibiting sub-Gaussian decay. The results of [8] appear to obtain this rate, but the constants involved depend on n and the ambient dimension of the space in a way that makes the results difficult to interpret.

We follow a different approach, which more clearly emphasizes the dependence of the tail on n and the dimension. The results of Sections 4 and 5, above, yield bounds on the expected value $\mathbb{E}[W_p^p(\mu, \hat{\mu}_n)]$. As we have seen, the convergence of this quantity to 0 may be slow when

the dimension is large. On the other hand, we show below that, as long as X is bounded, the quantity $W_p^p(\mu, \hat{\mu}_n)$ concentrates well around its expectation independent of the dimension. The argument is standard [51] and is significantly easier to obtain than the above bounds on the expected value.

We require the following dual formulation [44,45].

Definition 9. Given a bounded continuous function $f : X \to \mathbb{R}$, the *c-transform* of f (with respect to $D(\cdot,\cdot)^p$) is the function $f^c : X \to \mathbb{R}$ defined by

$$f^c(y) = \sup_{x \in X} \big(f(x) - D(x,y)^p \big).$$

The following claims are standard, and we provide a proof in the supplement [57] for completeness.

Proposition 19 (Kantorovich duality). *Given any pair of probability measures μ and ν on X and any $p \in [1,\infty)$, the following duality holds*:

$$W_p^p(\mu,\nu) = \sup_{f \in \mathcal{C}_b(X)} \mathbb{E}_\mu f - \mathbb{E}_\nu f^c, \tag{4}$$

where the supremum is taken over all bounded continuous functions on X and f^c is the c-transform of f with respect to $D(\cdot,\cdot)^p$. Moreover, if $\operatorname{diam}(X) \le 1$, *then we can take* $0 \le f(x) \le 1$ *for all* $x \in X$.

We then obtain a concentration result via a standard bounded difference argument.

Proposition 20. *For all $n \ge 0$ and $1 \le p < \infty$,*

$$\mathbb{P}\big[W_p^p(\mu,\hat{\mu}_n) \ge \mathbb{E} W_p^p(\mu,\hat{\mu}_n) + t \big] \le \exp\big(-2nt^2\big).$$

Proof. Let $\hat{\mu}_n$ be the empirical distribution corresponding to the i.i.d. samples $X_1,\dots,X_n \sim \mu$. We abbreviate $W_p^p(\mu,\hat{\mu}_n)$ by W. By Proposition 19, we can write

$$W = \sup_{0 \le f \le 1} \mathbb{E}_{\hat{\mu}_n} f - \mathbb{E}_\mu f^c,$$

or, writing W explicitly as a function of $X_1,\dots,X_n$,

$$W(X_1,\dots,X_n) = \frac{1}{n} \sup_{0 \le f \le 1} \sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f^c.$$

For any $x_1, \dots, x_n, x_n' \in X$, we have

$$W(x_1, \dots, x_n) - W\big(x_1, \dots, x_n'\big) = \frac{1}{n}\left\{ \sup_{0 \le f \le 1} \sum_{i=1}^{n} \big(f(x_i) - \mathbb{E}_\mu f^c\big) \right.$$
$$\left. - \sup_{0 \le f' \le 1} \sum_{i=1}^{n-1} \big(f'(x_i) - \mathbb{E}_\mu f'^c\big) + f'\big(x_n'\big) - \mathbb{E}_\mu f'^c \right\}$$
$$\le \frac{1}{n} \sup_{0 \le f \le 1} f(x_n) - f\big(x_n'\big) \le \frac{1}{n}.$$

Applying McDiarmid's inequality [35] yields the bound. □

7. Applications

In this section, we sketch two applications of our work to machine learning and statistics.

7.1. Quadrature

Numerical integration, or *quadrature*, refers to the technique of approximating integrals by finite sums for the purpose of evaluating them at low computational cost. Given a measure μ, the goal is to choose points $x_1, \dots, x_n \in X$ and weights $\alpha_1, \dots, \alpha_n \in \mathbb{R}_+$ such that the approximation

$$\int f(x)\, \mathrm{d}\mu(x) \approx \sum_{i=1}^{n} \alpha_i f(x_i)$$

is as good as possible for a wide class of functions f. This problem possesses close connections to optimal quantization, since the points $x_1, \dots, x_n$ naturally serve as a finite approximation to the underlying measure [22,40].

When only a single function f is considered, one can show that a Monte Carlo method which chooses $x_1, \dots, x_n$ i.i.d. from μ with uniform weights $\alpha_i := n^{-1}$ for $1 \le i \le n$ is asymptotically suboptimal [39]. However, our results show that if we require that the approximation hold over the class of Lipschitz functions $\mathrm{Lip}(X)$, then this simple Monte Carlo scheme is asymptotically optimal for a wide class of measures.

Proposition 21. *Denote by* $\mathrm{Lip}(X)$ *the class of* 1*-Lipschitz on* X *If* μ *is a measure supported on a regular set of dimension* $d \ge 2$ *and* $\mu \ll \mathcal{H}^d$*, then for any* $s > d$*,*

$$\mathbb{E} \sup_{f \in \mathrm{Lip}(X)} \left| \int f(x)\, \mathrm{d}\mu(x) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| \lesssim n^{-1/s},$$

where $X_1, \dots, X_n \sim \mu$ are independent. On the other hand, for any $t < d$, $x_1, \dots, x_n \in X$, and $\alpha_i, \dots, \alpha_n \in \mathbb{R}_+$,

$$\sup_{f \in \mathrm{Lip}(X)} \left| \int f(x)\, \mathrm{d}\mu(x) - \sum_{i=1}^{n} \alpha_i f(x_i) \right| \gtrsim n^{-1/t}.$$

Proof. Recalling (1), we immediately see that the first claim corresponds to the fact that $\mathbb{E} W_1(\mu, \hat{\mu}_n) \lesssim n^{-1/s}$, which follows from Corollary 1 and Proposition 8.

For the second claim, we first note that by choosing $f(x) := c$ to be a constant function, we have

$$\left| \int f(x)\, \mathrm{d}\mu(x) - \sum_{i=1}^{n} \alpha_i f(x_i) \right| = c \left| 1 - \sum_{i=1}^{n} \alpha_i \right|.$$

Since such an f is Lipschitz, by choosing c arbitrarily large we obtain that

$$\sup_{f \in \mathrm{Lip}(X)} \left| \int f(x)\, \mathrm{d}\mu(x) - \sum_{i=1}^{n} \alpha_i f(x_i) \right| = \infty$$

unless $\sum_{i=1}^{n} \alpha_i = 1$. It therefore suffices to prove the claim when $\sum_{i=1}^{n} \alpha_i = 1$. In this case, setting

$$\nu := \sum_{i=1}^{n} \alpha_i \delta_{x_i},$$

and again applying (1) yields that this claim is equivalent to the fact that $W_1(\mu, \nu) \gtrsim n^{-1/t}$. Since ν is supported on at most n points, this follows from Corollary 2 and Proposition 8. □

7.2. *k*-means clustering

The authors of [10] point out that many "unsupervised learning" techniques in machine learning involve constructing a simple approximation $\tilde{\mu}$ to a measure μ such that $W_2(\mu, \tilde{\mu})$ is small. One such example is the so-called k-means problem, where the goal is to find a set S with $|S| \le k$ minimizing the objective function

$$\mathbb{E} D(X, S)^2,$$

where $X \sim \mu$. It is not hard to see [10], Lemma 3.1, that this problem is equivalent to finding a measure $\tilde{\mu}$ supported on at most k points such that $W_2(\mu, \tilde{\mu})$ is as small as possible. Given such a measure, we obtain a clustering of μ into at most k pieces by constructing a Voronoi partition of $\mathrm{supp}(\mu)$ based on the k points in $\mathrm{supp}(\tilde{\mu})$.

The authors of [10] show that for k sufficiently large and for X a compact, smooth d-dimensional manifold, it is possible to find a measure $\tilde{\mu}$ with $|\mathrm{supp}(\tilde{\mu})| \le k$ satisfying

$$W_2(\mu, \tilde{\mu}) \le C_1 \tau k^{-1/d} \qquad \text{with probability } 1 - e^{-\tau^2}$$

on the basis of $C_2 k^{2+\frac{4}{d}}$ samples. Corollary 2 implies that this dependence on k is asymptotically optimal.

Our results show that a much simpler procedure suffices in high dimensions. As long as $d \geq 4$, the empirical measure $\hat{\mu}_k$ satisfies

$$\mathbb{E}W_2(\mu, \hat{\mu}_k) \lesssim k^{-1/s}$$

for any $s > d$, and Proposition 20 then implies that there exist universal constants C and C' such that

$$W_2(\mu, \hat{\mu}_k) \leq C\tau k^{-1/s} \qquad \text{with probability } 1 - C'e^{-\tau^4}.$$

This shows that clustering a measure μ into k pieces on the basis of k i.i.d. samples from μ is asymptotically optimal, and enjoys concentration properties even better than the ones implied by [10].

8. Conclusion and future work

Our focus in this work has been to obtain sharper rates than previously available for the convergence of $\hat{\mu}_n$ to $\hat{\mu}$ in Wasserstein distance, both in asymptotic and finite-sample settings. Our results give theoretical support to a phenomenon observed in practice: even though $W_p(\mu, \hat{\mu}_n)$ can converge very slowly for measures supported on a high-dimensional metric space, many measures arising in applications are intrinsically low dimensional, at least approximately, and therefore enjoy reasonably fast rates of convergence.

Our work leaves open whether slightly different versions of the Wasserstein distance can converge faster in general. Recently, a version of the Wasserstein distance with an entropic penalty has been proposed and shown to have attractive theoretical properties and practical performance [11,14,42,49]. It is possible that these objects achieve better rates than the vanilla Wasserstein distance in the high-dimensional setting.

We also do not consider here empirical measures other than the simple $\hat{\mu}_n$. In practice, a technique known as importance sampling [9] is often used to reduce the variance of estimates produced on the basis of random samples from a distribution. As noted in Section 7.1, if μ is sufficiently regular, then no discrete measure on n points can achieve better asymptotic performance than the empirical measure $\hat{\mu}_n$. However, we conjecture that many reasonable sampling techniques should produce measures that are also asymptotically no worse than $\hat{\mu}_n$. We leave this question for future work.

Acknowledgements

JW and FB acknowledge support from the Chaire Économie des nouvelles données, with the data science Joint Research Initiative with the Fonds AXA pour la recherche, and the Initiative de Recherche "Machine Learning for Large-Scale Insurance" from the Institut Louis Bachelier. JW acknowledges support from NSF Graduate Research Fellowship 1122374 and FB's hospitality at INRIA, where this research was conducted.

We thank Guillaume Carlier, Marco Cuturi, and Gabriel Peyré for discussions related to this work, and we thank the anonymous reviewers for suggesting several references and improvements.

Supplementary Material

Supplement to "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance" (DOI: 10.3150/18-BEJ1065SUPP; .pdf). Additional proofs and technical lemmas.

References

[1] Ajtai, M., Komlós, J. and Tusnády, G. (1984). On optimal matchings. *Combinatorica* **4** 259–264. MR0779885

[2] Ba, K.D., Nguyen, H.L., Nguyen, H.N. and Rubinfeld, R. (2011). Sublinear time algorithms for Earth mover's distance. *Theory Comput. Syst.* **48** 428–442. MR2763110

[3] Bellman, R. (1961). *Adaptive Control Processes*: *A Guided Tour*. Princeton, NJ: Princeton Univ. Press. MR0134403

[4] Berend, D. and Kontorovich, A. (2013). A sharp estimate of the binomial mean absolute deviation with applications. *Statist. Probab. Lett.* **83** 1254–1259. MR3041401

[5] Bobkov, S.G. and Götze, F. (1999). Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal.* **163** 1–28. MR1682772

[6] Boissard, E. (2011). Simple bounds for convergence of empirical and occupation measures in 1-Wasserstein distance. *Electron. J. Probab.* **16** 2296–2333. MR2861675

[7] 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

[8] Bolley, F., Guillin, A. and Villani, C. (2007). Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probab. Theory Related Fields* **137** 541–593. MR2280433

[9] Bucklew, J.A. (2013). *Introduction to Rare Event Simulation*. *Springer Series in Statistics*. New York: Springer. MR2045385

[10] Cañas, G.D. and Rosasco, L. (2012). Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems* 25: 26*th Annual Conference on Neural Information Processing Systems* 2012. *Proceedings of a Meeting Held December* 3–6, 2012, *Lake Tahoe*, *Nevada*, *United States* (P.L. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou and K.Q. Weinberger, eds.) 2501–2509.

[11] Carlier, G., Duval, V., Peyré, G. and Schmitzer, B. (2017). Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.* **49** 1385–1418. MR3635459

[12] Corlay, S. and Pagès, G. (2015). Functional quantization-based stratified sampling methods. *Monte Carlo Methods Appl.* **21** 1–32. MR3318550

[13] Cover, T.M. and Thomas, J.A. (2012). *Elements of Information Theory*. *Wiley Series in Telecommunications*. New York: Wiley. MR1122806

[14] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems* 26: 27*th Annual Conference on Neural Information Processing Systems* 2013. *Proceedings of a Meeting Held December* 5–8, 2013, *Lake Tahoe*, *Nevada*, *United States* (C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds.) 2292–2300.

[15] David, G. (1988). Morceaux de graphes lipschitziens et intégrales singulières sur une surface. *Rev. Mat. Iberoam.* **4** 73–114. MR1009120
[16] Dereich, S., Scheutzow, M. and Schottstedt, R. (2013). Constructive quantization: Approximation by empirical measures. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 1183–1203. MR3127919
[17] Dobrić, V. and Yukich, J.E. (1995). Asymptotics for transportation cost in high dimensions. *J. Theoret. Probab.* **8** 97–118. MR1308672
[18] Dudley, R.M. (1968). The speed of mean Glivenko–Cantelli convergence. *Ann. Math. Stat.* **40** 40–50. MR0236977
[19] Falconer, K. (2003). *Fractal Geometry*: *Mathematical Foundations and Applications*, 2nd ed. Hoboken, NJ: Wiley. MR2118797
[20] Fortet, R. and Mourier, E. (1953). Convergence de la répartition empirique vers la répartition théorique. *Ann. Sci. Éc. Norm. Supér.* (3) **70** 267–285. MR0061325
[21] 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
[22] Graf, S. and Luschgy, H. (2007). *Foundations of Quantization for Probability Distributions*. *Lecture Notes in Math.* **1730**. Berlin: Springer. MR1764176
[23] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. *Springer Series in Statistics*. New York: Springer. MR1851606
[24] Indyk, P. and Thaper, N. (2003). Fast image retrieval via embeddings. In 3*rd International Workshop on Statistical and Computational Theories of Vision*. ICCV.
[25] Kaimanovich, V.A. and Le Prince, V. (2011). Matrix random products with singular harmonic measure. *Geom. Dedicata* **150** 257–279. MR2753707
[26] Kantorovič, L.V. and Rubinšteĭn, G.Š. (1958). On a space of completely additive functions. *Vestn. Leningrad Univ.* **13** 52–59. MR0102006
[27] Kantorovitch, L. (1942). On the translocation of masses. *C. R.* (*Dokl.*) *Acad. Sci. URSS* **37** 199–201. MR0009619
[28] Kloeckner, B. (2012). Approximation by finitely supported measures. *ESAIM Control Optim. Calc. Var.* **18** 343–359. MR2954629
[29] Kusner, M.J., Sun, Y., Kolkin, N.I. and Weinberger, K.Q. (2015). From word embeddings to document distances. In *Proceedings of the* 32*nd International Conference on Machine Learning*, *ICML* 2015, *Lille*, *France*, 6–11 *July* 2015 (F.R. Bach and D.M. Blei, eds.). *JMLR Workshop and Conference Proceedings* **37** 957–966. JMLR.org.
[30] Ledoux, M. (1995). Remarks on logarithmic Sobolev constants, exponential integrability and bounds on the diameter. *J. Math. Kyoto Univ.* **35** 211–220. MR1346225
[31] Ledoux, M. (2005). *The Concentration of Measure Phenomenon*. *Mathematical Surveys and Monographs* **89**. Providence, RI: Amer. Math. Soc. MR1849347
[32] Ledrappier, F. (1981). Some relations between dimension and Lyapounov exponents. *Comm. Math. Phys.* **81** 229–238. MR0632758
[33] Little, A.V., Maggioni, M. and Rosasco, L. (2017). Multiscale geometric methods for data sets I: Multiscale SVD, noise and curvature. *Appl. Comput. Harmon. Anal.* **43** 504–567. MR3683673
[34] Mattila, P. (1999). *Geometry of Sets and Measures in Euclidean Spaces*: *Fractals and Rectifiability*. *Cambridge Studies in Advanced Mathematics* **44**. Cambridge: Cambridge Univ. Press. MR1333890
[35] McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, 1989 (*Norwich*, 1989). *London Mathematical Society Lecture Note Series* **141** 148–188. Cambridge: Cambridge Univ. Press. MR1036755
[36] Milman, E. (2010). Isoperimetric and concentration inequalities: Equivalence under curvature lower bound. *Duke Math. J.* **154** 207–239. MR2682183

[37] Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Hist. Acad. R. Sci.* **1** 666–704.
[38] Nova, D. and Estévez, P.A. (2014). A review of learning vector quantization classifiers. *Neural Comput. Appl.* **25** 511–524.
[39] Novak, E. (1988). *Deterministic and Stochastic Error Bounds in Numerical Analysis*. *Lecture Notes in Math.* **1349**. Berlin: Springer. MR0971255
[40] Pagès, G. (1998). A space quantization method for numerical integration. *J. Comput. Appl. Math.* **89** 1–38. MR1625987
[41] Posner, E.C., Rodemich, E.R. and Rumsey, H. Jr. (1967). Epsilon entropy of stochastic processes. *Ann. Math. Stat.* **38** 1000–1020. MR0211457
[42] 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*, *AISTATS* 2016, *Cadiz*, *Spain*, *May* 9–11, 2016 (A. Gretton and C.C. Robert, eds.). *JMLR Workshop and Conference Proceedings* **51** 630–638. JMLR.org.
[43] 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.
[44] Rüschendorf, L. (1991). Fréchet-bounds and their applications. In *Advances in Probability Distributions with Given Marginals* (*Rome*, 1990). *Math. Appl.* **67** 151–187. Dordrecht: Kluwer Academic. MR1215951
[45] Rüschendorf, L. and Rachev, S.T. (1990). A characterization of random variables with minimum L^2-distance. *J. Multivariate Anal.* **32** 48–54. MR1035606
[46] Sandler, R. and Lindenbaum, M. (2011). Nonnegative matrix factorization with Earth mover's distance metric for image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 1590–1602.
[47] Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*: *Calculus of Variations*, *PDEs*, *and Modeling*. *Progress in Nonlinear Differential Equations and Their Applications* **87**. Cham: Birkhäuser/Springer. MR3409718
[48] Shannon, C.E. (1960). Coding theorems for a discrete source with a fidelity criterion. In *Information and Decision Processes* 93–126. New York: McGraw-Hill. MR0122612
[49] Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T. and Guibas, L.J. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.* **34** 66:1–66:11. DOI:10.1145/2766963.
[50] Starck, J.-L., Murtagh, F. and Bijaoui, A. (1998). *Image Processing and Data Analysis*: *The Multiscale Approach*. Cambridge: Cambridge Univ. Press. MR1637482
[51] Talagrand, M. (1992). Matching random samples in many dimensions. *Ann. Appl. Probab.* **2** 846–856. MR1189420
[52] Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. Inst. Hautes Études Sci.* **81** 73–205. MR1361756
[53] Varadarajan, V.S. (1958). On the convergence of sample probability distributions. *Sankhyā* **19** 23–26. MR0094839
[54] Vershik, A.M. (2013). Long history of the Monge–Kantorovich transportation problem. *Math. Intelligencer* **35** 1–9. MR3133757
[55] Villani, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Berlin: Springer. MR2459454
[56] Wakin, M., Donoho, D., Choi, H. and Baraniuk, R.G. (2005). The multiscale structure of non-differentiable image manifolds. In *Proc. SPIE* SPIE.
[57] Weed, J. and Bach, F. (2018). Supplement to "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance." DOI:10.3150/18-BEJ1065SUPP.

[58] Young, L.S. (1982). Dimension, entropy and Lyapunov exponents. *Ergod. Theory Dyn. Syst.* **2** 109–124. MR0684248

[59] Zhang, M., Liu, Y., Luan, H., Sun, M., Izuha, T. and Hao, J. (2016). Building Earth mover's distance on bilingual word embeddings for machine translation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, *February* 12–17, 2016, *Phoenix*, *Arizona*, *USA* (D. Schuurmans and M.P. Wellman, eds.) 2870–2876. AAAI Press.

Received August 2017 and revised May 2018