Learning Probability Measures with Respect to Optimal Transport Metrics

Guillermo D. Canas*,† **Lorenzo A. Rosasco***,†
$\star$ Laboratory for Computational and Statistical Learning - MIT-IIT
† CBCL, McGovern Institute - Massachusetts Institute of Technology
{guilledc,lrosasco}@mit.edu

Abstract

We study the problem of estimating, in the sense of optimal transport metrics, a measure which is assumed supported on a manifold embedded in a Hilbert space. By establishing a precise connection between optimal transport metrics, optimal quantization, and learning theory, we derive new probabilistic bounds for the performance of a classic algorithm in unsupervised learning (k-means), when used to produce a probability measure derived from the data. In the course of the analysis, we arrive at new lower bounds, as well as probabilistic upper bounds on the convergence rate of empirical to population measures, which, unlike existing bounds, are applicable to a wide class of measures.

1 Introduction and Motivation

In this paper we study the problem of learning from random samples a probability distribution supported on a manifold, when the learning error is measured using transportation metrics.

The problem of learning a probability distribution is classic in statistics, and is typically analyzed for distributions in $\mathcal{X} = \mathbb{R}^d$ that have a density with respect to the Lebesgue measure, with total variation, and L_2 among the common distances used to measure closeness of two densities (see for instance [10, 32] and references therein.) The setting in which the data distribution is supported on a low dimensional manifold embedded in a high dimensional space has only been considered more recently. In particular, kernel density estimators on manifolds have been described in [36], and their pointwise consistency, as well as convergence rates, have been studied in [25, 23, 18]. A discussion on several topics related to statistics on a Riemannian manifold can be found in [26].

Interestingly, the problem of approximating measures with respect to transportation distances has deep connections with the fields of optimal quantization [14, 16], optimal transport [35] and, as we point out in this work, with unsupervised learning (see Sec. 4.) In fact, as described in the sequel, some of the most widely-used algorithms for unsupervised learning, such as k-means (but also others such as PCA and k-flats), can be shown to be performing exactly the task of estimating the data-generating measure in the sense of the 2-Wasserstein distance. This close relation between learning theory, and optimal transport and quantization seems novel and of interest in its own right. Indeed, in this work, techniques from the above three fields are used to derive the new probabilistic bounds described below.

Our technical contribution can be summarized as follows:

(a) we prove uniform lower bounds for the distance between a measure and estimates based on discrete sets (such as the empirical measure or measures derived from algorithms such as k-means);

(b) we provide new probabilistic bounds for the rate of convergence of empirical to population measures which, unlike existing probabilistic bounds, hold for a very large class of measures;

(c) we provide probabilistic bounds for the rate of convergence of measures derived from k-means to the data measure.

The structure of the paper is described at the end of Section 2, where we discuss the exact formulation of the problem as well as related previous works.

2 Setup and Previous work

Consider the problem of learning a probability measure ρ supported on a space $\mathcal{M}$, from an i.i.d. sample $X_n = (x_1, \ldots, x_n) \sim \rho^n$ of size n. We assume $\mathcal{M}$ to be a compact, smooth d-dimensional manifold of bounded curvature, with $\mathcal{C}^1$ metric and volume measure $\lambda_{\mathcal{M}}$, embedded in the unit ball of a separable Hilbert space $\mathcal{X}$ with inner product $\langle \cdot, \cdot \rangle$, induced norm $\| \cdot \|$, and distance d (for instance $\mathcal{M} = B_2^d(1)$ the unit ball in $\mathcal{X} = \mathbb{R}^d$.) Following [35, p. 94], let $P_p(\mathcal{M})$ denote the Wasserstein space of order $1 \leq p < \infty$:

$$P_p(\mathcal{M}) := \left\{ \rho \in P(\mathcal{M}) : \int_{\mathcal{M}} \|x\|^p d\rho(x) < \infty \right\}$$

of probability measures $P(\mathcal{M})$ supported on $\mathcal{M}$, with finite p-th moment. The p-Wasserstein distance

$$W_p(\rho, \mu) = \inf_{X,Y} \left\{ [\mathbb{E}\|X - Y\|^p]^{1/p} : \text{Law}(X) = \rho, \text{Law}(Y) = \mu \right\} \tag{1}$$

where the random variables X and Y are distributed according to ρ and μ respectively, is the optimal expected cost of transporting points generated from ρ to those generated from μ, and is guaranteed to be finite in $P_p(\mathcal{M})$ [35, p. 95]. The space $P_p(\mathcal{M})$ with the W_p metric is itself a complete separable metric space [35]. We consider here the problem of learning probability measures $\rho \in P_2(\mathcal{M})$, where the performance is measured by the distance W_2.

There are many possible choices of distances between probability measures [13]. Among them, W_p metrizes weak convergence (see [35] theorem 6.9), that is, in $P_p(\mathcal{M})$, a sequence $(\mu_i)_{i \in \mathbb{N}}$ of measures converges weakly to μ iff $W_p(\mu_i, \mu) \to 0$ and their p-th order moments converge to that of μ. There are other distances, such as the Lévy-Prokhorov, or the weak-* distance, that also metrize weak convergence. However, as pointed out by Villani in his excellent monograph [35, p. 98],

1. "Wasserstein distances are rather strong, [...]a definite advantage over the weak-* distance".
2. "It is not so difficult to combine information on convergence in Wasserstein distance with some smoothness bound, in order to get convergence in stronger distances."

Wasserstein distances have been used to study the mixing and convergence of Markov chains [22], as well as concentration of measure phenomena [20]. To this list we would add the important fact that existing and widely-used algorithms for unsupervised learning can be easily extended (see Sec. 4) to compute a measure ρ' that minimizes the distance $W_2(\hat{\rho}_n, \rho')$ to the empirical measure

$$\hat{\rho}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i},$$

a fact that will allow us to prove, in Sec. 5, bounds on the convergence of a measure induced by k-means to the population measure ρ.

The most useful versions of Wasserstein distance are $p = 1, 2$, with $p = 1$ being the weaker of the two (by Hölder's inequality, $p \leq q \Rightarrow W_p \leq W_q$.) In particular, "results in W_2 distance are usually stronger, and more difficult to establish than results in W_1 distance" [35, p. 95]. A discussion of $p = \infty$ would take us out of topic, since its behavior is markedly different.

2.1 Closeness of Empirical and Population Measures

By the strong law of large numbers, the empirical measure converges almost surely to the population measure: $\hat{\rho}_n \to \rho$ in the sense of the weak topology [34]. Since weak convergence and convergence in W_p plus convergence of p-th moments are equivalent in $P_p(\mathcal{M})$, this means that, in the W_p sense, the empirical measure $\hat{\rho}_n$ converges to ρ, as $n \to \infty$. A fundamental question is therefore how fast the rate of convergence of $\hat{\rho}_n \to \rho$ is.

2.1.1 Convergence in expectation

The rate of convergence of $\hat{\rho}_n \to \rho$ *in expectation* has been widely studied in the past, resulting in upper bounds of order $\mathbb{E}W_2(\rho, \hat{\rho}_n) = O(n^{-1/(d+2)})$ [19, 8], and lower bounds of order $\mathbb{E}W_2(\rho, \hat{\rho}_n) = \Omega(n^{-1/d})$ [29] (both assuming that the absolutely continuous part of ρ is $\rho_A \neq 0$, with possibly better rates otherwise).

More recently, an upper bound of order $\mathbb{E}W_p(\rho, \hat{\rho}_n) = O(n^{-1/d})$ has been proposed [2] by proving a bound for the Optimal Bipartite Matching (OBM) problem [1], and relating this problem to the expected distance $\mathbb{E}W_p(\rho, \hat{\rho}_n)$. In particular, given two independent samples X_n, Y_n, the OBM problem is that of finding a permutation σ that minimizes the matching cost $n^{-1} \sum \|x_i - y_{\sigma(i)}\|^p$ [24, 30]. It is not hard to show that the optimal matching cost is $W_p(\hat{\rho}_{X_n}, \hat{\rho}_{Y_n})^p$, where $\hat{\rho}_{X_n}, \hat{\rho}_{Y_n}$ are the empirical measures associated to X_n, Y_n. By Jensen's inequality, the triangle inequality, and $(a+b)^p \leq 2^{p-1}(a^p + b^p)$, it holds

$$\mathbb{E}W_p(\rho, \hat{\rho}_n)^p \leq \mathbb{E}W_p(\hat{\rho}_{X_n}, \hat{\rho}_{Y_n})^p \leq 2^{p-1}\mathbb{E}W_p(\rho, \hat{\rho}_n)^p,$$

and therefore a bound of order $O(n^{-p/d})$ for the OBM problem [2] implies a bound $\mathbb{E}W_p(\rho, \hat{\rho}_n) = O(n^{-1/d})$. The matching lower bound is only known for a special case: ρ_A constant over a bounded set of non-null measure [2] (e.g. ρ_A uniform.) Similar results, with matching lower bounds are found for W_1 in [11].

2.1.2 Convergence in probability

Results for convergence in probability, one of the main results of this work, appear to be considerably harder to obtain. One fruitful avenue of analysis has been the use of so-called *transportation*, or *Talagrand inequalities* T_p, which can be used to prove concentration inequalities on W_p [20]. In particular, we say that ρ satisfies a $T_p(C)$ inequality with $C > 0$ iff $W_p(\rho, \mu)^2 \leq CH(\mu|\rho), \forall \mu \in P_p(\mathcal{M})$, where $H(\cdot|\cdot)$ is the relative entropy [20]. As shown in [6, 5], it is possible to obtain probabilistic upper bounds on $W_p(\rho, \hat{\rho}_n)$, with $p = 1, 2$, if ρ is known to satisfy a T_p inequality of the same order, thereby reducing the problem of bounding $W_p(\rho, \hat{\rho}_n)$ to that of obtaining a T_p inequality. Note that, by Jensen's inequality, and as expected from the behavior of W_p, the inequality T_2 is stronger than T_1 [20].

While it has been shown that ρ satisfies a T_1 inequality iff it has a finite square-exponential moment ($\mathbb{E}[e^{\alpha\|x\|^2}]$ finite for some $\alpha > 0$) [4, 7], no such general conditions have been found for T_2. As an example, consider that, if $\mathcal{M}$ is compact with diameter D then, by theorem 6.15 of [35], and the celebrated Csiszár-Kullback-Pinsker inequality [27], for all $\rho, \mu \in P_p(\mathcal{M})$, it is

$$W_p(\rho, \mu)^{2p} \leq (2D)^{2p}\|\rho - \mu\|_{\text{TV}}^2 \leq 2^{2p-1}D^{2p}H(\mu|\rho),$$

where $\|\cdot\|_{\text{TV}}$ is the total variation norm. Clearly, this implies a $T_{p=1}$ inequality, but for $p \geq 2$ it does not.

The T_2 inequality has been shown by Talagrand to be satisfied by the Gaussian distribution [31], and then slightly more generally by strictly log-concave measures (see [20, p. 123], and [3].) However, as noted in [6], "contrary to the T_1 case, there is no hope to obtain T_2 inequalities from just integrability or decay estimates."

Structure of this paper. In this work we obtain bounds in probability (learning rates) for the problem of learning a probability measure in the sense of W_2. We begin by establishing (lower) bounds for the convergence of empirical to population measures, which serve to set up the problem and introduce the connection between quantization and measure learning (sec. 3.) We then describe how existing unsupervised learning algorithms that compute a set (k-means, k-flats, PCA,...) can be easily extended to produce a measure (sec. 4.) Due to its simplicity and widespread use, we focus here on k-means. Since the two measure estimates that we consider are the empirical measure, and the measure induced by k-means, we next set out to prove upper bounds on their convergence to the data-generating measure (sec. 5.) We arrive at these bounds by means of intermediate measures, which are related to the problem of optimal quantization. The bounds apply in a very broad setting (unlike existing bounds based on transportation inequalities, they are not restricted to log-concave measures [20, 3].)

3 Learning probability measures, optimal transport and quantization

We address the problem of learning a probability measure ρ when the only observations we have at our disposal are n i.i.d. samples $X_n = (x_1, \ldots, x_n)$. We begin by establishing some notation and useful intermediate results.

Given a closed set $S \subseteq \mathcal{X}$, let $\{V_q : q \in S\}$ be a Borel Voronoi partition of $\mathcal{X}$ composed of sets V_q closest to each $q \in S$, that is, such that each $V_q \subseteq \{x \in \mathcal{X} : \|x - q\| = \min_{r \in S} \|x - r\|\}$ is measurable (see for instance [15].) Consider the projection function $\pi_{_S} : \mathcal{X} \to S$ mapping each $x \in V_q$ to q. By virtue of $\{V_q\}_{q \in S}$ being a Borel Voronoi partition, the map $\pi_{_S}$ is measurable [15], and it is $d(x, \pi_{_S}(x)) = \min_{q \in S} \|x - q\|$ for all $x \in \mathcal{X}$.

For any $\rho \in P_p(\mathcal{M})$, let $\pi_{_S}\rho$ be the pushforward, or image measure of ρ under the mapping $\pi_{_S}$, which is defined to be $(\pi_{_S}\rho)(A) := \rho(\pi_{_S}^{-1}(A))$ for all Borel measurable sets A. From its definition, it is clear that $\pi_{_S}\rho$ is supported on S.

We now establish a connection between the expected distance to a set S, and the distance between ρ and the set's induced pushforward measure. Notice that, for discrete sets S, the expected L_p distance to S is exactly the expected quantization error

$$\mathcal{E}_{p,\rho}(S) := \mathbb{E}_{x \sim \rho} d(x, S)^p = \mathbb{E}_{x \sim \rho} \|x - \pi_{_S}(x)\|^p$$

incurred when encoding points x drawn from ρ by their closest point $\pi_{_S}(x)$ in S [14]. This close connection between optimal quantization and Wasserstein distance has been pointed out in the past in the statistics [28], optimal quantization [14, p. 33], and approximation theory [16] literatures.

The following two lemmas are key tools in the reminder of the paper. The first highlights the close link between quantization and optimal transport.

Lemma 3.1. *For closed $S \subseteq \mathcal{X}$, $\rho \in P_p(\mathcal{M})$, $1 \leq p < \infty$, it holds $\mathbb{E}_{x \sim \rho} d(x, S)^p = W_p(\rho, \pi_{_S}\rho)^p$.*

Note that the key element in the above lemma is that the two measures in the expression $W_p(\rho, \pi_{_S}\rho)$ must match. When there is a mismatch, the distance can only increase. That is, $W_p(\rho, \pi_{_S}\mu) \geq W_p(\rho, \pi_{_S}\rho)$ for all $\mu \in P_p(\mathcal{M})$. In fact, the following lemma shows that, among all the measures with support in S, $\pi_{_S}\rho$ is closest to ρ.

Lemma 3.2. *For closed $S \subseteq \mathcal{X}$, and all $\mu \in P_p(\mathcal{M})$ with $\operatorname{supp}(\mu) \subseteq S$, $1 \leq p < \infty$, it holds $W_p(\rho, \mu) \geq W_p(\rho, \pi_{_S}\rho)$.*

When combined, lemmas 3.1 and 3.2 indicate that the behavior of the measure learning problem is limited by the performance of the optimal quantization problem. For instance, $W_p(\rho, \hat{\rho}_n)$ can only be, in the best-case, as low as the optimal quantization cost with codebook of size n. The following section makes this claim precise.

3.1 Lower bounds

Consider the situation depicted in fig. 1, in which a sample $X_4 = \{x_1, x_2, x_3, x_4\}$ is drawn from a distribution ρ which we assume here to be absolutely continuous on its support. As shown, the projection map $\pi_{_{X_4}}$ sends points x to their closest point in X_4. The resulting Voronoi decomposition of $\operatorname{supp}(\rho)$ is drawn in shades of blue. By lemma 5.2 of [9], the pairwise intersections of Voronoi regions have null ambient measure, and since ρ is absolutely continuous, the pushforward measure can be written in this case as $\pi_{_{X_4}}\rho = \sum_{j=1}^{4} \rho(V_{x_j})\delta_{x_j}$, where V_{x_j} is the Voronoi region of x_j. Note that, even for finite sets S, this particular decomposition is not always possible if the $\{V_q\}_{q \in S}$ form a Borel Voronoi tiling, instead of a Borel Voronoi partition. If, for instance, ρ has an atom falling on two Voronoi regions in a tiling, then both regions would count the atom as theirs, and double-counting would imply $\sum_q \rho(V_q) > 1$. The technicalities required to correctly define a Borel Voronoi partition are such that, in general, it is simpler to write $\pi_S\rho$, even though (if S is discrete) this measure can clearly be written as a sum of deltas with appropriate masses.

By lemma 3.1, the distance $W_p(\rho, \pi_{_{X_4}}\rho)^p$ is the (expected) quantization cost of ρ when using X_4 as codebook. Clearly, this cost can never be lower than the *optimal* quantization cost of size 4. This reasoning leads to the following lower bound between empirical and population measures.

Theorem 3.3. *For $\rho \in P_p(\mathcal{M})$ with absolutely continuous part $\rho_A \neq 0$, and $1 \leq p < \infty$, it holds $W_p(\rho, \hat{\rho}_n) = \Omega(n^{-1/d})$ uniformly over $\hat{\rho}_n$, where the constants depend on d and ρ_A only.*

Proof: Let $V_{n,p}(\rho) := \inf_{S \subset \mathcal{M}, |S|=n} \mathbb{E}_{x\sim\rho} d(x, S)^p$ be the optimal quantization cost of ρ of order p with n centers. Since $\rho_A \neq 0$, and since ρ has a finite $(p+\delta)$-th order moment, for some $\delta > 0$ (since it is supported on the unit ball), then it is $V_{n,p}(\rho) = \Theta(n^{-p/d})$, with constants depending on d and ρ_A (see [14, p. 78] and [16].) Since $\operatorname{supp}(\hat{\rho}_n) = X_n$, it follows that

$$W_p(\rho, \hat{\rho}_n)^p \underset{\text{lemma 3.2}}{\geq} W_p(\rho, \pi_{X_n}\rho)^p \underset{\text{lemma 3.1}}{=} \mathbb{E}_{x\sim\rho} d(x, X_n)^p \geq V_{n,p}(\rho) = \Theta(n^{-p/d}) \quad \square$$

Note that the bound of theorem 3.3 holds for $\hat{\rho}_n$ derived from *any* sample X_n, and is therefore stronger than the existing lower bounds on the convergence rates of $\mathbb{E}W_p(\rho, \hat{\rho}_n) \to 0$. In particular, it trivially induces the known lower bound $\Omega(n^{-1/d})$ on the rate of convergence in expectation.

4 Unsupervised learning algorithms for learning a probability measure

As described in [21], several of the most widely used unsupervised learning algorithms can be interpreted to take as input a sample X_n and output a set $\hat{S}_k$, where k is typically a free parameter of the algorithm, such as the number of means in k-means[1], the dimension of affine spaces in PCA, etc. Performance is measured by the empirical quantity $n^{-1}\sum_{i=1}^n d(x_i, \hat{S}_k)^2$, which is minimized among all sets in some class (e.g. sets of size k, affine spaces of dimension k,...) This formulation is general enough to encompass k-means and PCA, but also k-flats, non-negative matrix factorization, and sparse coding (see [21] and references therein.)

Using the discussion of Sec. 3, we can establish a clear connection between unsupervised learning and the problem of learning probability measures with respect to W_2. Consider as a running example the k-means problem, though the argument is general. Given an input X_n, the k-means problem is to find a set $|\hat{S}_k| = k$ minimizing its average distance from points in X_n. By associating to $\hat{S}_k$ the pushforward measure $\pi_{\hat{S}_k}\hat{\rho}_n$, we find that

$$\frac{1}{n}\sum_{i=1}^n d(x_i, \hat{S}_k)^2 = \mathbb{E}_{x\sim\hat{\rho}_n} d(x, \hat{S}_k)^2 \underset{\text{lemma 3.1}}{=} W_2(\hat{\rho}_n, \pi_{\hat{S}_k}\hat{\rho}_n)^2. \tag{2}$$

Since k-means minimizes equation 2, it also finds the measure that is closest to $\hat{\rho}_n$, among those with support of size k. This connection between k-means and W_2 measure approximation was, to the best of the authors' knowledge, first suggested by Pollard [28] though, as mentioned earlier, the argument carries over to many other unsupervised learning algorithms.

Unsupervised measure learning algorithms. We briefly clarify the steps involved in using an existing unsupervised learning algorithm for probability measure learning. Let $\mathcal{U}_k$ be a parametrized algorithm (e.g. k-means) that takes a sample X_n and outputs a set $\mathcal{U}_k(X_n)$. The measure learning algorithm $\mathcal{A}_k : \mathcal{M}^n \to P_p(\mathcal{M})$ corresponding to $\mathcal{U}_k$ is defined as follows:

1. $\mathcal{A}_k$ takes a sample X_n and outputs the measure $\pi_{\hat{S}_k}\hat{\rho}_n$, supported on $\hat{S}_k = \mathcal{U}_k(X_n)$;
2. since $\hat{\rho}_n$ is discrete, then so must $\pi_{\hat{S}_k}\hat{\rho}_n$ be, and thus $\mathcal{A}_k(X_n) = \frac{1}{n}\sum_{i=1}^n \delta_{\pi_{\hat{S}_k}(x_i)}$;
3. in practice, we can simply store an n-vector $\left[\pi_{\hat{S}_k}(x_1), \ldots, \pi_{\hat{S}_k}(x_n)\right]$, from which $\mathcal{A}_k(X_n)$ can be reconstructed by placing atoms of mass $1/n$ at each point.

In the case that $\mathcal{U}_k$ is the k-means algorithm, only k points and k masses need to be stored.

Note that any algorithm $\mathcal{A}'$ that attempts to output a measure $\mathcal{A}'(X_n)$ close to $\hat{\rho}_n$ can be cast in the above framework. Indeed, if S' is the support of $\mathcal{A}'(X_n)$ then, by lemma 3.2, $\pi_{S'}\hat{\rho}_n$ is the measure closest to $\hat{\rho}_n$ with support in S'. This effectively reduces the problem of learning a measure to that of

[1] In a slight abuse of notation, we refer to the k-means algorithm here as an ideal algorithm that solves the k-means problem, even though in practice an approximation algorithm may be used.

finding a set, and is akin to how the fact that every optimal quantizer is a nearest-neighbor quantizer (see [15], [12, p. 350], and [14, p. 37–38]) reduces the problem of finding an optimal quantizer to that of finding an optimal quantizing *set*.

Clearly, the minimum of equation 2 over sets of size k (the output of k-means) is monotonically non-increasing with k. In particular, since $\hat{S}_n = X_n$ and $\pi_{\hat{S}_n}\hat{\rho}_n = \hat{\rho}_n$, it is $\mathbb{E}_{x\sim\hat{\rho}_n} d(x, \hat{S}_n)^2 = W_2(\hat{\rho}_n, \pi_{\hat{S}_n}\hat{\rho}_n)^2 = 0$. That is, we can always make the learned measure arbitrarily close to $\hat{\rho}_n$ by increasing k. However, as pointed out in Sec. 2, the problem of measure learning is concerned with minimizing the 2-Wasserstein distance $W_2(\rho, \pi_{\hat{S}_k}\hat{\rho}_n)$ to the data-generating measure. The actual performance of k-means is thus not necessarily guaranteed to behave in the same way as the empirical one, and the question of characterizing its behavior as a function of k and n naturally arises.

Finally, we note that, while it is $\mathbb{E}_{x\sim\hat{\rho}_n} d(x, \hat{S}_k)^2 = W_2(\hat{\rho}_n, \pi_{\hat{S}_k}\hat{\rho}_n)^2$ (the empirical performances are the same in the optimal quantization, and measure learning problem formulations), the actual performances satisfy

$$\mathbb{E}_{x\sim\rho} d(x, \hat{S}_k)^2 \underset{\text{lemma 3.1}}{=} W_2(\rho, \pi_{\hat{S}_k}\rho)^2 \underset{\text{lemma 3.2}}{\leq} W_2(\rho, \pi_{\hat{S}_k}\hat{\rho}_n)^2, \quad 1 \leq k \leq n.$$

Consequently, with the identification between sets S and measures $\pi_S\hat{\rho}_n$, the measure learning problem is, in general, *harder* than the set-approximation problem (for example, if $\mathcal{M} = \mathbb{R}^d$ and ρ is absolutely continuous over a set of non-null volume, it is not hard to show that the inequality is almost surely strict: $\mathbb{E}_{x\sim\rho} d(x, \hat{S}_k)^2 < W_2(\rho, \pi_{\hat{S}_k}\hat{\rho}_n)^2$ for $1 < k < n$.)

In the remainder, we characterize the performance of k-means on the measure learning problem, for varying k, n. Although other unsupervised learning algorithms could have been chosen as basis for our analysis, k-means is one of the oldest and most widely used, and the one for which the deep connection between optimal quantization and measure approximation is most clearly manifested. Note that, by setting $k = n$, our analysis includes the problem of characterizing the behavior of the distance $W_2(\rho, \hat{\rho}_n)$ between empirical and population measures which, as indicated in Sec. 2.1, is a fundamental question in statistics (i.e. the speed of convergence of empirical to population measures.)

5 Learning rates

In order to analyze the performance of k-means as a measure learning algorithm, and the convergence of empirical to population measures, we propose the decomposition shown in fig. 2. The diagram includes all the measures considered in the paper, and shows the two decompositions used to prove upper bounds. The upper arrow (green), illustrates the decomposition used to bound the distance $W_2(\rho, \hat{\rho}_n)$. This decomposition uses the measures $\pi_{S_k}\rho$ and $\pi_{S_k}\hat{\rho}_n$ as intermediates to arrive at $\hat{\rho}_n$, where S_k is a k-point optimal quantizer of ρ, that is, a set S_k minimizing $\mathbb{E}_{x\sim\rho} d(x, S)^2$ over all sets of size $|S| = k$. The lower arrow (blue) corresponds to the decomposition of $W_2(\rho, \pi_{\hat{S}_k}\hat{\rho}_n)$ (the performance of k-means), whereas the labelled black arrows correspond to individual terms in the bounds. We begin with the (slightly) simpler of the two results.

5.1 Convergence rates for the empirical to population measures

Let S_k be the optimal k-point quantizer of ρ of order two [14, p. 31]. By the triangle inequality and the identity $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$, it follows that

$$W_2(\rho, \hat{\rho}_n)^2 \leq 3\left[W_2(\rho, \pi_{S_k}\rho)^2 + W_2(\pi_{S_k}\rho, \pi_{S_k}\hat{\rho}_n)^2 + W_2(\pi_{S_k}\hat{\rho}_n, \hat{\rho}_n)^2\right]. \tag{3}$$

This is the decomposition depicted in the upper arrow of fig. 2.

By lemma 3.1, the first term in the sum of equation 3 is the optimal k-point quantization error of ρ over a d-manifold $\mathcal{M}$ which, using recent techniques from [16] (see also [17, p. 491]), is shown in the proof of theorem 5.1 (part a) to be of order $\Theta(k^{-2/d})$. The remaining terms, b) and c), are slightly more technical and are bounded in the proof of theorem 5.1.

Since equation 3 holds for all $1 \leq k \leq n$, the best bound on $W_2(\rho, \hat{\rho}_n)$ can be obtained by optimizing the right-hand side over all possible values of k, resulting in the following probabilistic bound for the rate of convergence of the empirical to population measures.

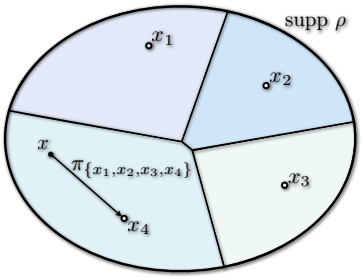

Figure 1: A sample $\{x_1, x_2, x_3, x_4\}$ is drawn from a distribution ρ with support in $\operatorname{supp}\rho$. The projection map $\pi_{\{x_1,x_2,x_3,x_4\}}$ sends points x to their closest one in the sample. The induced Voronoi tiling is shown in shades of blue.

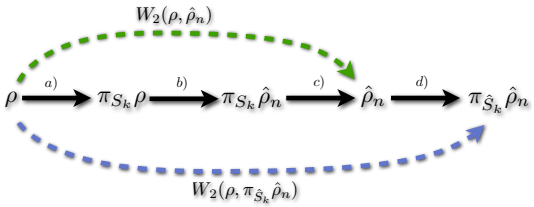

Figure 2: The measures considered in this paper are linked by arrows for which upper bounds for their distance are derived. Bounds for the quantities of interest $W_2(\rho, \hat{\rho}_n)^2$, and $W_2(\rho, \pi_{\hat{S}_k}\hat{\rho}_n)^2$, are decomposed by following the top and bottom colored arrows.

Theorem 5.1. *Given $\rho \in P_p(\mathcal{M})$ with absolutely continuous part $\rho_A \neq 0$, sufficiently large n, and $\tau > 0$, it holds*

$$W_2(\rho, \hat{\rho}_n) \leq C \cdot m(\rho_A) \cdot n^{-1/(2d+4)} \cdot \tau, \quad \text{with probability } 1 - e^{-\tau^2}.$$

where $m(\rho_A) := \int_{\mathcal{M}} \rho_A(x)^{d/(d+2)} d\lambda_{\mathcal{M}}(x)$, and C depends only on d.

5.2 Learning rates of k-means

The key element in the proof of theorem 5.1 is that the distance between population and empirical measures can be bounded by choosing an intermediate optimal quantizing measure of an appropriate size k. In the analysis, the best bounds are obtained for k smaller than n. If the output of k-means is close to an optimal quantizer (for instance if sufficient data is available), then we would similarly expect that the best bounds for k-means correspond to a choice of $k < n$.

The decomposition of the bottom (blue) arrow in figure 2 leads to the following bound in probability.

Theorem 5.2. *Given $\rho \in P_p(\mathcal{M})$ with absolutely continuous part $\rho_A \neq 0$, and $\tau > 0$, then for all sufficiently large n, and letting*

$$k = C \cdot m(\rho_A) \cdot n^{d/(2d+4)},$$

it holds

$$W_2(\rho, \pi_{\hat{S}_k}\hat{\rho}_n) \leq C \cdot m(\rho_A) \cdot n^{-1/(2d+4)} \cdot \tau, \quad \text{with probability } 1 - e^{-\tau^2}.$$

where $m(\rho_A) := \int_{\mathcal{M}} \rho_A(x)^{d/(d+2)} d\lambda_{\mathcal{M}}(x)$, and C depends only on d.

Note that the upper bounds in theorem 5.1 and 5.2 are exactly the same. Although this may appear surprising, it stems from the following fact. Since $S = \hat{S}_k$ is a minimizer of $W_2(\pi_S\hat{\rho}_n, \hat{\rho}_n)^2$, the bound d) of figure 2 satisfies:

$$W_2(\pi_{\hat{S}_k}\hat{\rho}_n, \hat{\rho}_n)^2 \leq W_2(\pi_{S_k}\hat{\rho}_n, \hat{\rho}_n)^2$$

and therefore (by the definition of c), the term d) is of the same order as c). It follows then that adding term d) to the bound only affects the constants, but otherwise leaves it unchanged. Since d) is the term that takes the output measure of k-means to the empirical measure, this implies that the rate of convergence of k-means (for suitably chosen k) cannot be worse than that of $\hat{\rho}_n \to \rho$. Conversely, bounds for $\hat{\rho}_n \to \rho$ are obtained from best rates of convergence of optimal quantizers, whose convergence to ρ cannot be slower than that of k-means (since the quantizers that k-means produces are suboptimal.)

Since the bounds obtained for the convergence of $\hat{\rho}_n \to \rho$ are the same as those for k-means with k of order $k = \Theta(n^{d/(2d+4)})$, this suggests that estimates of ρ that are as accurate as those derived from an n point-mass measure $\hat{\rho}_n$ can be derived from k point-mass measures with $k \ll n$.

Finally, we note that the introduced bounds are currently limited by the statistical bound

$$\sup_{|S|=k} |W_2(\pi_S\hat{\rho}_n, \hat{\rho}_n)^2 - W_2(\pi_S\rho, \rho)^2| \underset{\text{lemma 3.1}}{=} \sup_{|S|=k} |\mathbb{E}_{x\sim\hat{\rho}_n} d(x,S)^2 - \mathbb{E}_{x\sim\rho} d(x,S)^2| \quad (4)$$

(see for instance [21]), for which non-matching lower bounds are known. This means that, if better upper bounds can be obtained for equation 4, then both bounds in theorems 5.1 and 5.2 would automatically improve (would become closer to the lower bound.)

References

[1] M. Ajtai, J. Komls, and G. Tusndy. On optimal matchings. *Combinatorica*, 4:259–264, 1984.

[2] Franck Barthe and Charles Bordenave. Combinatorial optimization over two random point sets. Technical Report arXiv:1103.2734, Mar 2011.

[3] Gordon Blower. The Gaussian isoperimetric inequality and transportation. *Positivity*, 7:203–224, 2003.

[4] S. G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, April 1999.

[5] Emmanuel Boissard. Simple bounds for the convergence of empirical and occupation measures in 1-wasserstein distance. *Electron. J. Probab.*, 16(83):2296–2333, 2011.

[6] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3):541–593, 2007.

[7] F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculte des Sciences de Toulouse*, 14(3):331–352, 2005.

[8] Claire Caillerie, Frédéric Chazal, Jérôme Dedecker, and Bertrand Michel. Deconvolution for the Wasserstein metric and geometric inference. Rapport de recherche RR-7678, INRIA, July 2011.

[9] Kenneth L. Clarkson. Building triangulations using ϵ-nets. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, STOC '06, pages 326–335, New York, NY, USA, 2006. ACM.

[10] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[11] V. Dobri and J. Yukich. Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8:97–118, 1995.

[12] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression*. Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, 1992.

[13] Alison L. Gibbs and Francis E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–435, 2002.

[14] Siegfried Graf and Harald Luschgy. *Foundations of quantization for probability distributions*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2000.

[15] Siegfried Graf, Harald Luschgy, and Gilles Pageès. Distortion mismatch in the quantization of probability measures. *Esaim: Probability and Statistics*, 12:127–153, 2008.

[16] Peter M. Gruber. Optimum quantization and its applications. *Adv. Math*, 186:2004, 2002.

[17] P.M. Gruber. *Convex and discrete geometry*. Grundlehren der mathematischen Wissenschaften. Springer, 2007.

[18] Guillermo Henry and Daniela Rodriguez. Kernel density estimation on riemannian manifolds: Asymptotic results. *J. Math. Imaging Vis.*, 34(3):235–239, July 2009.

[19] Joseph Horowitz and Rajeeva L. Karandikar. Mean rates of convergence of empirical measures in the Wasserstein metric. *J. Comput. Appl. Math.*, 55(3):261–273, November 1994.

[20] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, 2001.

[21] A. Maurer and M. Pontil. K–dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839 –5846, nov. 2010.

[22] Yann Ollivier. Ricci curvature of markov chains on metric spaces. *J. Funct. Anal.*, 256(3):810–864, 2009.

[23] Arkadas Ozakin and Alexander Gray. Submanifold density estimation. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1375–1382. 2009.

[24] C. Papadimitriou. The probabilistic analysis of matching heuristics. In *Proc. of the 15th Allerton Conf. on Communication, Control and Computing*, pages 368–378, 1978.

[25] Bruno Pelletier. Kernel density estimation on Riemannian manifolds. *Statist. Probab. Lett.*, 73(3):297–304, 2005.

[26] Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *J. Math. Imaging Vis.*, 25(1):127–154, July 2006.

[27] M. S. Pinsker. Information and information stability of random variables and processes. *San Francisco: Holden-Day*, 1964.

[28] David Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, 28(2):199–204, 1982.

[29] S.T. Rachev. *Probability metrics and the stability of stochastic models*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1991.

[30] J.M. Steele. *Probability Theory and Combinatorial Optimization*. Cbms-Nsf Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1997.

[31] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geometric And Functional Analysis*, 6:587–600, 1996.

[32] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[33] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, 1996.

[34] V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics*, 19(1/2):23–26, Feb. 1958.

[35] C. Villani. *Optimal Transport: Old and New*. Grundlehren der Mathematischen Wissenschaften. Springer, 2009.

[36] P. Vincent and Y. Bengio. Manifold Parzen Windows. In *Advances in Neural Information Processing Systems 22*, pages 849–856. 2003.