Learning Probability Measures with respect to Optimal Transport Metrics: Supplementary Material

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

A Proofs

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

Proof. Consider a random variable X with $\text{Law}(X) = \rho$, and the random variable $Y = \pi_S(X)$, which satisfies $\text{Law}(Y) = \pi_S\rho$. It is

$$\begin{aligned}\mathbb{E}_{x\sim\rho} d(x,S)^p &= \mathbb{E}\|X - \pi_S X\|^p = \mathbb{E}\|X - Y\|^p \\ &\geq \inf\left\{\mathbb{E}\|X-Y\|^p, \text{Law}(X) = \rho, \text{Law}(Y) = \pi_S\rho\right\} = W_p(\rho, \pi_S\rho)^p\end{aligned}$$

Let $\epsilon > 0$ be arbitrary, and define X and Y to be random variables with $\text{Law}(X) = \rho$, and $\text{Law}(Y) = \pi_S\rho$ (but not necessarily related in a deterministic way), and such that they minimize eq. 1 (up to ϵ) for $\mu = \pi_S\rho$. Then, it holds

$$W_p(\rho, \pi_S\rho)^p + \epsilon \geq \mathbb{E}\|X - Y\|^p \underset{\text{supp}\,\mu\subseteq S}{\geq} \mathbb{E}\min_{q\in S}\|X - q\|^p = \mathbb{E}_{x\sim\rho} d(x,S)^p$$

Since $\epsilon > 0$ is arbitrary, it follows that $W_p(\rho, \pi_S\rho)^p \geq \mathbb{E}_{x\sim\rho} d(x,S)^p$. □

Lemma 3.2. For closed S, and all $\mu \in P_p(\mathcal{M})$ with $\text{supp}(\mu) \subseteq S$, it holds $W_p(\rho,\mu) \geq W_p(\rho, \pi_{S}\rho)$.

Proof. Let $\epsilon > 0$ be arbitrary, and define X, Y to be random variables with $\text{Law}(X) = \rho$ and $\text{Law}(Y) = \mu$, that minimize eq. 1 up to ϵ. It is

$$W_p(\rho,\mu)^p + \epsilon \geq \mathbb{E}\|X - Y\|^p \underset{\text{supp}\,\mu = S}{\geq} \mathbb{E}\min_{q\in S}\|X - q\|^p = \mathbb{E}_{x\sim\rho} d(x,S)^p \underset{\text{lemma 3.1}}{=} W_p(\rho, \pi_S\rho)^p$$

Since $\epsilon > 0$ is arbitrary, it follows that $W_p(\rho,\mu) \geq W_p(\rho, \pi_S\rho)$. □

In the reminder of the paper, we use C to denote a constant whose value may change each time it appears, but such that it only depends on the dimension d.

Theorem 5.1. Given $\rho \in P_p(\mathcal{M})$ with absolutely continuous part $\rho_A \neq 0$, sufficiently large n, and $0 < \delta < 1$, 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.

Proof. Given eq. 3, it is possible to bound from above each of the three terms in the sum. As stated in sec. 5, let S_k be an optimal quantizer of ρ of order 2 and size k. The terms' labels correspond to those in fig. 2.

(a) By lemma 3.1, the first term in fig. 2) is simply the quantization error associated to S_k, which, in $\mathbb{R}^d$, is know to be of order $\Theta(k^{-2/d})$ for measures with non-null absolutely continuous part [1]. More recently, the work of Gruber [2] (see [3] ch. 33 for an excellent account) extended these results to manifolds, effectively providing the same rates for quantization with respect to the geodesic distance on $\mathcal{M}$, and with quantization points lying on $\mathcal{M}$. Since the geodesic distance d_G of a d-manifold $\mathcal{M}$ embedded in $\mathcal{X}$ is never smaller than the natural distance in $\mathcal{X}$, and optimal quantization with points on the manifold $\mathcal{M} \subset \mathcal{X}$ can never perform better than choosing them from $\mathcal{X}$, it follows that, for sufficiently large k, it is

$$W_2(\rho, \pi_{S_k}\rho)^2 \underset{\text{lemma 3.1}}{=} \mathbb{E}_{x\sim\rho} d(x, S_k)^2 = \inf_{|S|=k, S\subset\mathcal{X}} \mathbb{E}_{x\sim\rho} d(x, S)^2 \leq \inf_{|S|=k, S\subset\mathcal{M}} \mathbb{E}_{x\sim\rho} d_G(x, S)^2$$

Therefore, by [2], if $\rho_A \neq 0$ (and thus $m(\rho_A) \neq 0$), it is

$$\lim_{k\to\infty} k^{2/d} W_2(\rho, \pi_{S_k}\rho)^2 \leq \lim_{k\to\infty} k^{2/d} \inf_{|S|=k, S\subset\mathcal{M}} \mathbb{E}_{x\sim\rho} d_G(x, S)^2 = C \cdot m(\rho_A)^{(d+2)/d}$$

and thus for sufficiently large k it is $W_2(\rho, \pi_{S_k}\rho)^2 \leq C \cdot m(\rho_A)^{(d+2)/d} \cdot k^{-2/d}$ (where the value of C may be slightly larger than that of the previous equation.)

(b) The second term ($W_2(\pi_{S_k}\rho, \pi_{S_k}\hat{\rho}_n)^2$) of eq. 3, can be bounded as follows. As pointed out in sec. 3.1, both $\pi_{S_k}\hat{\rho}_n$ and $\pi_{S_k}\rho$ are discrete distributions supported on S_k. In particular, if $S_k = \{m_1, \ldots, m_k\}$, they can be written as

$$\begin{aligned} \pi_{S_k}\rho &= \sum_{j=1}^{n} w_j \delta_{m_j} \\ \pi_{S_k}\hat{\rho}_n &= \sum_{i=1}^{n} \frac{1}{n} \delta_{\pi_{S_k}(x_i)} = \sum_{i=1}^{n} \hat{w}_j \delta_{x_j} \end{aligned} \tag{5}$$

for some probability masses $w_j, \hat{w}_j \in [0, 1]$. Let $\mathbf{w} := (w_j)_{j=1,\ldots,k} \in \mathbb{R}^k$, and $\hat{\mathbf{w}} := (\hat{w}_j)_{j=1,\ldots,k} \in \mathbb{R}^k$, and note that S_k is an optimal quantizer of ρ, and thus can be chosen deterministically as a function of ρ. Since $\hat{w}_j$ is simply the proportion of samples x_i that map to the same point $m_j \in S_k$, which is governed by the probabilities $(w_j)_{j=1,\ldots,k}$, the values $\hat{w}_j$ are distributed according to a scaled multinomial distribution $n^{-1}M(n; \mathbf{w})$ and, in particular, their expectation is $\mathbb{E}\hat{\mathbf{w}} = \mathbf{w}$.

The optimal cost $W_2(\pi_{S_k}\rho, \pi_{S_k}\hat{\rho}_n)^2$ of transporting $\pi_{S_k}\rho$ to $\pi_{S_k}\hat{\rho}_n$ corresponds to the cost of redistributing the excess probability masses $\hat{\mathbf{w}} - \mathbf{w}$, among the points in $S_k \subset \mathcal{M}$. Since $\mathcal{M}$ is constrained to lie in the unit ball of $\mathcal{X}$, the maximum (squared) distance that masses are transported by is 4, while the amount to transport is given by the excess mass at each m_j. It then follows that

$$W_2(\pi_{S_k}\rho, \pi_{S_k}\hat{\rho}_n)^2 \leq 4\|n^{-1}M(n; \hat{\mathbf{w}}) - \mathbf{w}\|_1$$

We obtain a bound on the L_1 norm of a multinomial from proposition A.6.6 of [5], although bounds of similar order could've been obtained by using known concentration inequalities for Hilbert space random variables, and combining them with a standard $\mathbb{R}^k$-norm inequality $\|\cdot\|_2 \leq \sqrt{k}\|\cdot\|_1$. The resulting bound is

$$W_2(\pi_{S_k}\rho, \pi_{S_k}\hat{\rho}_n)^2 \geq 8\sqrt{n}\lambda \quad \text{with probability } 2^k e^{-2\lambda^2},$$

or equivalently,

$$W_2(\pi_{S_k}\rho, \pi_{S_k}\hat{\rho}_n)^2 \leq \frac{8}{\sqrt{n}} \sqrt{k\frac{\ln 2}{2} + \frac{1}{2}\ln\frac{1}{\delta}} \quad \text{with probability } 1 - \delta.$$

(c) The third term of eq. 3 satisfies

$$\begin{aligned} W_2(\pi_{S_k}\hat{\rho}_n, \hat{\rho}_n)^2 &\leq W_2(\rho, \pi_{S_k}\rho)^2 + |W_2(\hat{\rho}_n, \pi_{S_k}\hat{\rho}_n)^2 - W_2(\rho, \pi_{S_k}\rho)^2| \\ &\underset{\text{lemma 3.1}}{=} \mathbb{E}_{x\sim\rho} d(x, S_k)^2 + |\mathbb{E}_{x\sim\hat{\rho}_n} d(x, S_k)^2 - \mathbb{E}_{x\sim\rho} d(x, S_k)^2| \\ &\leq \mathbb{E}_{x\sim\rho} d(x, S_k)^2 + \sup_{|S|=k} |\mathbb{E}_{x\sim\hat{\rho}_n} d(x, S)^2 - \mathbb{E}_{x\sim\rho} d(x, S)^2| \end{aligned} \tag{6}$$

The first term in the last line of eq. 6 is the optimal quantization error of size k, which was already bounded in part a). The second term is a uniform bound on the quantization error for sets of size k, which can be bounded by making use of theorem 6 in [4]:

$$\sup_{|S|=k} |\mathbb{E}_{x\sim\hat{\rho}_n} d(x,S)^2 - \mathbb{E}_{x\sim\rho} d(x,S)^2| \leq k\sqrt{\frac{72\pi}{n}} + \sqrt{\frac{8\ln 1/\delta}{n}} \quad \text{with probability } 1-\delta$$

By combining the above three bounds, it follows that, with probability at least $1-\delta$, with $0<\delta<1$, it is

$$\begin{aligned} W_2(\rho,\hat{\rho}_n)^2 &\leq C\cdot m(\rho_A)^{(d+2)/d} k^{-2/d} + \frac{8}{\sqrt{n}}\left(k\sqrt{\frac{9\pi}{2}} + \sqrt{2\ln\frac{1}{\delta}} + \sqrt{k\frac{\ln 2}{2} + \frac{1}{2}\ln\frac{1}{\delta}}\right) \\ &\underset{\delta<1,k\geq 1}{\leq} 2\sqrt{\ln\frac{1}{\delta}}\left(C\cdot m(\rho_A)^{(d+2)/d}k^{-2/d} + \frac{72k}{\sqrt{n}}\right) \end{aligned} \tag{7}$$

The minimizer of eq. 7 over all values of k can easily be seen to correspond to the choice $k = C\cdot n^{d/(2(d+2))}$. In particular, by setting

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

the two terms in the final sum of eq. 7 are balanced, yielding

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

The theorem follows by letting $\tau := \sqrt{\ln 1/\delta}$. □

Theorem 5.2. Given $\rho\in P_p(\mathcal{M})$ with absolutely continuous part $\rho_A\neq 0$, sufficiently large n, and $0<\delta<1$, 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.

Proof. Consider the decomposition of $W_2(\rho,\pi_{\hat{S}_k}\hat{\rho}_n)^2$ depicted in figure 2 (blue arrow):

$$W_2(\rho,\pi_{\hat{S}_k}\hat{\rho}_n)^2 \leq 2\left[W_2(\rho,\pi_{\hat{S}_k}\rho)^2 + W_2(\pi_{\hat{S}_k}\rho,\pi_{\hat{S}_k}\hat{\rho}_n)^2\right]$$

Letting S_k be, as before, an optimal quantizer of ρ of order 2 and size k, we may now simply reduce each of the above terms to those already analyzed in the proof of theorem 5.1:

(e) A bound for $W_2(\pi_{\hat{S}_k}\rho,\pi_{\hat{S}_k}\hat{\rho}_n)^2$ can be obtained in exactly the same way as the bound b) in theorem 5.1, by simply noticing that the same conditions apply, with the difference that the distributions whose distance we are bounding are supported on $\hat{S}_k$, rather than on the optimal quantizer S_k. Since the bound b) was obtained without assumptions on the support set (other than the fact that it is contained in the convex hull of $\operatorname{supp}\rho$, and this remains the case for the empirical minimizer $\hat{S}_k$), it is readily applicable to our case, and therefore

$$W_2(\pi_{\hat{S}_k}\rho,\pi_{\hat{S}_k}\hat{\rho}_n)^2 \leq \frac{8}{\sqrt{n}}\sqrt{k\frac{\ln 2}{2}+\frac{1}{2}\ln\frac{1}{\delta}} \quad \text{with probability } 1-\delta.$$

(f) Since $\hat{S}_k$ is a minimizer of $\mathbb{E}_{x\sim\hat{\rho}_n} d(x,\hat{S}_k)^2 \underset{\text{lemma 3.1}}{=} W_2(\hat{\rho}_n,\pi_{\hat{S}_k}\hat{\rho}_n)^2$ over sets of size k, and by part c) in the proof of theorem 5.1, it holds

$$\begin{aligned} W_2(\rho,\pi_{\hat{S}_k}\rho)^2 &\leq |W_2(\pi_{\hat{S}_k}\rho,\rho)^2 - W_2(\pi_{\hat{S}_k}\hat{\rho}_n,\hat{\rho}_n)^2| + W_2(\pi_{\hat{S}_k}\hat{\rho}_n,\hat{\rho}_n)^2 \\ &\leq |W_2(\pi_{\hat{S}_k}\rho,\rho)^2 - W_2(\pi_{\hat{S}_k}\hat{\rho}_n,\hat{\rho}_n)^2| + \\ &\quad |W_2(\pi_{S_k}\rho,\rho)^2 - W_2(\pi_{S_k}\hat{\rho}_n,\hat{\rho}_n)^2| + W_2(\rho,\pi_{S_k}\rho) \\ &\underset{\text{lemma 3.1}}{\leq} 2\sup_{|S|=k} |\mathbb{E}_{x\sim\hat{\rho}_n} d(x,S)^2 - \mathbb{E}_{x\sim\rho} d(x,S)^2| + \mathbb{E}_{x\sim\rho} d(x,S_k)^2 \end{aligned}$$

where the $\sup$ has been bounded in part c) of theorem 5.1, and a bound for the optimal quantization cost $\mathbb{E}_{x\sim\rho} d(x,S_k)^2$ is discussed in part a) of the same theorem.

By putting together the above bounds, we obtain an equivalent bound to that of eq. 7:

$$W_2(\rho, \pi_{\hat{S}_k}\hat{\rho}_n)^2 \leq 4\sqrt{\ln\frac{1}{\delta}}\left(C \cdot m(\rho_A)^{(d+2)/d}k^{-2/d} + \frac{72k}{\sqrt{n}}\right) \tag{8}$$

where the constants differ by a factor of two. Since the bounds in theorem 5.1 are written up to a universal multiplicative constant C that depends only on the dimension, eq. 8 implies that the exact same analysis holds in the k-means case that concerns us here. Namely, a bound

$$W_2(\rho, \pi_{\hat{S}_k}\hat{\rho}_n)^2 \leq C \cdot \sqrt{\ln\frac{1}{\delta}} \cdot m(\rho_A) \cdot n^{-1/(d+2)} \quad \text{with probability } 1-\delta$$

holds, and the theorem follows again by letting $\tau := \sqrt{\ln 1/\delta}$. □

References

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

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

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

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

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