
Learning Manifolds with K-Means and K-Flats: Supplementary Material

Guillermo D. Canas^{*,†} Tomaso Poggio^{*,†} Lorenzo A. Rosasco^{*,†}

* Laboratory for Computational and Statistical Learning - MIT-IIT

† CBCL, McGovern Institute - Massachusetts Institute of Technology

guilledc@mit.edu tp@ai.mit.edu lrosasco@mit.edu

A Methodology and Derivation of Results

Although both k-means and k-flats optimize the same empirical risk, the performance measure we are interested in is that of Equation 1. We may bound it from above as follows:

$$\begin{aligned} \mathcal{E}_\rho(S_{n,k}) &\leq |\mathcal{E}_\rho(S_{n,k}) - \mathcal{E}_n(S_{n,k})| + \mathcal{E}_n(S_{n,k}) - \mathcal{E}_n(S_k^*) + |\mathcal{E}_n(S_k^*) - \mathcal{E}_{\rho,k}^*| + \mathcal{E}_{\rho,k}^* \\ &\leq 2 \cdot \underbrace{\sup_{S \in \mathcal{S}_k} |\mathcal{E}_\rho(S) - \mathcal{E}_n(S)|}_{\text{Statistical error}} + \underbrace{\mathcal{E}_{\rho,k}^*}_{\text{Approximation error}} \end{aligned} \quad (14)$$

where $\mathcal{E}_{\rho,k}^* := \inf_{S \in \mathcal{S}_k} \mathcal{E}_\rho(S)$ is the best attainable performance over \mathcal{S}_k , and S_k^* is a set for which the best performance is attained. Note that $\mathcal{E}_n(S_{n,k}) - \mathcal{E}_n(S_k^*) \leq 0$ by the definition of $S_{n,k}$. The same error decomposition can be considered for k-flats, by replacing $S_{n,k}$ by $F_{n,k}$ and \mathcal{S}_k by \mathcal{F}_k .

Equation 14 decomposes the total learning error into two terms: a uniform (over all sets in the class C_k) bound on the difference between the empirical, and true error measures, and an *approximation error* term. The uniform statistical error bound will depend on the samples, and thus may hold with a certain probability.

In this setting, the approximation error will typically tend to zero as the class C_k becomes larger (as k increases.) Note that this is true, for instance, if C_k is the class of discrete sets of size k , as in the k-means problem.

The performance of Equation 14 is, through its dependence on the samples, a random variable. We will thus set out to find probabilistic bounds on its performance, as a function of the number n of samples, and the size k of the approximation. By choosing the approximation size parameter k to minimize these bounds, we obtain performance bounds as a function of the sample size.

B K-Means

We use the above decomposition to derive sample complexity bounds for the performance of the k-means algorithm. To derive explicit bounds on the different error terms we have to combine in a novel way some previous results and some new observations.

Approximation error. The error $\mathcal{E}_{\rho,k}^* = \inf_{S_k \in \mathcal{S}_k} \mathcal{E}_\rho(S_k)$ is related to the problem of optimal quantization. The classical optimal quantization problem is quite well understood, going back to the fundamental work of [21, 20] on optimal quantization for data transmission, and more recently by the work of [10, 13, 12, 6]. In particular, it is known that, for distributions with finite moment of order $2 + \lambda$, for some $\lambda > 0$, it is [10]

$$\lim_{k \rightarrow \infty} \mathcal{E}_{\rho,k}^* \cdot k^{2/d} = C \left\{ \int d\nu(x) p_a(x)^{d/(d+2)} \right\}^{(d+2)/d} \quad (15)$$

where ν is the Lebesgue measure, p_a is the density of the absolutely continuous part of the distribution (according to its Lebesgue decomposition), and C is a constant that depends only on the dimension. Therefore, the approximation error decays *at least* as fast as $k^{-2/d}$.

We note that, by setting μ to be the uniform distribution over the unit cube $[0, 1]^d$, it clearly is

$$\lim_{k \rightarrow \infty} \mathcal{E}_{\mu,k}^* \cdot k^{2/d} = C$$

and thus, by making use of Zador's asymptotic formula [21], and combining it with a result of Böröczky (see [13], p. 491), we observe that $C \sim (d/(2\pi e))^{r/2}$ with $d \rightarrow \infty$, for the r -th order quantization problem. In particular, this shows that the constant C only depends on the dimension, and, in our case ($r = 2$), has only linear growth in d , a fact that will be used in the sequel.

The approximation error $\mathcal{E}_{\rho,k}^* = \inf_{S_k \in \mathcal{S}_k} \mathcal{E}_\rho(S_k)$ of k -means is related to the problem of optimal quantization on manifolds, for which some results are known [12]. By calling $\mathcal{E}_{\mathcal{M},p,k}^*$ the approximation error only among sets of means contained in \mathcal{M} , Theorem 1 in Appendix C, implies in this case (letting $r = 2$) that

$$\lim_{k \rightarrow \infty} \mathcal{E}_{\rho,k}^* \cdot k^{2/d} = C \left\{ \int_{\mathcal{M}} d\mu_1(x) p(x)^{d/(d+2)} \right\}^{(d+2)/d} \quad (16)$$

where p is absolutely continuous over \mathcal{M} and, by replacing \mathcal{M} with a d -dimensional domain in \mathbb{R}^d , it is clear that the constant C is the same as above.

Since restricting the means to be on \mathcal{M} cannot decrease the approximation error, it is $\mathcal{E}_{\rho,k}^* \leq \mathcal{E}_{\mathcal{M},p,k}^*$, and therefore the right-hand side of Equation 16 provides an (asymptotic) upper bound to $\mathcal{E}_{\rho,k}^* \cdot k^{2/d}$.

For the statistical error we use available bounds.

Statistical error. The statistical error of Equation 14, which uniformly bounds the difference between the empirical, and expected error, has been widely-studied in recent years in the literature [16, 17, 3]. In particular, it has been shown that, for a distribution p over the unit ball in \mathbb{R}^d , it is

$$\sup_{S \in \mathcal{S}_k} |\mathcal{E}_\rho(S) - \mathcal{E}_n(S)| \leq \frac{k\sqrt{18\pi}}{\sqrt{n}} + \sqrt{\frac{8 \ln 1/\delta}{n}} \quad (17)$$

with probability $1 - \delta$ [16]. Clearly, this implies convergence $\mathcal{E}_n(S) \rightarrow \mathcal{E}_\rho(S)$ almost surely, as $n \rightarrow \infty$; although this latter result was proven earlier in [18], under the less restrictive condition that p have finite second moment.

By bringing together the above results, we obtain the bound in Theorem 1 on the performance of k -means, whose proof is postponed to Appendix A.

Further, we can consider the error incurred by the actual optimization algorithm used to compute the k -means solution.

Computational error. In practice, the k -means problem is NP-hard [1, 8, 15], with the original Lloyd relaxation algorithm providing no guarantees of closeness to the global minimum of Equation 2. However, practical approximations, such as the k -means++ algorithm [2], exist. When using k -means++, means are inserted one by one at samples selected with probability proportional to their squared distance to the set of previously-inserted means. This randomized seeding has been shown by [2] to output a set that is, in expectation, within a $8(\ln k + 2)$ -factor of the optimal. Once again, by combining these results, we obtain Theorem 2, whose proof is also in Appendix A.

We use the results discussed in Section A to obtain the proof of Theorem 1 as follows.

Proof. Letting $\|p\|_{d/(d+2)} := \left\{ \int d\mu_1(x) p(x)^{d/(d+2)} \right\}^{(d+2)/d}$, then with probability $1 - \delta$, it is

$$\begin{aligned} \mathcal{E}_\rho(S_{n,k}) &\leq 2n^{-1/2} \left(k\sqrt{18\pi} + \sqrt{8\ln 1/\delta} \right) + Ck^{-2/d} \cdot \|p\|_{d/(d+2)} \\ &\leq 2n^{-1/2} k\sqrt{18\pi} \cdot \sqrt{8\ln 1/\delta} + Ck^{-2/d} \cdot \|p\|_{d/(d+2)} \\ &= 24\sqrt{\pi} k n^{-1/2} \sqrt{\ln 1/\delta} + Ck^{-2/d} \cdot \|p\|_{d/(d+2)} \\ &= 2\sqrt{\ln 1/\delta} n^{-1/(d+2)} C^{d/(d+2)} (24\sqrt{\pi})^{2/(d+2)} \cdot \left\{ \int d\mu_1(x) p(x)^{d/(d+2)} \right\} \end{aligned} \quad (18)$$

where the parameter

$$k_n = n^{\frac{d}{2(d+2)}} \cdot \left(\frac{C}{24\sqrt{\pi}} \right)^{d/(d+2)} \cdot \left\{ \int d\mu_1(x) p(x)^{d/(d+2)} \right\} \quad (19)$$

has been chosen to balance the summands in the third line of Equation 18. \square

The proof of Theorem 2 follows a similar argument.

Proof. In the case of Theorem 2, the additional multiplicative term $A_k = 8(\ln k + 2)$ corresponding to the computational error incurred by the k-means++ algorithm does not affect the choice of parameter k_n since both summands in the third line of Equation 18 are multiplied by A_k in this case. Therefore, we may simply use the same choice of k_n as in Equation 19 in this case to obtain

$$\begin{aligned} \mathbb{E}_Z \mathcal{E}_\rho(S_{n,k}) &\leq 2n^{-1/2} \left(k\sqrt{18\pi} + \sqrt{8\ln 1/\delta} \right) + Ck^{-2/d} \cdot \|p\|_{d/(d+2)} \cdot 8(\ln k + 2) \\ &\leq 16\sqrt{\ln 1/\delta} n^{-1/(d+2)} C^{d/(d+2)} (24\sqrt{\pi})^{2/(d+2)} \cdot \left\{ \int d\mu_1(x) p(x)^{d/(d+2)} \right\} \\ &\quad \cdot \left[2 + \frac{d}{d+2} \left(\frac{1}{2} \ln n + \ln \frac{C}{12\sqrt{\pi}} + \ln \|p\|_{d/(d+2)} \right) \right] \end{aligned} \quad (20)$$

with probability $1 - \delta$, where the expectation is with respect to the random choice Z in the algorithm. From this the bound of Theorem 2 follows. \square

C K-Flats

Here we state a series of lemma that we prove in the next section. For the k-flats problem, we begin by introducing a uniform bound on the difference between empirical (Equation 2) and expected risk (Equation 1.)

Lemma 1. *If \mathcal{F}_k is the class of sets of k d -dimensional affine spaces then, with probability $1 - \delta$ on the sampling of $X_n \sim p$, it is*

$$\sup_{X' \in \mathcal{F}_k} |\mathcal{E}_\rho(X') - \mathcal{E}_n(X')| \leq k \sqrt{\frac{2\pi d}{n}} + \sqrt{\frac{\ln 1/\delta}{2n}}$$

By combining the above result with approximation error bounds, we may produce performance bounds on the expected risk for the k-flats problem, with appropriate choice of parameter k_n . We distinguish between the codimension one hypersurface case, and the more general case of a smooth manifold \mathcal{M} embedded in a Hilbert space. We begin with an approximation error bound for hypersurfaces in Euclidean space.

Lemma 2. *Assume given \mathcal{M} smooth with metric of class \mathcal{C}^3 in \mathbb{R}^{d+1} . If \mathcal{F}_k is the class of sets of k d -dimensional affine spaces, and $\mathcal{E}_{\rho,k}^*$ is the minimizer of Equation 1 over \mathcal{F}_k , then there is a constant C that depends on d only, such that*

$$\lim_{k \rightarrow \infty} \mathcal{E}_{\rho,k}^* \cdot k^{4/d} \leq C \cdot (\kappa_{\mathcal{M}})^{4/d}$$

where $\kappa_{\mathcal{M}} := \mu_{|H|}(\mathcal{M})$ is the total root curvature of \mathcal{M} , and $\mu_{|H|}$ is the measure associated with the (positive) second fundamental form. The constant C grows as $C \sim (d/(2\pi e))^2$ with $d \rightarrow \infty$.

For the more general problem of approximation of a smooth manifold in a separable Hilbert space, we begin by considering the definitions in Section 4 the second fundamental form Π and its operator norm $|\Pi_q|$ at a point $q \in \mathcal{M}$. Then we have:

Lemma 3. *Assume given a d -manifold \mathcal{M} with metric in \mathcal{C}^3 embedded in a separable Hilbert space \mathcal{X} . If \mathcal{F}_k is the class of sets of k d -dimensional affine spaces, and $\mathcal{E}_{\rho,k}^*$ is the minimizer of Equation 1 over \mathcal{F}_k , then there is a constant C that depends on d only, such that*

$$\lim_{k \rightarrow \infty} \mathcal{E}_{\rho,k}^* \cdot k^{4/d} \leq C \cdot (\kappa_{\mathcal{M}})^{4/d}$$

where $\kappa_{\mathcal{M}} := \int_{\mathcal{M}} d\mu_I(x) \frac{1}{4} |\Pi_x|^2$ and μ_I is the volume measure over \mathcal{M} . The constant C grows as $C \sim (d/(2\pi e))^2$ with $d \rightarrow \infty$.

We combine these two results into Theorems 3 and 4, whose derivation is in Appendix B.

C.1 Proofs

We begin proving the bound on the statistical error given in Lemma 1.

Proof. We begin by finding uniform upper bounds on the difference between Equations 1 and 2 for the class \mathcal{F}_k of sets of k d -dimensional affine spaces. To do this, we will first bound the Rademacher complexity $\mathcal{R}_n(\mathcal{F}_k, p)$ of the class \mathcal{F}_k .

Let Φ and Ψ be Gaussian processes indexed by \mathcal{F}_k , and defined by

$$\begin{aligned} \Phi_{X'} &= \sum_{i=1}^n \gamma_i \min_{j=1}^k d_{\mathcal{X}}^2(x_i, \pi'_j x_i) \\ \Psi_{X'} &= \sum_{i=1}^n \gamma_i \sum_{j=1}^k d_{\mathcal{X}}^2(x_i, \pi'_j x_i) \end{aligned} \tag{21}$$

$X' \in \mathcal{F}_k$, X' is the union of k d -subspaces: $X' = \cup_{j=1}^k F_j$, where each π'_j is an orthogonal projection onto F_j , and γ_i are independent Gaussian sequences of zero mean and unit variance.

Noticing that $d_{\mathcal{X}}^2(x, \pi x) = \|x\|^2 - \|\pi x\|^2 = \|x\|^2 - \langle x x^t, \pi \rangle_F$ for any orthogonal projection π (see for instance [5], Sec. 2.1), where $\langle \cdot, \cdot \rangle_F$ is the Hilbert-Schmidt inner product, we may verify that:

$$\begin{aligned} \mathbb{E}_{\gamma} (\Phi_{X'} - \Phi_{X''})^2 &= \sum_{i=1}^n \left[\min_{j=1}^k \|x_i\|^2 - \langle x_i x_i^t, \pi'_j \rangle_F - \left(\min_{j=1}^k \|x_i\|^2 - \langle x_i x_i^t, \pi''_j \rangle_F \right) \right]^2 \\ &\leq \sum_{i=1}^n \max_{j=1}^k \left(\langle x_i x_i^t, \pi'_j \rangle_F - \langle x_i x_i^t, \pi''_j \rangle_F \right)^2 \\ &\leq \sum_{i=1}^n \sum_{j=1}^k \left(\langle x_i x_i^t, \pi'_j \rangle_F - \langle x_i x_i^t, \pi''_j \rangle_F \right)^2 = \mathbb{E}_{\gamma} (\Psi_{X'} - \Psi_{X''})^2 \end{aligned} \tag{22}$$

Since it is,

$$\begin{aligned} \mathbb{E}_{\gamma} \sup_{X' \in \mathcal{F}_k} \sum_{i=1}^n \gamma_i \sum_{j=1}^k \langle x_i x_i^t, \pi'_j \rangle_F &= \mathbb{E}_{\gamma} \sup_{X' \in \mathcal{F}_k} \sum_{j=1}^k \left\langle \sum_{i=1}^n \gamma_i x_i x_i^t, \pi'_j \right\rangle_F \\ &\leq k \mathbb{E}_{\gamma} \sup_{\pi} \left\langle \sum_{i=1}^n \gamma_i x_i x_i^t, \pi \right\rangle_F \\ &\leq k \sup_{\pi} \|\pi\|_F \mathbb{E}_{\gamma} \left\| \sum_{i=1}^n \gamma_i x_i x_i^t \right\|_F \leq k \sqrt{dn} \end{aligned} \tag{23}$$

we may bound the Gaussian complexity $\Gamma_n(\mathcal{F}_k, p)$ as follows:

$$\begin{aligned}\Gamma_n(\mathcal{F}_k, p) &= \frac{2}{n} \mathbb{E}_\gamma \sup_{X' \in \mathcal{F}_k} \sum_{i=1}^n \gamma_i \min_{j=1}^k d_{\mathcal{X}}^2(x_i, \pi'_j x_i) \\ &\leq \frac{2}{n} \mathbb{E}_\gamma \sup_{X' \in \mathcal{F}_k} \sum_{i=1}^n \gamma_i \sum_{j=1}^k \langle x_i x_i^t, \pi'_j \rangle_F \leq 2k \sqrt{\frac{d}{n}}\end{aligned}\tag{24}$$

where the first inequality follows from Equation 22 and Slepian's Lemma [19], and the second from Equation 23.

Therefore the Rademacher complexity is bounded by

$$\mathcal{R}_n(\mathcal{F}_k, p) \leq \sqrt{\pi/2} \Gamma_n(\mathcal{F}_k, p) \leq k \sqrt{\frac{2\pi d}{n}}\tag{25}$$

Finally, by Theorem 8 of [4], it is:

$$\sup_{X' \in \mathcal{F}_k} |\mathcal{E}_\rho(X') - \mathcal{E}_n(X')| \leq \mathcal{R}_n(\mathcal{F}_k, p) + \sqrt{\frac{\ln 1/\delta}{2n}} \leq k \sqrt{\frac{2\pi d}{n}} + \sqrt{\frac{\ln 1/\delta}{2n}}\tag{26}$$

as desired. \square

C.2 Approximation Error

In order to prove approximation bounds for the k-flats problem, we will begin by first considering the simpler setting of a smooth d -manifold in \mathbb{R}^{d+1} space (codimension 1), and later we will extend the analysis to the general case.

Approximation Error: Codimension One

Assume that it is $\mathcal{X} = \mathbb{R}^{d+1}$ with the natural metric, and \mathcal{M} is a compact, smooth d -manifold with metric of class \mathcal{C}^2 . Since \mathcal{M} is of codimension one, the second fundamental form at each point is a map from the tangent space to the reals. Assume given $\alpha > 0$ and $\lambda > 0$. At every point $x \in \mathcal{M}$, define the metric $Q_x := |\Pi_x| + \alpha'(x)I_x$, where

- a) I and II are, respectively, the first and second fundamental forms on \mathcal{M} [9].
 - b) $|\Pi|$ is the *convexified* second fundamental form, whose eigenvalues are those of Π but in absolute value. If the second fundamental form Π is written in coordinates (with respect to an orthonormal basis of the tangent space) as $S\Lambda S^T$, with S orthonormal, and Λ diagonal, then $|\Pi|$ is $S|\Lambda|S^T$ in coordinates. Because $|\Pi|$ is continuous and positive semi-definite, it has an associated measure $\mu_{|\Pi|}$ (with respect to the volume measure μ_I).
 - c) $\alpha'(x) > 0$ is chosen such that $d\mu_{Q_x}/d\mu_I = d\mu_{|\Pi|}/d\mu_I + \alpha$. Note that such $\alpha'(x) > 0$ always exists since:
 - $\alpha'(x) = 0$ implies $d\mu_{Q_x}/d\mu_I = d\mu_{|\Pi|}/d\mu_I$, and
 - $d\mu_{Q_x}/d\mu_I$ can be made arbitrarily large by increasing $\alpha'(x)$.
- and therefore there is some intermediate value of $\alpha'(x) > 0$ that satisfies the constraint.

In particular, from condition c), it is clear that Q is everywhere positive definite.

Let μ_I and μ_Q be the measures over \mathcal{M} , associated with I and Q . Since, by its definition, μ_I is absolutely continuous with respect to I, then so must Q be. Therefore, we may define

$$\omega_Q := d\mu_Q/d\mu_I$$

to be the density of μ_Q with respect to μ_I .

Consider the discrete set $P_k \subset \mathcal{M}$ of size k that minimizes the quantity

$$f_{Q,p}(P_k) = \int_{\mathcal{M}} d\mu_Q(x) \left[\frac{p(x)}{\omega_Q(x)} \right] \min_{p \in P_k} d_Q^4(x, p)\tag{27}$$

among all sets of k points on \mathcal{M} . $f_{Q,p}(P_k)$ is the (fourth-order) quantization error over \mathcal{M} , with metric Q , and with respect to a weight function p/ω_Q . Note that, in the definition of $f_{Q,p}(P_k)$, it is crucial that the measure (μ_Q) , and distance (d_Q) match, in the sense that d_Q is the geodesic distance with respect to the metric Q , whose associated measure is μ_Q .

The following theorem, adapted from [12], characterizes the relation between k and the quantization error $f_{Q,p}(P_k)$ on a Riemannian manifold.

Theorem 1. [[12]] *Given a smooth compact Riemannian d -manifold \mathcal{M} with metric Q of class \mathcal{C}^1 , and a continuous function $w : \mathcal{M} \rightarrow \mathbb{R}^+$, then*

$$\min_{P \in \mathcal{P}_k} \int_{\mathcal{M}} d\mu_Q(x) w(x) \min_{p \in P} d_Q^r(x, p) \sim C \left\{ \int_{\mathcal{M}} d\mu_Q(x) w(x)^{d/(d+r)} \right\}^{(d+r)/d} \cdot k^{-r/d} \quad (28)$$

as $k \rightarrow \infty$, where the constant C depends only on d .

Furthermore, for each connected \mathcal{M} , there is a number $\xi > 1$ such that each set P_k that minimizes Equation 28 is a $(k^{-1/d}/\xi)$ -packing and $(\xi k^{-1/d})$ -cover of \mathcal{M} , with respect to d_Q .

This last result, which shows that a minimizing set P_k of size k must be a $(\xi k^{-1/d})$ -cover, clearly implies, by the definition of Voronoi diagram and the triangle inequality, the following key corollary.

Corollary 1. *Given \mathcal{M} , there is $\xi > 1$ such that each set P_k that minimizes Equation 28 has Voronoi regions of diameter no larger than $2\xi k^{-1/d}$, as measured by the distance d_Q .*

Let each $P_k \subset \mathcal{M}$ be a minimizer of Equation 27 of size k , then, for each k , define F_k to be the union of $(d$ -dimensional affine) tangent spaces to \mathcal{M} at each $q \in P_k$, that is, $F_k := \cup_{q \in P_k} T_q \mathcal{M}$. We may now use the definition of P_k to bound the approximation error $\mathcal{E}_\rho(F_k)$ on this set.

We begin by establishing some results that link distance to tangent spaces on manifolds to the geodesic distance d_Q associated with Q . The following lemma appears (in a slightly different form) as Lemma 4.1 in [7], and is borrowed from [12, 11].

Lemma 4. [[12, 11], [7]] *Given \mathcal{M} as above, and $\lambda > 0$ then, for every $p \in \mathcal{M}$ there is an open neighborhood $V_\lambda(p) \ni p$ in \mathcal{M} such that, for all $x, y \in V_\lambda(p)$, it is*

$$d_{\mathcal{X}}^2(x, T_y \mathcal{M}) \leq (1 + \lambda) d_{|\text{II}|}^4(x, y) \quad (29)$$

where $d_{\mathcal{X}}(x, T_y \mathcal{M})$ is the distance from x to the tangent plane $T_y \mathcal{M}$ at y , and $d_{|\text{II}|}$ is the geodesic distance associated with the convexified second fundamental form.

From the definition of Q , it is clear that, because Q strictly dominates $|\text{II}|$ then, for points x, y satisfying the conditions of Equation 29, it must be $d_{\mathcal{X}}(x, T_y \mathcal{M}) \leq (1 + \lambda) d_{|\text{II}|}(x, y) \leq (1 + \lambda) d_Q(x, y)$.

Given our choice of $\lambda > 0$, Lemma 4 implies that there is a collection of k neighborhoods, centered around the points $p \in P_k$, such that Equation 29 holds inside each. However, these neighborhoods may be too small for our purposes. In order to apply Lemma 4 to our problem, we will need to prove a stronger condition. We begin by considering the Dirichlet-Voronoi regions $D_{\mathcal{M},Q}(p; P_k)$ of points $p \in P_k$, with respect to the distance d_Q . That is,

$$D_{\mathcal{M},Q}(p; P_k) = \{x \in \mathcal{M} : d_Q(x, p) \leq d_Q(x, q), \forall q \in P_k\}$$

where, as before, P_k is a set of size k minimizing Equation 27.

Lemma 5. *For each $\lambda > 0$, there is k' such that, for all $k \geq k'$, and all $q \in P_k$, Equation 29 holds for all $x, y \in D_{\mathcal{M},Q}(q; P_k)$.*

Remark Note that, if it were $P'_k \subset P_k$ with $k > k'$ (if each P_{k+1} were constructed by adding one point to P_k), then Lemma 5 would follow automatically from Lemma 4 and Corollary 1. Since, in general, this not the case, the following proof is needed.

Proof. It suffices to show that every Voronoi region $D_{\mathcal{M},Q}(q; P_k)$, for sufficiently large k , is contained in a neighborhood $V_\lambda(v_q)$ of the type described in Lemma 4, for some $v_q \in \mathcal{M}$.

Clearly, by Lemma 4, the set $C = \{V_\lambda(x) : x \in \mathcal{M}\}$ is an open cover of \mathcal{M} . Since \mathcal{M} is compact, C admits a finite subcover C' . By the Lebesgue number lemma, there is $\delta > 0$ such that every set in \mathcal{M} of diameter less than δ is contained in some open set of C' .

Now let $k' = \lceil (\delta/2\xi)^{-d} \rceil$. By Corollary 1, every Voronoi region $D_{\mathcal{M},Q}(q; P_k)$, with $q \in P_k$, $k \geq k'$, has diameter less than δ , and is therefore contained in some set of C' . Since Equation 29 holds inside every set of C' then, in particular, it holds inside $D_{\mathcal{M},Q}(q; P_k)$. \square

We now have all the tools needed to prove:

Lemma 2 *If \mathcal{F}_k is the class of sets of k d -dimensional affine spaces, and $\mathcal{E}_{\rho,k}^*$ is the minimizer of Equation 1 over \mathcal{F}_k , then there is a constant C that depends on d only, such that*

$$\lim_{k \rightarrow \infty} \mathcal{E}_{\rho,k}^* \cdot k^{4/d} \leq C \cdot (\kappa_{\mathcal{M}})^{4/d}$$

where $\kappa_{\mathcal{M}} := \mu_{|\Pi|}(\mathcal{M})$ is the total root curvature of \mathcal{M} . The constant C grows as $C \sim (d/(2\pi e))^2$ with $d \rightarrow \infty$.

Proof. Pick $\alpha > 0$ and $\lambda > 0$. Given P_k minimizing Equation 27, if F_k is the union of tangent spaces at each $p \in P_k$, by Lemmas 4 and 5, it is

$$\begin{aligned} \mathcal{E}_\rho(F_k) &= \int_{\mathcal{M}} d\mu_1(x) p(x) \min_{p \in P_k} d_x^2(x, T_p \mathcal{M}) \\ &\leq (1 + \lambda) \int_{\mathcal{M}} d\mu_1(x) p(x) \min_{p \in P_k} d_Q^4(x, p) \\ &= (1 + \lambda) \int_{\mathcal{M}} d\mu_Q(x) \frac{p(x)}{\omega_Q(x)} \min_{p \in P_k} d_Q^4(x, p) \\ &\stackrel{\text{Thm. 1, } r=4}{\leq} (1 + \lambda) C \left\{ \int_{\mathcal{M}} d\mu_Q(x) \left[\frac{p(x)}{\omega_Q(x)} \right]^{d/(d+4)} \right\}^{(d+4)/d} \cdot k^{-4/d} \end{aligned} \tag{30}$$

where the last line follows from the fact that P_k has been chosen to minimize Equation 27, and where, in order to apply Theorem 1, we use the fact that p is absolutely continuous in \mathcal{M} .

By the definition of ω_Q , it follows that

$$\begin{aligned} \left\{ \int_{\mathcal{M}} d\mu_Q(x) \left[\frac{p(x)}{\omega_Q(x)} \right]^{d/(d+4)} \right\}^{(d+4)/d} &= \left\{ \int_{\mathcal{M}} d\mu_1(x) \omega_Q(x)^{4/(d+4)} p(x)^{d/(d+4)} \right\}^{(d+4)/d} \\ &\leq \left\{ \int_{\mathcal{M}} d\mu_1(x) \omega_Q(x) \right\}^{4/d} \end{aligned} \tag{31}$$

where the last line follows from Hölder's inequality ($\|fg\|_1 \leq \|f\|_p \|g\|_q$ with $p = (d+4)/d > 1$, and $q = (d+4)/4$.)

Finally, by the definition of Q and α' , it is

$$\int_{\mathcal{M}} d\mu_1(x) \omega_Q(x) \leq \int_{\mathcal{M}} d\mu_1(x) \alpha + \int_{\mathcal{M}} d\mu_{|\Pi|}(x) = \alpha \mathcal{V}_{\mathcal{M}} + \kappa_{\mathcal{M}} \tag{32}$$

where $\mathcal{V}_{\mathcal{M}}$ is the total volume of \mathcal{M} , and $\kappa_{\mathcal{M}} := \mu_{|\Pi|}(\mathcal{M})$ is the total root curvature of \mathcal{M} . Therefore

$$\mathcal{E}_\rho(F_k) \leq (1 + \lambda) C \{ \alpha \mathcal{V}_{\mathcal{M}} + \kappa_{\mathcal{M}} \}^{4/d} \cdot k^{-4/d} \tag{33}$$

Since $\alpha > 0$ and $\lambda > 0$ are arbitrary, Lemma 2 follows.

Finally, we discuss an important technicality in the proof that we hadn't mentioned before in the interest of clarity of exposition. Because we are taking absolute values in its definition, Q is not necessarily of class \mathcal{C}^1 , even if Π is. Therefore, we may not apply Theorem 1 directly. We may, however, use Whitney's approximation theorem (see for example [14] p. 252), to obtain a smooth

ϵ -approximation to Q , which can be enforced to be positive definite by relating the choice of ϵ to that of α , and with $\epsilon \rightarrow 0$ as $\alpha \rightarrow 0$. Since the ϵ -approximation Q only affects the final performance (Equation 33) by at most a constant times ϵ , then the fact that α is arbitrarily (and hence so is ϵ) implies the lemma. \square

Approximation Error: General Case

Assume given a d -manifold \mathcal{M} with metric in \mathcal{C}^3 embedded in a separable Hilbert space \mathcal{X} . Consider the definition in Section 4 of the second fundamental form Π and its operator norm $|\Pi|$.

We begin extending the results of Lemma 4 to the general case, where the manifold is embedded in a possibly infinite-dimensional ambient space. In this case, the orthogonal complement $(T_x\mathcal{M})^\perp$ to the tangent space at $x \in \mathcal{M}$ may be infinite-dimensional (although, by the separability of \mathcal{X} , it has a countable basis.)

For each $x \in \mathcal{M}$, consider the largest x -centered ball $B_x(\varepsilon)$ for which there is a smooth one-to-one Monge patch $m_x : B_x(\varepsilon_x) \subset T_x\mathcal{M} \rightarrow \mathcal{M}$. Since \mathcal{M} is smooth, and Π bounded, by the inverse function theorem it holds $\varepsilon_x > 0$. Because $\Pi \in \mathcal{C}^1$, we can always choose ε_x to be continuous in \mathcal{M} , and thus by the compactness of \mathcal{M} there is a minimum $0 < \varepsilon$ such that $0 < \varepsilon \leq \varepsilon_x$ with $x \in \mathcal{M}$. Let $N_x(\delta)$ denote the geodesic neighborhood around $x \in \mathcal{M}$ of radius δ . We begin by proving the following technical Lemma.

Lemma 6. *For every $q \in \mathcal{M}$, there is δ_q such that, for all $x, y \in N_q(\delta_q)$, it is $x \in m_y(B_y(\varepsilon))$ (x is in the Monge patch of y).*

Proof. The Monge function $m_y : B_y(\varepsilon) \rightarrow \mathcal{M}$ is such that $r \in B_y(\varepsilon)$ implies $m_y(r) - (y + r) \in (T_y\mathcal{M})^\perp$ (with the appropriate identification of vectors in \mathcal{X} and in $(T_y\mathcal{M})^\perp$), and therefore for all $r \in B_y(\varepsilon)$ it holds

$$d_I(y, m_y(r)) \geq \|m_y(r) - y\|_{\mathcal{X}} = \|m_y(r) - (y + r) + (y + r) - y\|_{\mathcal{X}} = \|m_y(r) - (y + r)\|_{\mathcal{X}} + \|r\|_{\mathcal{X}} \geq \|r\|_{\mathcal{X}}$$

Therefore $N_y(\varepsilon) \subset m_y(B_y(\varepsilon))$.

For each $q \in \mathcal{M}$, the geodesic ball $N_q(\varepsilon/2)$ is such that, by the triangle inequality, for all $x, y \in N_q(\varepsilon/2)$ it is $d_I(x, y) \leq \varepsilon$. Therefore $x \in N_y(\varepsilon) \subset m_y(B_y(\varepsilon))$. \square

Lemma 7. *For all $\lambda > 0$ and $q \in \mathcal{M}$, there is a neighborhood $V \ni q$ such that, for all $x, y \in V$ it is*

$$d_{\mathcal{X}}^2(x, T_y\mathcal{M}) \leq (1 + \lambda) d_I^4(x, y) |\Pi_x|^2 \quad (34)$$

Proof. Let V be a geodesic neighborhood of radius smaller than ε , so that Lemma 6 holds. Define the extension $\Pi_x^*(r) = \Pi_x^*(r^t + r^\perp) := \Pi_x(r^t)$ of the second fundamental form to \mathcal{X} , where $r^t \in T_x\mathcal{M}$ and $r^\perp \in (T_x\mathcal{M})^\perp$ is the unique decomposition of $r \in \mathcal{X}$ into tangent and orthogonal components.

By Lemma 6, given $x, y \in V$, x is in the (one-to-one) Monge patch m_y of y . Let $x' \in T_y\mathcal{M}$ be the unique point such that $m_y(x') = x$, and let $r := (x' - y)/\|x' - y\|_{\mathcal{X}}$. Since the domain of m_y is convex, the curve $\gamma_{y,r} : [0, \|x' - y\|_{\mathcal{X}}] \rightarrow \mathcal{M}$ given by

$$\gamma_{y,r}(t) = y + tr + m_y(tr) = y + tr + \frac{1}{2}t^2\Pi_y(r) + o(t^2)$$

is well-defined, where the last equality follows from the smoothness of Π . Clearly, $\gamma_{y,r}(\|x' - y\|_{\mathcal{X}}) = x$.

For $0 \leq t \leq \|x' - y\|_{\mathcal{X}}$ the length of $\gamma_{y,r}([0, t])$ is

$$L(\gamma_{y,r}([0, t])) = \int_0^t d\tau \|\dot{\gamma}_{y,r}(\tau)\|_{\mathcal{X}} = \int_0^t d\tau (\|r\|_{\mathcal{X}} + O(t)) = t \cdot (1 + o(1)) \quad (35)$$

(where $o(1) \rightarrow 0$ as $t \rightarrow 0$.) This establishes the closeness of distances in $T_y\mathcal{M}$ to geodesic distance on \mathcal{M} . In particular, for any $\alpha > 0$, $y \in \mathcal{M}$, there is a sufficiently small geodesic neighborhood $N \ni y$ such that, for $x \in N$, it holds

$$\|x' - y\|_{\mathcal{X}} \leq \|x - y\|_{\mathcal{X}} \leq d_I(x, y) \leq (1 + \lambda) \|x' - y\|_{\mathcal{X}}$$

By the smoothness of Π , for $y \in \mathcal{M}$ and $x \in N_y(\delta_y)$, with $0 < \delta_y < \varepsilon$, it is

$$\begin{aligned} d_{\mathcal{X}}^2(x, T_y \mathcal{M}) &= d_{\mathcal{X}}^2(\gamma_{y,r}(\|x' - y\|_{\mathcal{X}}), T_y \mathcal{M}) = \left\| \frac{1}{2} \Pi_y(r) \|x' - y\|_{\mathcal{X}}^2 + o(\|x' - y\|_{\mathcal{X}}^2) \right\|^2 \\ &= \left\| \frac{1}{2} \Pi_y^*(x - y) + o(\delta_y^2) \right\|^2 \end{aligned}$$

and therefore for any $\alpha > 0$, there is a sufficiently small $0 < \delta_{y,\alpha} < \varepsilon$ such that, given any $x \in N_y(\delta_{y,\alpha})$, it is

$$d_{\mathcal{X}}^2(x, T_y \mathcal{M}) \leq (1 + \alpha) \left\| \frac{1}{2} \Pi_y^*(x - y) \right\|^2 \quad (36)$$

By the smoothness of Π , and the same argument as in Lemma 6, there is a continuous choice of $0 < \delta_{y,\alpha}$, and therefore a minimum value $0 < \delta_{\alpha} \leq \delta_{y,\alpha}$, for $y \in \mathcal{M}$.

Similarly, by the smoothness of Π^* , for any $\alpha > 0$ and $y \in \mathcal{M}$, there is a sufficiently small $\beta_{y,\alpha} > 0$ such that, for all $x \in N_y(\beta_{y,\alpha})$, it holds

$$\left\| \frac{1}{2} \Pi_y^*(y - x) \right\|^2 \leq (1 + \alpha) \left\| \frac{1}{2} \Pi_x^*(y - x) \right\|^2 \quad (37)$$

By the argument of Lemma 6, there is a continuous choice of $0 < \beta_{y,\alpha}$, and therefore a minimum value $0 < \beta_{\alpha} \leq \beta_{y,\alpha}$, for $y \in \mathcal{M}$.

Finally, let $\alpha = \lambda/4$, and restrict $0 < \lambda < 1$ (larger λ are simply less restrictive.) For each $q \in \mathcal{M}$, let $V = N_q(\min\{\delta_{\alpha}, \beta_{\alpha}\}/2) \ni q$ be a sufficiently small geodesic neighborhood such that, for all $x, y \in V$, Eqs. 36 and 37 hold.

Since $\alpha = \lambda/4 < 1/4$, it is clearly $(1 + \alpha)^2 \leq (1 + \lambda)$, and therefore

$$\begin{aligned} d_{\mathcal{X}}^2(x, T_y \mathcal{M}) &\leq (1 + \alpha) \left\| \frac{1}{2} \Pi_y^*(y - x) \right\|^2 \leq (1 + \alpha)^2 \left\| \frac{1}{2} \Pi_x^*(y - x) \right\|^2 \\ &\leq (1 + \lambda) \frac{1}{4} \|y - x\|^4 |\Pi_x|^2 \leq (1 + \lambda) \frac{1}{4} d_1^4(x, y) |\Pi_x|^2 \end{aligned} \quad (38)$$

where the second-to-last inequality follows from the definition of $|\Pi|$. \square

Note that the same argument as that of Lemma 5 can be used here, with the goal of making sure that, for sufficiently large k , every Voronoi region of each $p \in P_k$ in the approximation satisfies Equation 34. We may now finish the proof by using a similar argument to that of the codimension-one case.

Let $\lambda > 0$. Consider a discrete set $P_k \subset \mathcal{M}$ of size k that minimizes

$$g(P_k) = \int_{\mathcal{M}} d\mu_1(x) \frac{1}{4} p(x) |\Pi_x|^2 \min_{p \in P_k} d_1^4(x, p) \quad (39)$$

Note once again that the distance and measure in Equation 39 match and therefore, since $p(x) |\Pi_x|^2 / 4$ is continuous, we can apply Theorem 1 (with $r = 4$) in this case.

Let $F_k := \cup_{q \in P_k} T_q \mathcal{M}$. By Lemma 7 and Lemma 5, adapted to this case, there is k' such that for all $k \geq k'$ it is

$$\begin{aligned} \mathcal{E}_{\rho}(F_k) &= \int_{\mathcal{M}} d\mu_1(x) \frac{1}{4} p(x) \min_{p \in P_k} d_x^2(x, T_p \mathcal{M}) \\ &\leq (1 + \lambda) \int_{\mathcal{M}} d\mu_1(x) \frac{1}{4} p(x) |\Pi_x|^2 \min_{p \in P_k} d_1^4(x, p) \\ &\stackrel{\text{Thm. 1, } r=4}{\leq} (1 + \lambda) C \left\{ \int_{\mathcal{M}} d\mu_1(x) \left[\frac{1}{4} p(x) |\Pi_x|^2 \right]^{d/(d+4)} \right\}^{(d+4)/d} \cdot k^{-4/d} \end{aligned} \quad (40)$$

where the last line follows from the fact that P_k has been chosen to minimize Equation 39.

Finally, by Hölder's inequality, it is

$$\begin{aligned} \left\{ \int_{\mathcal{M}} d\mu_1(x) \left[\frac{1}{4} p(x) |\Pi_x|^2 \right]^{d/(d+4)} \right\}^{(d+4)/d} &\leq \left\{ \int_{\mathcal{M}} d\mu_1(x) p(x) \right\} \left\{ \int_{\mathcal{M}} d\mu_1(x) \left(\frac{1}{4} |\Pi_x|^2 \right)^{d/4} \right\}^{4/d} \\ &= \left\| \frac{1}{4} |\Pi|^2 \right\|_{d/4} \end{aligned}$$

and thus

$$\mathcal{E}_\rho(F_k) \leq (1 + \lambda)C \cdot (\kappa_{\mathcal{M}}/k)^{4/d}$$

where the total curvature $\kappa_{\mathcal{M}} := \int_{\mathcal{M}} d\mu_1(x) \frac{1}{4} |\Pi_x|^{d/2}$ is the geometric invariant of the manifold (aside from the dimension) that controls the constant in the bound.

Since $\alpha > 0$ and $\lambda > 0$ are arbitrary, Lemma 3 follows.

Proofs of Theorems 3 and 4

We use the results discussed in Section A to obtain the proof of Theorem 3 as follows. The proof of Theorem 4 follows from the derivation in Section A, as well as the argument below, with $\kappa_{\mathcal{M}}^1$ substituted by $\kappa_{\mathcal{M}}$, and is omitted in the interest of brevity.

Proof. By Lemmas 1 and 2, with probability $1 - \delta$, it is

$$\begin{aligned} \mathcal{E}_\rho(F_{n,k}) &\leq 2n^{-1/2} \left(k\sqrt{2\pi d} + \sqrt{\frac{1}{2} \ln 1/\delta} \right) + C(\kappa_{\mathcal{M}}^1/k)^{4/d} \\ &\leq 2n^{-1/2} k\sqrt{2\pi d} \cdot \sqrt{\frac{1}{2} \ln 1/\delta} + C(\kappa_{\mathcal{M}}^1/k)^{4/d} \\ &= 2(8\pi d)^{2/(d+4)} C^{d/(d+4)} \cdot n^{-2/(d+4)} \cdot \sqrt{\frac{1}{2} \ln 1/\delta} \cdot (\kappa_{\mathcal{M}}^1)^{4/(d+4)} \end{aligned} \quad (41)$$

where the last line follows from choosing k to balance the two summands of the second line, as:

$$k_n = n^{\frac{d}{2(d+4)}} \cdot \left(\frac{C}{2\sqrt{2\pi d}} \right)^{d/(d+4)} \cdot (\kappa_{\mathcal{M}}^1)^{4/(d+4)}$$

□

References

- [1] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Mach. Learn.*, 75:245–248, May 2009.
- [2] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. SIAM.
- [3] Peter L. Bartlett, Tamas Linder, and Gabor Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44:1802–1813, 1998.
- [4] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
- [5] Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Mach. Learn.*, 66:259–294, March 2007.
- [6] E V Chernaya. On the optimization of weighted cubature formulae on certain classes of continuous functions. *East J. Approx.*, 1995.
- [7] 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.
- [8] Sanjoy Dasgupta and Yoav Freund. Random projection trees for vector quantization. *IEEE Trans. Inf. Theor.*, 55:3229–3242, July 2009.
- [9] M.P. DoCarmo. *Riemannian geometry*. Theory and Applications Series. Birkhäuser, 1992.
- [10] Siegfried Graf and Harald Luschgy. *Foundations of quantization for probability distributions*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2000.
- [11] P. M. Gruber. Asymptotic estimates for best and stepwise approximation of convex bodies i. *Forum Mathematicum*, 15:281–297, 1993.
- [12] Peter M. Gruber. Optimum quantization and its applications. *Adv. Math.*, 186:2004, 2002.

- [13] P.M. Gruber. *Convex and discrete geometry*. Grundlehren der mathematischen Wissenschaften. Springer, 2007.
- [14] J.M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, 2002.
- [15] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. In *Proceedings of the 3rd International Workshop on Algorithms and Computation, WALCOM '09*, pages 274–285, Berlin, Heidelberg, 2009. Springer-Verlag.
- [16] A. Maurer and M. Pontil. K-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, nov. 2010.
- [17] Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems 23*, pages 1786–1794. MIT Press, 2010.
- [18] David Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9(1):135–140, 1981.
- [19] David Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.*, 41:463–501, 1962.
- [20] G Fejes Toth. Sur la representation d’une population in par une nombre d’elements. *Acta Math. Acad. Sci. Hungaricae*, 1959.
- [21] Paul L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–148, 1982.