Learning Manifolds with K-Means and K-Flats

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

February 20, 2013

Abstract

We study the problem of estimating a manifold from random samples. In particular, we consider piecewise constant and piecewise linear estimators induced by k-means and k-flats, and analyze their performance. We extend previous results for k-means in two separate directions. First, we provide new results for k-means reconstruction on manifolds and, secondly, we prove reconstruction bounds for higher-order approximation (k-flats), for which no known results were previously available. While the results for k-means are novel, some of the technical tools are well-established in the literature. In the case of k-flats, both the results and the mathematical tools are new.

1 Introduction

Our study is broadly motivated by questions in high-dimensional learning. As is well known, learning in high dimensions is feasible only if the data distribution satisfies suitable prior assumptions. One such assumption is that the data distribution lies on, or is close to, a low-dimensional set embedded in a high dimensional space, for instance a low dimensional manifold. This latter assumption has proved to be useful in practice, as well as amenable to theoretical analysis, and it has led to a significant amount of recent work. Starting from [29, 40, 7], this set of ideas, broadly referred to as *manifold learning*, has been applied to a variety of problems from supervised [42] and semi-supervised learning [8], to clustering [45] and dimensionality reduction [7], to name a few.

Interestingly, the problem of learning the manifold itself has received less attention: given samples from a d-manifold $\mathcal{M}$ embedded in some ambient space $\mathcal{X}$, the problem is to learn a set that approximates $\mathcal{M}$ in a suitable sense. This problem has been considered in computational geometry, but in a setting in which typically the manifold is a hyper-surface in a low-dimensional space (e.g. $\mathbb{R}^3$), and the data are typically not sampled probabilistically, see for instance [32, 30]. The problem of learning a manifold is also related to that of estimating the support of a distribution, (see [17, 18] for recent surveys.) In this context, some of the distances considered to measure approximation quality are the Hausforff distance, and the so-called *excess mass* distance.

The reconstruction framework that we consider is related to the work of [1, 38], as well as to the framework proposed in [37], in which a manifold is approximated by a set, with performance measured by an expected distance to this set. This setting is similar to the problem of dictionary learning (see for instance [36], and extensive references therein), in which a dictionary is found by minimizing a similar reconstruction error, perhaps with additional constraints on an associated encoding of the data. Crucially, while the dictionary is learned on the empirical data, the quantity of interest is the expected reconstruction error, which is the focus of this work.

We analyze this problem by focusing on two important, and widely-used algorithms, namely k-means and k-flats. The k-means algorithm can be seen to define a piecewise constant approximation of $\mathcal{M}$. Indeed, it induces a Voronoi decomposition on $\mathcal{M}$, in which each Voronoi region is effectively approximated by a fixed mean. Given this, a natural extension is to consider higher order approximations, such as those induced by discrete collections of k d-dimensional affine spaces (k-flats), with possibly better

resulting performance. Since $\mathcal{M}$ is a d-manifold, the k-flats approximation naturally resembles the way in which a manifold is locally approximated by its tangent bundle.

Our analysis extends previous results for k-means to the case in which the data-generating distribution is supported on a manifold, and provides analogous results for k-flats. We note that the k-means algorithm has been widely studied, and thus much of our analysis in this case involves the combination of known facts to obtain novel results. The analysis of k-flats, however, requires developing substantially new mathematical tools.

The rest of the paper is organized as follows. In section 2, we describe the formal setting and the algorithms that we study. We begin our analysis by discussing the reconstruction properties of k-means in section 3. In section 4, we present and discuss our main results, whose proofs are postponed to the appendices.

2 Learning Manifolds

Let $\mathcal{X}$ by a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, endowed with a Borel probability measure ρ supported over a compact, smooth d-manifold $\mathcal{M}$. We assume the data to be given by a training set, in the form of samples $X_n = (x_1, \ldots, x_n)$ drawn identically and independently with respect to ρ.
Our goal is to *learn* a set S_n that approximates well the manifold. The approximation (learning error) is measured by the expected reconstruction error

$$\mathcal{E}_\rho(S_n) := \int_{\mathcal{M}} d\rho(x)\ d^2_{\mathcal{X}}(x, S_n), \tag{1}$$

where the distance to a set $S \subseteq \mathcal{X}$ is $d^2_{\mathcal{X}}(x, S) = \inf_{x' \in S} d^2_{\mathcal{X}}(x, x')$, with $d_{\mathcal{X}}(x, x') = \|x - x'\|$. This is the same reconstruction measure that has been the recent focus of [37, 5, 38].

It is easy to see that any set such that $S \supset \mathcal{M}$ will have zero risk, with $\mathcal{M}$ being the "smallest" such set (with respect to set containment.) In other words, the above error measure does not introduce an explicit penalty on the "size" of S_n: enlarging any *given* S_n can never increase the learning error.
With this observation in mind, we study specific learning algorithms that, given the data, produce a set belonging to some restricted hypothesis space $\mathcal{H}$ (e.g. sets of size k for k-means), which effectively introduces a constraint on the size of the sets. Finally, note that the risk of Equation 1 is non-negative and, if the hypothesis space is sufficiently *rich*, the risk of an unsupervised algorithm may converge to zero under suitable conditions.

2.1 Using K-Means and K-Flats for Piecewise Manifold Approximation

In this work, we focus on two specific algorithms, namely k-means [34, 33] and k-flats [12]. Although typically discussed in the Euclidean space case, their definition can be easily extended to a Hilbert space setting. The study of manifolds embedded in a Hilbert space is of special interest when considering non-linear (kernel) versions of the algorithms [20]. More generally, this setting can be seen as a limit case when dealing with high dimensional data. Naturally, the more classical setting of an absolutely continuous distribution over d-dimensional Euclidean space is simply a particular case, in which $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{M}$ is a domain with positive Lebesgue measure.
K-Means. Let $\mathcal{H} = \mathcal{S}_k$ be the class of sets of size k in $\mathcal{X}$. Given a training set X_n and a choice of k, k-means is defined by the minimization over $S \in \mathcal{S}_k$ of the empirical reconstruction error

$$\mathcal{E}_n(S) := \frac{1}{n} \sum_{i=1}^{n} d^2_{\mathcal{X}}(x_i, S). \tag{2}$$

where, for any fixed set S, $\mathcal{E}_n(S)$ is an unbiased empirical estimate of $\mathcal{E}_\rho(S)$, so that k-means can be seen to be performing a kind of empirical risk minimization [13, 9, 37, 10, 37].

A minimizer of Equation 2 on $\mathcal{S}_k$ is a discrete set of k *means* $S_{n,k} = \{m_1, \ldots, m_k\}$, which induces a Dirichlet-Voronoi tiling of $\mathcal{X}$: a collection of k regions, each closest to a common mean [4] (in our notation, the subscript n denotes the dependence of $S_{n,k}$ on the sample, while k refers to its size.) By virtue of $S_{n,k}$ being a minimizing set, each mean must occupy the center of mass of the samples in its Voronoi region.

These two facts imply that it is possible to compute a local minimum of the empirical risk by using a greedy coordinate-descent relaxation, namely Lloyd's algorithm [33]. Furthermore, given a finite sample X_n, the number of locally-minimizing sets $S_{n,k}$ is also finite since (by the center-of-mass condition) there cannot be more than the number of possible partitions of X_n into k groups, and therefore the global minimum must be attainable. Even though Lloyd's algorithm provides no guarantees of closeness to the global minimizer, in practice it is possible to use a randomized approximation algorithm, such as kmeans++ [3], which provides guarantees of approximation to the global minimum in expectation with respect to the randomization.

K-Flats. Let $\mathcal{H} = \mathcal{F}_k$ be the class of collections of k *flats* (affine spaces) of dimension d. For any value of k, k-flats, analogously to k-means, aims at finding the set $F_k \in \mathcal{F}_k$ that minimizes the empirical reconstruction (2) over $\mathcal{F}_k$. By an argument similar to the one used for k-means, a global minimizer must be attainable, and a Lloyd-type relaxation converges to a local minimum. Note that, in this case, given a Voronoi partition of $\mathcal{M}$ into regions closest to each d-flat, new optimizing flats for that partition can be computed by a d-truncated PCA solution on the samples falling in each region.

2.2 Learning a Manifold with K-means and K-flats

In practice, k-means is often interpreted to be a clustering algorithm, with clusters defined by the Voronoi diagram of the set of means $S_{n,k}$. In this interpretation, Equation 2 is simply rewritten by summing over the Voronoi regions, and adding all pairwise distances between samples in the region (the intra-cluster distances.) For instance, this point of view is considered in [14] where k-means is studied from an information theoretic persepective. K-means can also be interpreted to be performing vector quantization, where the goal is to minimize the encoding error associated to a nearest-neighbor quantizer [23]. Interestingly, in the limit of increasing sample size, this problem coincides, in a precise sense [39], with the problem of optimal quantization of probability distributions (see for instance the excellent monograph of [24].)

When the data-generating distribution is supported on a manifold $\mathcal{M}$, k-means can be seen to be approximating points on the manifold by a discrete set of means. Analogously to the Euclidean setting, this induces a Voronoi decomposition of $\mathcal{M}$, in which each Voronoi region is effectively approximated by a fixed mean (in this sense k-means produces a piecewise constant approximation of $\mathcal{M}$.) As in the Euclidean setting, the limit of this problem with increasing sample size is precisely the problem of optimal quantization of distributions on manifolds, which is the subject of significant recent work in the field of optimal quantization [26, 27].

In this paper, we take the above view of k-means as defining a (piecewise constant) approximation of the manifold $\mathcal{M}$ supporting the data distribution. In particular, we are interested in the behavior of the expected reconstruction error $\mathcal{E}_\rho(S_{n,k})$, for varying k and n. This perspective has an interesting relation with dictionary learning, in which one is interested in finding a dictionary, and an associated representation, that allows to approximately reconstruct a finite set of data-points/signals. In this interpretation, the set of means can be seen as a dictionary of size k that produces a maximally sparse representation (the k-means encoding), see for example [36] and references therein. Crucially, while the dictionary is learned on the available empirical data, the quantity of interest is the expected reconstruction error, and the question of characterizing the performance with respect to this latter quantity naturally arises.

Since k-means produces a piecewise constant approximation of the data, a natural idea is to consider higher orders of approximation, such as approximation by discrete collections of k d-dimensional affine spaces (k-flats), with possibly better performance. Since $\mathcal{M}$ is a d-manifold, the approximation induced by k-flats may more naturally resemble the way in which a manifold is locally approximated by its tangent bundle. We provide in Sec. 4.2 a partial answer to this question.

3 Reconstruction Properties of k-Means

Since we are interested in the behavior of the expected reconstruction (1) of k-means and k-flats for *varying k and n*, before analyzing this behavior, we consider what is currently known about this problem,

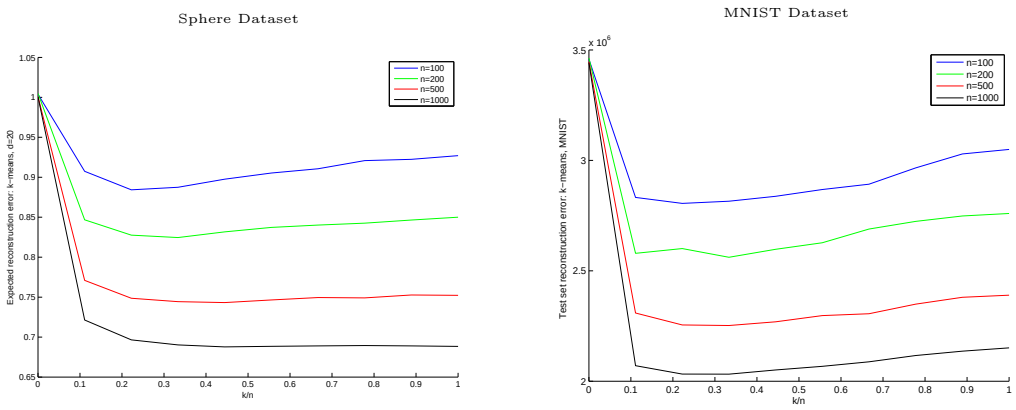


Figure 1: We consider the behavior of k-means for data sets obtained by sampling uniformly a 19 dimensional sphere embedded in $\mathbb{R}^{20}$ (left). For each value of k, k-means (with k-means++ seeding) is run 20 times, and the best solution kept. The reconstruction performance on a (large) hold-out set is reported as a function of k. The results for four different training set cardinalities are reported: for small number of points, the reconstruction error decreases sharply for small k and then increases, while it is simply decreasing for larger data sets. A similar experiment, yielding similar results, is performed on subsets of the MNIST (`http://yann.lecun.com/exdb/mnist`) database (right). In this case the data might be thought to be concentrated around a low dimensional manifold. For example [28] report an average intrinsic dimension d for each digit to be between 10 and 13.

based on previous work. While k-flats is a relatively new algorithm whose behavior is not yet well understood, several properties of k-means are currently known.

Recall that k-means find an discrete set $S_{n,k}$ of size k that best approximates the samples in the sense of (2). Clearly, as k increases, the empirical reconstruction error $\mathcal{E}_n(S_{n,k})$ cannot increase, and typically decreases. However, we are ultimately interested in the expected reconstruction error, and therefore would like to understand the behavior of $\mathcal{E}_\rho(S_{n,k})$ with varying k, n.

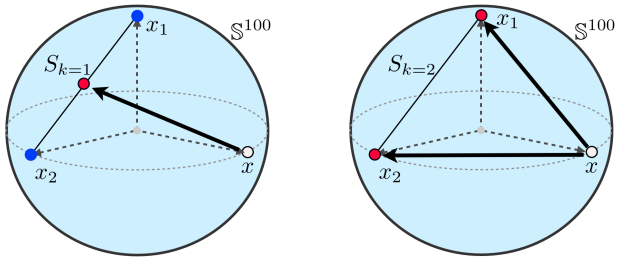
In the context of optimal quantization, the behavior of the expected reconstruction error $\mathcal{E}_\rho$ has been considered for an approximating set S_k obtained by minimizing the *expected* reconstruction error itself over the hypothesis space $\mathcal{H} = \mathcal{S}_k$. The set S_k can thus be interpreted as the output of a *population*, or infinite sample version of k-means. In this case, it is possible to show that $\mathcal{E}_\rho(S_k)$ is a non increasing function of k and, in fact, to derive explicit rates. For example in the case $\mathcal{X} = \mathbb{R}^d$, and under fairly general technical assumptions, it is possible to show that $\mathcal{E}_\rho(S_k) = \Theta(k^{-2/d})$, where the constants depend on ρ and d [24].

In machine learning, the properties of k-means have been studied, *for fixed* k, by considering the *excess* reconstruction error $\mathcal{E}_\rho(S_{n,k}) - \mathcal{E}_\rho(S_k)$. In particular, this quantity has been studied for $\mathcal{X} = \mathbb{R}^d$, and shown to be, with high probability, of order $\sqrt{kd/n}$, up-to logarithmic factors [37]. The case where $\mathcal{X}$ is a Hilbert space has been considered in [37, 10], where an upper-bound of order $k/\sqrt{n}$ is proven to hold with high probability. The more general setting where $\mathcal{X}$ is a metric space has been studied in [9].

When analyzing the behavior of $\mathcal{E}_\rho(S_{n,k})$, and in the particular case that $\mathcal{X} = \mathbb{R}^d$, the above results can be combined to obtain, with high probability, a bound of the form

$$\begin{aligned}\mathcal{E}_\rho(S_{n,k}) &\leq |\mathcal{E}_\rho(S_{n,k}) - \mathcal{E}_n(S_{n,k})| + \mathcal{E}_n(S_{n,k}) - \mathcal{E}_n(S_k) + |\mathcal{E}_n(S_k) - \mathcal{E}_\rho(S_k)| + \mathcal{E}_\rho(S_k) \\ &\leq C\left(\sqrt{\frac{kd}{n}} + k^{-2/d}\right) \end{aligned} \tag{3}$$

up to logarithmic factors, where the constant C does not depend on k or n (a complete derivation is given in the Appendix.) The above inequality suggests a somewhat surprising effect: the expected reconstruction properties of k-means may be described by a *trade-off* between a statistical error (of order $\sqrt{\frac{kd}{n}}$) and a geometric approximation error (of order $k^{-2/d}$.)



(a) $\mathcal{E}_\rho(S_{k=1}) \simeq 1.5$ (b) $\mathcal{E}_\rho(S_{k=2}) \simeq 2$

Figure 2: The optimal k-means (red) computed from $n = 2$ samples drawn uniformly on $\mathbb{S}^{100}$ (blue.) For a) $k = 1$, the expected squared-distance to a random point $x \in \mathbb{S}^{100}$ is $\mathcal{E}_\rho(S_{k=1}) \simeq 1.5$, while for b) $k = 2$, it is $\mathcal{E}_\rho(S_{k=2}) \simeq 2$.

The existence of such a tradeoff between the approximation, and the statistical errors may itself not be entirely obvious, see the discussion in [5]. For instance, in the k-means problem, it is intuitive that, as more means are inserted, the expected distance from a random sample to the means should decrease, and one might expect a similar behavior for the expected reconstruction error. This observation naturally begs the question of whether and when this trade-off really exists or if it is simply a result of the looseness in the bounds. In particular, one could ask how tight the bound (3) is.

While the bound on $\mathcal{E}_\rho(S_k)$ is known to be tight for k sufficiently large [24], the remaining terms (which are dominated by $|\mathcal{E}_\rho(S_{n,k}) - \mathcal{E}_n(S_{n,k})|$) are derived by controlling the supremum of an empirical process

$$\sup_{S \in S_k} |\mathcal{E}_n(S) - \mathcal{E}_\rho(S)| \tag{4}$$

and it is unknown whether available bounds for it are tight [37]. Indeed, it is not clear how close the *distortion redundancy* $\mathcal{E}_\rho(S_{n,k}) - \mathcal{E}_\rho(S_k)$ is to its known lower bound of order $d\sqrt{\frac{k^{1-\frac{4}{d}}}{n}}$ (in expectation) [5]. More importantly, we are not aware of a lower bound for $\mathcal{E}_\rho(S_{n,k})$ itself. Indeed, as pointed out in [5], "The exact dependence of the minimax distortion redundancy on k and d is still a challenging open problem".

Finally, we note that, whenever a trade-off can be shown to hold, it may be used to justify a heuristic for choosing k empirically as the value that minimizes the reconstruction error in a hold-out set.

In Figure 1 we perform some simple numerical simulations showing that the trade-off indeed occurs in certain regimes. The following example provides a situation where a trade-off can be easily shown to occur.

Example 1. *Consider a setup in which $n = 2$ samples are drawn from a uniform distribution on the unit $d = 100$-sphere, though the argument holds for other n much smaller than d. Because $d \gg n$, with high probability, the samples are nearly orthogonal: $< x_1, x_2 >_{\mathcal{X}} \simeq 0$, while a third sample x drawn uniformly on $\mathbb{S}^{100}$ will also very likely be nearly orthogonal to both x_1, x_2 [31]. The k-means solution on this dataset is clearly $S_{k=1} = \{(x_1 + x_2)/2\}$ (Fig 2(a)). Indeed, since $S_{k=2} = \{x_1, x_2\}$ (Fig 2(b)), it is $\mathcal{E}_\rho(S_{k=1}) \simeq 1.5 < 2 \simeq \mathcal{E}_\rho(S_{k=2})$ with very high probability. In this case, it is better to place a single mean closer to the origin (with $\mathcal{E}_\rho(\{0\}) = 1$), than to place two means at the sample locations. This example is sufficiently simple that the exact k-means solution is known, but the effect can be observed in more complex settings.*

4 Main Results

Contributions. Our work extends previous results in two different directions:

(a) We provide an analysis of k-means for the case in which the data-generating distribution is supported on a manifold embedded in a Hilbert space. In particular, in this setting: 1) we derive new results

on the approximation error, and 2) new sample complexity results (learning rates) arising from the choice of k by optimizing the resulting bound. We analyze the case in which a solution is obtained from an approximation algorithm, such as k-means++ [3], to include this computational error in the bounds.

(b) We generalize the above results from k-means to k-flats, deriving learning rates obtained from new bounds on both the statistical and the approximation errors. To the best of our knowledge, these results provide the first theoretical analysis of k-flats in either sense.

We note that the k-means algorithm has been widely studied in the past, and much of our analysis in this case involves the combination of known facts to obtain novel results. However, in the case of k-flats, there is currently no known analysis, and we provide novel results as well as new performance bounds for each of the components in the bounds.

Throughout this section we make the following technical assumption:

Assumption 1. *$\mathcal{M}$ is a smooth d-manifold with metric of class $\mathcal{C}^1$, contained in the unit ball in $\mathcal{X}$, and with volume measure denoted by μ_I. The probability measure ρ is absolutely continuous with respect to μ_I, with density p.*

4.1 Learning Rates for k-Means

The first result considers the idealized case where we have access to an exact solution for k-means.

Theorem 1. *Under Assumption 1, if $S_{n,k}$ is a solution of k-means then, for $0 < \delta < 1$, there are constants C and γ dependent only on d, and sufficiently large n' such that, by setting*

$$k_n = n^{\frac{d}{2(d+2)}} \cdot \left(\frac{C}{24\sqrt{\pi}}\right)^{d/(d+2)} \cdot \left\{\int_{\mathcal{M}} d\mu_I(x) p(x)^{d/(d+2)}\right\}, \tag{5}$$

and $S_n = S_{n,k_n}$, it is

$$\mathbb{P}\left[\mathcal{E}_\rho(S_n) \leq \gamma \cdot n^{-1/(d+2)} \cdot \sqrt{\ln 1/\delta} \cdot \left\{\int_{\mathcal{M}} d\mu_I(x) p(x)^{d/(d+2)}\right\}\right] \geq 1 - \delta, \tag{6}$$

for all $n \geq n'$, where $C \sim d/(2\pi e)$ and γ grows sublinearly with d.

Remark 1. *Note that the distinction between distributions with density in $\mathcal{M}$, and singular distributions is important. The bound of Equation (6) holds only when the absolutely continuous part of ρ over $\mathcal{M}$ is non-vanishing. the case in which the distribution is singular over $\mathcal{M}$ requires a different analysis, and may result in faster convergence rates.*

The following result considers the case where the k-means++ algorithm is used to compute the estimator.

Theorem 2. *Under Assumption 1, if $S_{n,k}$ is the solution of k-means++ , then for $0 < \delta < 1$, there are constants C and γ that depend only on d, and a sufficiently large n' such that, by setting*

$$k_n = n^{\frac{d}{2(d+2)}} \cdot \left(\frac{C}{24\sqrt{\pi}}\right)^{d/(d+2)} \cdot \left\{\int_{\mathcal{M}} d\mu_I(x) p(x)^{d/(d+2)}\right\}, \tag{7}$$

and $S_n = S_{n,k_n}$, it is

$$\mathbb{P}\left[\mathbb{E}_Z\, \mathcal{E}_\rho(S_n) \leq \gamma \cdot n^{-1/(d+2)} \left(\ln n + \ln \|p\|_{d/(d+2)}\right) \cdot \sqrt{\ln 1/\delta} \cdot \left\{\int_{\mathcal{M}} d\mu_I(x) p(x)^{d/(d+2)}\right\}\right] \geq 1 - \delta, \tag{8}$$

for all $n \geq n'$, where the expectation is with respect to the random choice Z in the algorithm, and $\|p\|_{d/(d+2)} = \left\{\int_{\mathcal{M}} d\mu_I(x) p(x)^{d/(d+2)}\right\}^{(d+2)/d}$, $C \sim d/(2\pi e)$, and γ grows sublinearly with d.

Remark 2. *In the particular case that $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{M}$ is contained in the unit ball, we may further bound the distribution-dependent part of Equations 6 and 8. Using Hölder's inequality, one obtains*

$$\begin{aligned}\int d\nu(x)p(x)^{d/(d+2)} &\leq \left[\int_{\mathcal{M}} d\nu(x)p(x)\right]^{d/(d+2)} \cdot \left[\int_{\mathcal{M}} d\nu(x)\right]^{2/(d+2)} \\ &\leq Vol(\mathcal{M})^{2/(d+2)} \leq \omega_d^{2/(d+2)},\end{aligned} \tag{9}$$

where ν is the Lebesgue measure in $\mathbb{R}^d$, and ω_d is the volume of the d-dimensional unit ball.

It is clear from the proof of Theorem 1 that, in this case, we may choose

$$k_n = n^{\frac{d}{2(d+2)}} \cdot \left(\frac{C}{24\sqrt{\pi}}\right)^{d/(d+2)} \cdot \omega_d^{2/d},$$

independently of the density p, to obtain a bound $\mathcal{E}_\rho(S_n^*) = O\left(n^{-1/(d+2)} \cdot \sqrt{\ln 1/\delta}\right)$ with probability $1-\delta$ (and similarly for Theorem 2, except for an additional $\ln n$ term), where the constant only depends on the dimension.

Remark 3. *Note that according to the above theorems, choosing k requires knowledge of properties of the distribution ρ underlying the data, such as the intrinsic dimension of the support. In fact, following the ideas in [43] Section 6.3-5, it is easy to prove that choosing k to minimize the reconstruction error on a hold-out set, allows to achieve the same learning rates (up to a logarithmic factor), adaptively in the sense that knowledge of properties of ρ are not needed.*

4.2 Learning Rates for k-Flats

To study k-flats, we need to slightly strengthen Assumption 1 by adding to it by the following:

Assumption 2. *Assume the manifold $\mathcal{M}$ to have metric of class $\mathcal{C}^3$, and finite second fundamental form II [22].*

One reason for the higher-smoothness assumption is that k-flats uses higher order approximation, whose analysis requires a higher order of differentiability.
We begin by providing a result for k-flats on hypersurfaces (codimension one), and next extend it to manifolds in more general spaces.

Theorem 3. *Let, $\mathcal{X} = \mathbb{R}^{d+1}$. Under Assumptions 1,2, if $F_{n,k}$ is a solution of k-flats, then there is a constant C that depends only on d, and sufficiently large n' such that, by setting*

$$k_n = n^{\frac{d}{2(d+4)}} \cdot \left(\frac{C}{2\sqrt{2\pi d}}\right)^{d/(d+4)} \cdot \left(\kappa_{\mathcal{M}}\right)^{4/(d+4)}, \tag{10}$$

and $F_n = \mathrm{F}_{n,k_n}$, then for all $n \geq n'$ it is

$$\mathbb{P}\left[\mathcal{E}_\rho(F_n) \leq 2\left(8\pi d\right)^{2/(d+4)} C^{d/(d+4)} \cdot n^{-2/(d+4)} \cdot \sqrt{\frac{1}{2}\ln 1/\delta} \cdot \left(\kappa_{\mathcal{M}}\right)^{4/(d+4)}\right] \geq 1-\delta, \tag{11}$$

where $\kappa_{\mathcal{M}} := \mu_{|\mathrm{II}|}(\mathcal{M}) = \int_{\mathcal{M}} d\mu_{\mathrm{I}}(x)|\kappa_G^{1/2}(x)|$ is the total root curvature of $\mathcal{M}$, $\mu_{|\mathrm{II}|}$ is the measure associated with the (positive) second fundamental form, and κ_G is the Gaussian curvature on $\mathcal{M}$.

In the more general case of a d-manifold $\mathcal{M}$ (with metric in $\mathcal{C}^3$) embedded in a separable Hilbert space $\mathcal{X}$, we cannot make any assumption on the codimension of $\mathcal{M}$ (the dimension of the orthogonal complement to the tangent space at each point.) In particular, the second fundamental form II, which is an extrinsic quantity describing how the tangent spaces bend locally is, at every $x \in \mathcal{M}$, a map $\mathrm{II}_x : T_x\mathcal{M} \mapsto (T_x\mathcal{M})^\perp$ (in this case of class $\mathcal{C}^1$ by Assumption 2) from the tangent space to its orthogonal complement ($\mathrm{II}(x) := B(x,x)$ in the notation of [22, p. 128].) Crucially, in this case, we may no longer assume the dimension of the orthogonal complement $(T_x\mathcal{M})^\perp$ to be finite.
Denote by $|\mathrm{II}_x| = \sup_{\substack{r \in T_x\mathcal{M} \\ \|r\| \leq 1}} \|\mathrm{II}_x(r)\|_{\mathcal{X}}$, the operator norm of II_x. We have:

Theorem 4. *Under Assumptions 1,2, if $F_{n,k}$ is a solution to the k-flats problem, then there is a constant C that depends only on d, and sufficiently large n' such that, by setting*

$$k_n = n^{\frac{d}{2(d+4)}} \cdot \left(\frac{C}{2\sqrt{2\pi d}}\right)^{d/(d+4)} \cdot \kappa_{\mathcal{M}}^{4/(d+4)}, \tag{12}$$

and $F_n = \mathrm{F}_{n,k_n}$, then for all $n \geq n'$ it is

$$\mathbb{P}\left[\mathcal{E}_\rho(F_n) \leq 2\left(8\pi d\right)^{2/(d+4)} C^{d/(d+4)} \cdot n^{-2/(d+4)} \cdot \sqrt{\frac{1}{2}\ln 1/\delta} \cdot \kappa_{\mathcal{M}}^{4/(d+4)}\right] \geq 1-\delta, \tag{13}$$

where $\kappa_{\mathcal{M}} := \int_{\mathcal{M}} d\mu_I(x)\ |\mathrm{II}_x|^2$

Note that the better k-flats bounds stem from the higher approximation power of d-flats over points. Although this greatly complicates the setup and proofs, as well as the analysis of the constants, the resulting bounds are of order $O\left(n^{-2/(d+4)}\right)$, compared with the slower order $O\left(n^{-1/(d+2)}\right)$ of k-means.

4.3 Discussion

In all the results, the final performance does not depend on the dimensionality of the embedding space (which in fact can be infinite), but only on the intrinsic dimension of the space on which the data-generating distribution is defined. The key to these results is an approximation construction in which the Voronoi regions on the manifold (points closest to a given mean or flat) are guaranteed to have vanishing diameter in the limit of k going to infinity. Under our construction, a hypersurface is approximated efficiently by tracking the variation of its tangent spaces by using the second fundamental form. Where this form vanishes, the Voronoi regions of an approximation will not be ensured to have vanishing diameter with k going to infinity, unless certain care is taken in the analysis.

An important point of interest is that the approximations are controlled by averaged quantities, such as the total root curvature (k-flats for surfaces of codimension one), total curvature (k-flats in arbitrary codimensions), and $d/(d+2)$-norm of the probability density (k-means), which are integrated over the domain where the distribution is defined. Note that these types of quantities have been linked to provably tight approximations in certain cases, such as for convex manifolds [25, 16], in contrast with worst-case methods that place a constraint on a maximum curvature, or minimum injectivity radius (for instance [1, 38].) Intuitively, it is easy to see that a constraint on an average quantity may be arbitrarily less restrictive than one on its maximum. A small difficult region (e.g. of very high curvature) may cause the bounds of the latter to substantially degrade, while the results presented here would not be adversely affected so long as the region is small.

Additionally, care has been taken throughout to analyze the behavior of the constants. In particular, there are no constants in the analysis that grow exponentially with the dimension, and in fact, many have polynomial, or slower growth. We believe this to be an important point, since this ensures that the asymptotic bounds do not hide an additional exponential dependence on the dimension.

References

[1] William K Allard, Guangliang Chen, and Mauro Maggioni. Multiscale geometric methods for data sets ii: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 1:1–38, 2011.

[2] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np–hardness of euclidean sum-of-squares clustering. *Mach. Learn.*, 75:245–248, May 2009.

[3] 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.

[4] Franz Aurenhammer. Voronoi diagrams: A survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23:345–405, September 1991.

[5] 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.

[6] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[7] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[8] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006.

[9] Shai Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Mach. Learn.*, 66(2-3):243–257, March 2007.

[10] Gérard Biau, Luc Devroye, and Gábor Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

[11] Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Mach. Learn.*, 66:259–294, March 2007.

[12] P. S. Bradley and O. L. Mangasarian. k-plane clustering. *J. of Global Optimization*, 16:23–32, January 2000.

[13] Joachim M. Buhmann. Empirical risk approximation: An induction principle for unsupervised learning. Technical report, University of Bonn, 1998.

[14] Joachim M. Buhmann. Information theoretic model validation for clustering. In *International Symposium on Information Theory, Austin Texas*. IEEE, 2010. (in press).

[15] E V Chernaya. On the optimization of weighted cubature formulae on certain classes of continuous functions. *East J. Approx*, 1995.

[16] 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.

[17] A. Cuevas and R. Fraiman. Set estimation. In *New perspectives in stochastic geometry*, pages 374–397. Oxford Univ. Press, Oxford, 2010.

[18] A. Cuevas and A. Rodríguez-Casal. Set estimation: an overview and some recent developments. In *Recent advances and trends in nonparametric statistics*, pages 251–264. Elsevier B. V., Amsterdam, 2003.

[19] Sanjoy Dasgupta and Yoav Freund. Random projection trees for vector quantization. *IEEE Trans. Inf. Theor.*, 55:3229–3242, July 2009.

[20] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 551–556, New York, NY, USA, 2004. ACM.

[21] J. Dieudonne. *Foundations of Modern Analysis*. Pure and Applied Mathematics. Hesperides Press, 2008.

[22] M.P. DoCarmo. *Riemannian geometry*. Theory and Applications Series. Birkhäuser, 1992.

[23] Allen Gersho and Robert M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991.

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

[25] P. M. Gruber. Asymptotic estimates for best and stepwise approximation of convex bodies i. *Forum Mathematicum*, 15:281–297, 1993.

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

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

[28] Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 289–296, 2005.

[29] V. De Silva J. B. Tenenbaum and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[30] Ravikrishna Kolluri, Jonathan Richard Shewchuk, and James F. O'Brien. Spectral surface reconstruction from noisy point clouds. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, SGP '04, pages 11–21, New York, NY, USA, 2004. ACM.

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

[32] David Levin. Mesh-independent surface interpolation. In Hamann Brunnett and Mueller, editors, *Geometric Modeling for Scientific Visualization*, pages 37–49. Springer-Verlag, 2003.

[33] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[34] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[35] 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.

[36] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 689–696, 2009.

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

[38] 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.

[39] David Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9(1):135–140, 1981.

[40] ST Roweis and LK Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[41] David Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.*, 41:463–501, 1962.

[42] Florian Steinke, Matthias Hein, and Bernhard Schölkopf. Nonparametric regression between general Riemannian manifolds. *SIAM J. Imaging Sci.*, 3(3):527–563, 2010.

[43] I. Steinwart and A. Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[44] G Fejes Toth. Sur la representation d´une population in par une nombre d´elements. *Acta Math. Acad. Sci. Hungaricae*, 1959.

[45] Ulrike von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007.

[46] Paul L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–148, 1982.

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:

$$\mathcal{E}_\rho(S_{n,k}) \leq |\mathcal{E}_\rho(S_{n,k}) - \mathcal{E}_n(S_{n,k})| + \mathcal{E}_n(S_{n,k}) - \mathcal{E}_n(S_k^*) + |\mathcal{E}_n(S_k^*) - \mathcal{E}_{\rho,k}^*| + \mathcal{E}_{\rho,k}^* \tag{14}$$

$$\leq 2 \cdot \underbrace{\sup_{S \in \mathcal{S}_k} |\mathcal{E}_\rho(S) - \mathcal{E}_n(S)|}_{\text{Statistical error}} + \underbrace{\mathcal{E}_{\rho,k}^*}_{\text{Approximation error}} \tag{15}$$

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 [46, 44] on optimal quantization for data transmission, and more recently by the work of [24, 27, 26, 15]. In particular, it is known that, for distributions with finite moment of order $2+\lambda$, for some $\lambda > 0$, it is [24]

$$\lim_{k\to\infty} \mathcal{E}^*_{\rho,k} \cdot k^{2/d} = C \left\{ \int d\nu(x) p_a(x)^{d/(d+2)} \right\}^{(d+2)/d} \tag{16}$$

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

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

and thus, by making use of Zador's asymptotic formula [46], and combining it with a result of Böröczky (see [27], p. 491), we observe that $C \sim (d/(2\pi e))^{r/2}$ with $d \to \infty$, 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 [26]. By calling $\mathcal{E}^*_{\mathcal{M},p,k}$ the approximation error only among sets of means contained in $\mathcal{M}$, Theorem 5 in Appendix C, implies in this case (letting $r = 2$) that

$$\lim_{k\to\infty} \mathcal{E}^*_{\rho,k} \cdot k^{2/d} = C \left\{ \int_{\mathcal{M}} d\mu_{\text{I}}(x)\ p(x)^{d/(d+2)} \right\}^{(d+2)/d} \tag{17}$$

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 17 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 [37, 38, 5]. In particular, it has been shown that, for a distribution p over the unit ball in $\mathbb{R}^d$, it is

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

with probability $1-\delta$ [37]. Clearly, this implies convergence $\mathcal{E}_n(S) \to \mathcal{E}_\rho(S)$ almost surely, as $n \to \infty$; although this latter result was proven earlier in [39], 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 [2, 19, 35], 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 [3], 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 [3] 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 $\|p\|_{d/(d+2)} := \left\{ \int d\mu_{\mathrm{I}}(x) p(x)^{d/(d+2)} \right\}^{(d+2)/d}$, then with probability $1-\delta$, it is

$$\begin{aligned}
\mathcal{E}_\rho(S_{n,k}) &\leq 2n^{-1/2} \left(k\sqrt{18\pi} + \sqrt{8 \ln 1/\delta} \right) + Ck^{-2/d} \cdot \|p\|_{d/(d+2)} \\
&\leq 2n^{-1/2} k \sqrt{18\pi} \cdot \sqrt{8 \ln 1/\delta} + Ck^{-2/d} \cdot \|p\|_{d/(d+2)} \\
&= 24\sqrt{\pi} k n^{-1/2} \sqrt{\ln 1/\delta} + Ck^{-2/d} \cdot \|p\|_{d/(d+2)} \\
&= 2\sqrt{\ln 1/\delta} n^{-1/(d+2)} C^{d/(d+2)} \left(24\sqrt{\pi}\right)^{2/(d+2)} \cdot \left\{ \int d\mu_{\mathrm{I}}(x) p(x)^{d/(d+2)} \right\}
\end{aligned} \tag{19}$$

where the parameter

$$k_n = n^{\frac{d}{2(d+2)}} \cdot \left(\frac{C}{24\sqrt{\pi}} \right)^{d/(d+2)} \cdot \left\{ \int d\mu_{\mathrm{I}}(x) p(x)^{d/(d+2)} \right\} \tag{20}$$

has been chosen to balance the summands in the third line of Equation 19. □

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 19 are multiplied by A_k in this case. Therefore, we may simply use the same choice of k_n as in Equation 20 in this case to obtain

$$\begin{aligned}
\mathbb{E}_Z\, \mathcal{E}_\rho(S_{n,k}) &\leq 2n^{-1/2} \left(k\sqrt{18\pi} + \sqrt{8 \ln 1/\delta} \right) + Ck^{-2/d} \cdot \|p\|_{d/(d+2)} \cdot 8(\ln k + 2) \\
&\leq 16\sqrt{\ln 1/\delta} n^{-1/(d+2)} C^{d/(d+2)} \left(24\sqrt{\pi}\right)^{2/(d+2)} \cdot \left\{ \int d\mu_{\mathrm{I}}(x) p(x)^{d/(d+2)} \right\} \\
&\quad \cdot \left[2 + \frac{d}{d+2} \left(\frac{1}{2} \ln n + \ln \frac{C}{12\sqrt{\pi}} + \ln \|p\|_{d/(d+2)} \right) \right]
\end{aligned} \tag{21}$$

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. □

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*

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

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

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 II and its operator norm $|\mathrm{II}_q|$ at a point $q \in \mathcal{M}$. The 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*

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

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

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

$$\begin{aligned} \Phi_{X'} &= \sum_{i=1}^{n} \gamma_i \min_{j=1}^{k} d^2_{\mathcal{X}}(x_i, \pi'_j x_i) \\ \Psi_{X'} &= \sum_{i=1}^{n} \gamma_i \sum_{j=1}^{k} d^2_{\mathcal{X}}(x_i, \pi'_j x_i) \end{aligned} \tag{22}$$

$X' \in \mathcal{F}_k$, 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 $d^2_{\mathcal{X}}(x, \pi x) = \|x\|^2 - \|\pi x\|^2 = \|x\|^2 - \langle xx^t, \pi \rangle_{_F}$ for any orthogonal projection π (see for instance [11], Sec. 2.1), where $\langle \cdot, \cdot \rangle_{_F}$ is the Hilbert-Schmidt inner product, we may verify that:

$$\begin{aligned} \mathbb{E}_\gamma \left(\Phi_{X'} - \Phi_{X''}\right)^2 &= \sum_{i=1}^{n} \left[\min_{j=1}^{k} \|x_i\|^2 - \left\langle x_i x_i^t, \pi'_j \right\rangle_{_F} - \left(\min_{j=1}^{k} \|x_i\|^2 - \left\langle x_i x_i^t, \pi''_j \right\rangle_{_F} \right) \right]^2 \\ &\leq \sum_{i=1}^{n} \max_{j=1}^{k} \left(\left\langle x_i x_i^t, \pi'_j \right\rangle_{_F} - \left\langle x_i x_i^t, \pi''_j \right\rangle_{_F} \right)^2 \\ &\leq \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\left\langle x_i x_i^t, \pi'_j \right\rangle_{_F} - \left\langle x_i x_i^t, \pi''_j \right\rangle_{_F} \right)^2 = \mathbb{E}_\gamma \left(\Psi_{X'} - \Psi_{X''}\right)^2 \end{aligned} \tag{23}$$

Since it is,

$$\begin{aligned}\mathbb{E}_\gamma \sup_{X'\in\mathcal{F}_k} \sum_{i=1}^{n}\gamma_i \sum_{j=1}^{k} \left\langle x_i x_i^t, \pi_j' \right\rangle_{_F} &= \mathbb{E}_\gamma \sup_{X'\in\mathcal{F}_k} \sum_{j=1}^{k} \left\langle \sum_{i=1}^{n}\gamma_i x_i x_i^t, \pi_j' \right\rangle_{_F} \\ &\leq k\,\mathbb{E}_\gamma \sup_{\pi} \left\langle \sum_{i=1}^{n}\gamma_i x_i x_i^t, \pi \right\rangle_{_F} \\ &\leq k \sup_{\pi} \|\pi\|_{_F}\, \mathbb{E}_\gamma \|\sum_{i=1}^{n}\gamma_i x_i x_i^t\|_{_F} \leq k\sqrt{dn}\end{aligned} \tag{24}$$

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

$$\begin{aligned}\Gamma_n(\mathcal{F}_k, p) &= \frac{2}{n}\,\mathbb{E}_\gamma \sup_{X'\in\mathcal{F}_k} \sum_{i=1}^{n}\gamma_i \min_{j=1}^{k} d_{\mathcal{X}}^2(x_i, \pi_j' x_i) \\ &\leq \frac{2}{n}\,\mathbb{E}_\gamma \sup_{X'\in\mathcal{F}_k} \sum_{i=1}^{n}\gamma_i \sum_{j=1}^{k} \left\langle x_i x_i^t, \pi_j' \right\rangle_{_F} \leq 2k\sqrt{\frac{d}{n}}\end{aligned} \tag{25}$$

where the first inequality follows from Equation 23 and Slepian's Lemma [41], and the second from Equation 24.

Therefore the Rademacher complexity is bounded by

$$\mathcal{R}_n(\mathcal{F}_k, p) \leq \sqrt{\pi/2}\Gamma_n(\mathcal{F}_k, p) \leq k\sqrt{\frac{2\pi d}{n}} \tag{26}$$

Finally, by Theorem 8 of [6], it is:

$$\sup_{X'\in\mathcal{F}_k} |\mathcal{E}_\rho(X') - \mathcal{E}_n(X')| \leq \mathcal{R}_n(\mathcal{F}_k, p) + \sqrt{\frac{\ln 1/\delta}{2n}} \leq k\sqrt{\frac{2\pi d}{n}} + \sqrt{\frac{\ln 1/\delta}{2n}} \tag{27}$$

as desired. □

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 $Q_x := |\mathrm{II}_x| + \alpha'(x)\mathrm{I}_x$, where

a) I and II are, respectively, the first and second fundamental forms on $\mathcal{M}$ [22].

b) $|\mathrm{II}|$ is the *convexified* second fundamental form, whose eigenvalues are those of II but in absolute value. If the second fundamental form II 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 $|\mathrm{II}|$ is $S|\Lambda|S^T$ in coordinates. Because $|\mathrm{II}|$ is continuous and positive semi-definite, it has an associated measure $\mu_{|\mathrm{II}|}$ (with respect to the volume measure μ_{I}.)

c) $\alpha'(x) > 0$ is chosen such that $d\mu_{_{Q_x}}/d\mu_{\mathrm{I}} = d\mu_{|\mathrm{II}|}/d\mu_{\mathrm{I}} + \alpha$. Note that such $\alpha'(x) > 0$ always exists since:

- $\alpha'(x) = 0$ implies $d\mu_{Q_x}/d\mu_{\mathrm{I}} = d\mu_{|\mathrm{II}|}/d\mu_{\mathrm{I}}$, and
- $d\mu_{Q_x}/d\mu_{\mathrm{I}}$ 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_{II} is absolutely continuous with respect to I, then so must Q be. Therefore, we may define

$$\omega_{Q} := d\mu_{Q}/d\mu_{\mathrm{I}}$$

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

$$f_{Q,p}(P_k) = \int_{\mathcal{M}} d\mu_{Q}(x) \left[\frac{p(x)}{\omega_{Q}(x)}\right] \min_{p \in P_k} d^4_{Q}(x,p) \tag{28}$$

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 [26], characterizes the relation between k and the quantization error $f_{Q,p}(P_k)$ on a Riemannian manifold.

Theorem 5. [[26]] *Given a smooth compact Riemannian d-manifold $\mathcal{M}$ with metric Q of class $\mathcal{C}^1$, and a continuous function $w : \mathcal{M} \to \mathbb{R}^+$, then*

$$\min_{P \in \mathcal{P}_k} \int_{\mathcal{M}} d\mu_{Q}(x) w(x) \min_{p \in P} d^r_{Q}(x,p) \sim C \left\{ \int_{\mathcal{M}} d\mu_{Q}(x) w(x)^{d/(d+r)} \right\}^{(d+r)/d} \cdot k^{-r/d} \tag{29}$$

as $k \to \infty$, 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 29 is a $\left(k^{-1/d}/\xi\right)$-packing and $\left(\xi k^{-1/d}\right)$-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 $\left(\xi k^{-1/d}\right)$-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 29 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 28 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 [16], and is borrowed from [26, 25].

Lemma 4. [[26, 25], [16]] *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*

$$d^2_{\mathcal{X}}(x, T_y\mathcal{M}) \leq (1+\lambda) d^4_{|II|}(x,y) \tag{30}$$

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 $d_{|II|}$ is the geodesic distance associated with the convexified second fundamental form.

From the definition of Q, it is clear that, because Q strictly dominates $|\mathrm{II}|$ then, for points x, y satisfying the conditions of Equation 30, it must be $d_{\mathcal{X}}(x, T_y\mathcal{M}) \leq (1+\lambda) d_{|\mathrm{II}|}(x,y) \leq (1+\lambda) d_{Q}(x,y)$.

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 30 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 28.

Lemma 5. *For each $\lambda > 0$, there is k' such that, for all $k \geq k'$, and all $q \in P_k$, Equation 30 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 30 holds inside every set of C' then, in particular, it holds inside $D_{\mathcal{M},Q}(q;P_k)$. □

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*

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

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

Proof. Pick $\alpha > 0$ and $\lambda > 0$. Given P_k minimizing Equation 28, if F_k is the union of tangent spaces at each $p \in P_k$, by Lemmas 4 and 5, it is

$$\begin{aligned}
\mathcal{E}_\rho(F_k) &= \int_{\mathcal{M}} d\mu_{\text{I}}(x) p(x) \min_{p\in P_k} d^2_{\mathcal{X}}(x, T_p\mathcal{M}) \\
&\leq (1+\lambda) \int_{\mathcal{M}} d\mu_{\text{I}}(x) p(x) \min_{p\in P_k} d^4_{Q}(x,p) \\
&= (1+\lambda) \int_{\mathcal{M}} d\mu_{Q}(x) \frac{p(x)}{\omega_Q(x)} \min_{p\in P_k} d^4_{Q}(x,p) \\
&\overset{\text{Thm. 5, r=4}}{\leq} (1+\lambda) C \left\{ \int_{\mathcal{M}} d\mu_{Q}(x) \left[\frac{p(x)}{\omega_Q(x)}\right]^{d/(d+4)} \right\}^{(d+4)/d} \cdot k^{-4/d}
\end{aligned} \tag{31}$$

where the last line follows from the fact that P_k has been chosen to minimize Equation 28, and where, in order to apply Theorem 5, we use the fact that p is absolutely continuous in $\mathcal{M}$.

By the definition of ω_Q, it follows that

$$\begin{aligned}
\left\{ \int_{\mathcal{M}} d\mu_{Q}(x) \left[\frac{p(x)}{\omega_Q(x)}\right]^{d/(d+4)} \right\}^{(d+4)/d} &= \left\{ \int_{\mathcal{M}} d\mu_{\text{I}}(x) \omega_Q(x)^{4/(d+4)} p(x)^{d/(d+4)} \right\}^{(d+4)/d} \\
&\leq \left\{ \int_{\mathcal{M}} d\mu_{\text{I}}(x) \omega_{Q}(x) \right\}^{4/d}
\end{aligned} \tag{32}$$

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

$$\int_{\mathcal{M}} d\mu_{\text{I}}(x)\omega_{Q}(x) \leq \int_{\mathcal{M}} d\mu_{\text{I}}(x)\alpha + \int_{\mathcal{M}} d\mu_{|\text{II}|}(x) = \alpha \mathcal{V}_{\mathcal{M}} + \kappa_{\mathcal{M}} \tag{33}$$

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

$$\mathcal{E}_{\rho}(F_k) \leq (1+\lambda)C \left\{\alpha \mathcal{V}_{\mathcal{M}} + \kappa_{\mathcal{M}}\right\}^{4/d} \cdot k^{-4/d} \tag{34}$$

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 absolutely values in its definition, Q is not necessarily of class $\mathcal{C}^1$, even if II is. Therefore, we may not apply Theorem 5 directly. We may, however, use Weierstrass' approximation theorem (see for example [21] p. 133), 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 $\epsilon \to 0$ as $\alpha \to 0$. Since the ϵ-approximation Q only affects the final performance (Equation 34) by at most a constant times ϵ, then the fact that α is arbitrarily small (and thus so is ϵ) implies the lemma. □

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 II and its operator norm |II|.

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 $m_x : B_x(\varepsilon_x) \subset T_x\mathcal{M} \to \mathcal{M}$. Since $\mathcal{M}$ is smooth, and II bounded, by the inverse function theorem it holds $\varepsilon_x > 0$. Because $\text{II} \in \mathcal{C}^1$, 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 $m_y : B_y(\epsilon) \to \mathcal{M}$ is such that $r \in B_y(\epsilon)$ 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 $r \in B_y(\epsilon)$ it holds

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

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 $d_{\text{I}}(x, y) \leq \varepsilon$. Therefore $x \in N_y(\varepsilon) \subset m_y(B_y(\varepsilon))$. □

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*

$$d^2_{\mathcal{X}}(x, T_y\mathcal{M}) \leq (1+\lambda) d^4_{\text{I}}(x, y) |\text{II}_x|^2 \tag{35}$$

Proof. Let V be a geodesic neighborhood of radius smaller than ε, so that Lemma 6 holds. Define the extension $\text{II}^*_x(r) = \text{II}^*_x(r^t + r^{\perp}) := \text{II}_x(r^t)$ 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 $\gamma_{y,r} : [0, \|x' - y\|_{\mathcal{X}}] \to \mathcal{M}$ given by

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

is well-defined, where the last equality follows from the smoothness of II. 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

$$L(\gamma_{y,r}([0,t])) = \int_0^t d\tau \|\dot{\gamma_{y,r}}(\tau)\|_{\mathcal{X}} = \int_0^t d\tau \left(\|r\|_{\mathcal{X}} + O(t)\right) = t \cdot (1 + o(1)) \tag{36}$$

(where $o(1) \to 0$ as $t \to 0$.) 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

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

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

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

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

$$d^2_{\mathcal{X}}(x, T_y\mathcal{M}) \leq (1 + \alpha)\|\frac{1}{2}\mathrm{II}^*_y(x - y)\|^2 \tag{37}$$

By the smoothness of II, 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 II^*, 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

$$\|\frac{1}{2}\mathrm{II}^*_y(y - x)\|^2 \leq (1 + \alpha)\|\frac{1}{2}\mathrm{II}^*_x(y - x)\|^2 \tag{38}$$

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. 37 and 38 hold.

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

$$\begin{aligned} d^2_{\mathcal{X}}(x, T_y\mathcal{M}) &\leq (1 + \alpha)\|\frac{1}{2}\mathrm{II}^*_y(y - x)\|^2 \leq (1 + \alpha)^2\|\frac{1}{2}\mathrm{II}^*_x(y - x)\|^2 \\ &\leq (1 + \lambda)\frac{1}{4}\|y - x\|^4|\mathrm{II}_x|^2 \leq (1 + \lambda)\frac{1}{4}d^4_{\mathrm{I}}(x, y)|\mathrm{II}_x|^2 \end{aligned} \tag{39}$$

where the second-to-last inequality follows from the definition of |II|. □

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 35. 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

$$g(P_k) = \int_{\mathcal{M}} d\mu_{\mathrm{I}}(x) \frac{1}{4} p(x) |\mathrm{II}_x|^2 \min_{p \in P_k} d^4_{\mathrm{I}}(x, p) \tag{40}$$

Note once again that the distance and measure in Equation 40 match and therefore, since $p(x)|\mathrm{II}_x|^2/4$ is continuous, we can apply Theorem 5 (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

$$\begin{aligned}
\mathcal{E}_\rho(F_k) &= \int_{\mathcal{M}} d\mu_{\text{I}}(x) \frac{1}{4} p(x) \min_{p \in P_k} d^2_{\mathcal{X}}(x, T_p\mathcal{M}) \\
&\leq (1+\lambda) \int_{\mathcal{M}} d\mu_{\text{I}}(x) \frac{1}{4} p(x) |\mathrm{II}_x|^2 \min_{p \in P_k} d^4_{\text{I}}(x, p) \\
&\overset{\text{Thm. 5}, r=4}{\leq} (1+\lambda) C \left\{ \int_{\mathcal{M}} d\mu_{\text{I}}(x) \left[\frac{1}{4} p(x) |\mathrm{II}_x|^2 \right]^{d/(d+4)} \right\}^{(d+4)/d} \cdot k^{-4/d}
\end{aligned} \tag{41}$$

where the last line follows from the fact that P_k has been chosen to minimize Equation 40.

Finally, by Hölder's inequality, it is

$$\begin{aligned}
\left\{ \int_{\mathcal{M}} d\mu_{\text{I}}(x) \left[\frac{1}{4} p(x) |\mathrm{II}_x|^2 \right]^{d/(d+4)} \right\}^{(d+4)/d} &\leq \left\{ \int_{\mathcal{M}} d\mu_{\text{I}}(x) p(x) \right\} \left\{ \int_{\mathcal{M}} d\mu_{\text{I}}(x) \left(\frac{1}{4} |\mathrm{II}_x\|^2 \right)^{d/4} \right\}^{4/d} \\
&= \| \frac{1}{4} |\mathrm{II}|^2 \|_{d/4}
\end{aligned}$$

and thus

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

where the total curvature $\kappa_{\mathcal{M}} := \int_{\mathcal{M}} d\mu_{\text{I}}(x) \frac{1}{4} |\mathrm{II}_x|^{d/2}$ 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 $\kappa^1_{\mathcal{M}}$ 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

$$\begin{aligned}
\mathcal{E}_\rho(F_{n,k}) &\leq 2n^{-1/2} \left(k\sqrt{2\pi d} + \sqrt{\frac{1}{2} \ln 1/\delta} \right) + C(\kappa^1_{\mathcal{M}}/k)^{4/d} \\
&\leq 2n^{-1/2} k \sqrt{2\pi d} \cdot \sqrt{\frac{1}{2} \ln 1/\delta} + C(\kappa^1_{\mathcal{M}}/k)^{4/d} \\
&= 2 \left(8\pi d\right)^{2/(d+4)} C^{d/(d+4)} \cdot n^{-2/(d+4)} \cdot \sqrt{\frac{1}{2} \ln 1/\delta} \cdot \left(\kappa^1_{\mathcal{M}}\right)^{4/(d+4)}
\end{aligned} \tag{42}$$

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

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

□