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

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 [23, 34, 5], this set of ideas, broadly referred to as *manifold learning*, has been applied to a variety of problems from supervised [35] and semi-supervised learning [6], to clustering [37] and dimensionality reduction [5], 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 [26, 24]. The problem of learning a manifold is also related to that of estimating the support of a distribution, (see [13, 14] 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, 32], as well as to the framework proposed in [30], 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 [29], 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 approxima-

tions, 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 [30, 4, 32].

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 [28, 27] and k-flats [9]. 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 [15]. 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 [10, 7, 30, 8, 31].

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 [3] (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 [27]. 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++ [2], 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 [11] 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 [17]. Interestingly, in the limit of increasing sample size, this problem coincides, in a precise sense [33], with the problem of optimal quantization of probability distributions (see for instance the excellent monograph of [18].)

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 [20, 21].

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 [29] 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, 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.

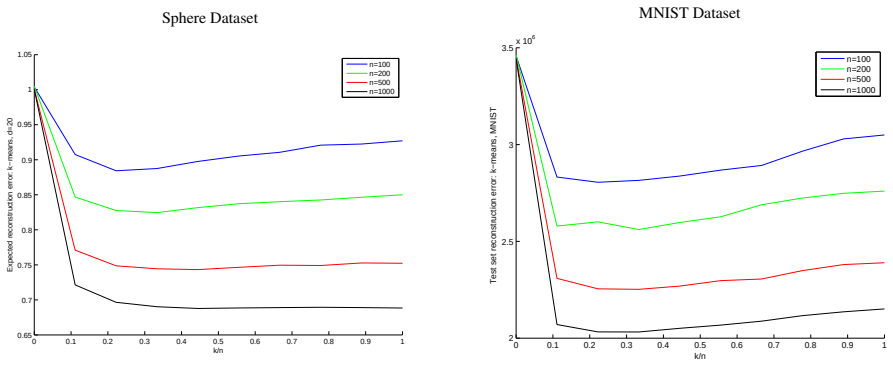


Figure 1: We consider the behavior of k-means for data sets obtained by sampling uniformly a 19 dimensional sphere embedded in $\mathbb{R}^{20}$ (left). For each value of k, k-means (with k-means++ seeding) is run 20 times, and the best solution kept. The reconstruction performance on a (large) hold-out 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 [22] report an average intrinsic dimension d for each digit to be between 10 and 13.

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 [18].

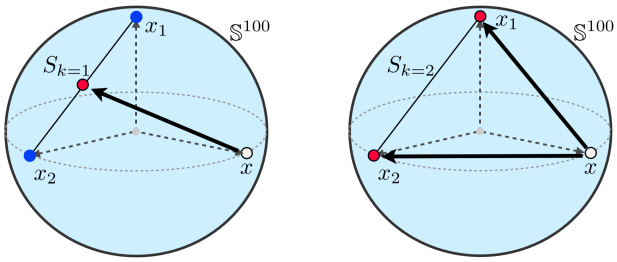
In machine learning, the properties of k-means have been studied, *for fixed* k, by considering the *excess* reconstruction error $\mathcal{E}_\rho(S_{n,k}) - \mathcal{E}_\rho(S_k)$. In particular, this quantity has been studied for $\mathcal{X} = \mathbb{R}^d$, and shown to be, with high probability, of order $\sqrt{kd/n}$, up-to logarithmic factors [31]. The case where $\mathcal{X}$ is a Hilbert space has been considered in [30, 8], 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 [7].

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 $k^{-2/d}$.)

The existence of such a tradeoff between the approximation, and the statistical errors may itself not be entirely obvious, see the discussion in [4]. 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



(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 $\mathcal{E}_\rho(S_{k=2}) \simeq 2$.

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 [18], 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 [30]. 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) [4]. More importantly, we are not aware of a lower bound for $\mathcal{E}_\rho(S_{n,k})$ itself. Indeed, as pointed out in [4], "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 [25]. 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++ [2], 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*

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

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 [36] 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 [16].*

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 $\kappa_{_G}$ 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 [16, 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 [19, 12], in contrast with worst-case methods that place a constraint on a maximum curvature, or minimum injectivity radius (for instance [1, 32].) 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] 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] Franz Aurenhammer. Voronoi diagrams: A survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23:345–405, September 1991.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

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

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

[12] 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.

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

[14] 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.

[15] 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.

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

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

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

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

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

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

[22] 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.

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

[24] 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.

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

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

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

[28] 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.

[29] 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.

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

[31] A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Trans.Inf.Th*, 56(11), 2010.

[32] 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.

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

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

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

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

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