On the Performance of Clustering in Hilbert Spaces *

Gérard BIAU [a], Luc DEVROYE [b], and Gábor LUGOSI [c,†]

[a] Institut de Mathématiques et de Modélisation de Montpellier
UMR CNRS 5149, Equipe de Probabilités et Statistique
Université Montpellier II, CC 051
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
biau@math.univ-montp2.fr

[b] School of Computer Science, McGill University
Montreal, Canada H3A 2K6
luc@cs.mcgill.ca

[c] ICREA and Department of Economics
Pompeu Fabra University
Ramon Trias Fargas 25-27, 08005 Barcelona, Spain
lugosi@upf.es

Abstract

Based on n randomly drawn vectors in a separable Hilbert space, one may construct a k-means clustering scheme by minimizing an empirical squared error. We investigate the risk of such a clustering scheme, defined as the expected squared distance of a random vector X from the set of cluster centers. Our main result states that, for an almost surely bounded X, the expected excess clustering risk is $O(\sqrt{1/n})$. Since clustering in high (or even infinite)-dimensional spaces may lead to severe computational problems, we examine the properties of a dimension reduction strategy for clustering based on Johnson-Lindenstrauss-type random projections. Our results reflect

*The second author's research was sponsored by NSERC Grant A3456 and FQRNT Grant 90-ER-0291. The third author acknowledges support by the Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324 and by the PASCAL Network of Excellence under EC grant no. 506778.

†Corresponding author.

a tradeoff between accuracy and computational complexity when one uses k-means clustering after random projection of the data to a low-dimensional space. We argue that random projections work better than other simplistic dimension reduction schemes.

Index Terms — Clustering, k-means, Vector quantization, Hilbert space, Empirical risk minimization, Random projections.

1 Introduction

Clustering is the problem of identifying groupings of similar points that are relatively isolated from each other, or, in other words, to partition the data into dissimilar groups of similar items (Duda, Hart, and Stork [14, Chapter 10]). This unsupervised learning paradigm is one of the most widely used techniques in exploratory data analysis. Across all disciplines, from social sciences to biology or computer science, practitioners try to get a first intuition about their data by identifying meaningful groups of observations. In data compression and information theory, the clustering problem is known as vector quantization or lossy data compression. Here, the goal is to find an efficient and compact representation from which the original observations can be reconstructed with a prescribed level of accuracy (see Gersho and Gray [18], Gray and Neuhoff [20], Linder [30]).

Whatever the terminology used, an observation is usually supposed to be a collection of numerical measurements represented by a d-dimensional vector. However, in some problems, input data items are in the form of random functions (speech recordings, spectra, images) rather than standard vectors, and this casts the clustering problem into the general class of functional data analysis. Even though in practice such observations are observed at discrete sampling points, the challenge in this context is to infer the data structure by exploiting the infinite-dimensional nature of the observations. The last few years have witnessed important developments in both the theory and practice of functional data analysis, and many traditional data analysis tools have been adapted to handle functional inputs. The book of Ramsay and Silverman [40] provides a comprehensive introduction to the area.

Interestingly, infinite-dimensional observations also arise naturally in the so-called kernel methods for general pattern analysis. These methods are based on the choice of a proper similarity measure, given by a positive definite kernel defined between pairs of objects of interest, to be used for inferring general types of relations. The key idea is to embed the observations at hand into a (possibly infinite-dimensional) Hilbert space, called the feature space, and to compute inner products efficiently directly from the original data items using the kernel function. The use of kernel methods for clustering is very natural, since the kernel defines similarities between observations, hence providing all the information needed to assess the quality of a clustering. For an exhaustive presentation of kernel methodologies and related algorithms, we refer the reader to Schölkopf and Smola [41], and Shawe-Taylor and Cristianini [42].

Motivated by this broad range of potential applications, we propose, in the present contribution, to investigate the general problem of clustering when observations take values in a separable Hilbert space $\mathcal{H}$. Thus, in our model, the data to be clustered is a sequence of independent $\mathcal{H}$-valued random observations $X_1, \ldots, X_n$ with the same distribution as a generic random variable X. The goal of clustering is to find an assignment of each variable to one of a finite number k of classes. Throughout, we will denote by $\langle \cdot, \cdot \rangle$ the inner product in $\mathcal{H}$, and by $\| \cdot \|$ the associated norm. In particular, we focus on the so-called k-means clustering, which prescribes a criterion for partitioning the sample $X_1, \ldots, X_n$ into k groups, or clusters, by minimizing the empirical squared norm criterion

$$W(\mathbf{c}, \mu_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|X_i - c_j\|^2 \tag{1}$$

over all possible choices of cluster centers $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{H}^k$. Here, μ_n is the empirical distribution of the data, defined by

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A\}}$$

for every Borel subset A of $\mathcal{H}$. Associated with each center c_j is the convex polyhedron S_j of all points in $\mathcal{H}$ closer to c_j than to any other center, called the Voronoi cell of c_j (ties are broken arbitrarily). Each X_i is assigned to its nearest center, and each empirically optimal center $c_{n1}, \ldots, c_{nk}$ is just the mean of those X_i's falling in the corresponding cluster.

The performance of a clustering scheme given by the collection $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{H}^k$ of cluster centers (and the associated Voronoi partition of $\mathcal{H}$) is measured by the *mean squared error* or *clustering risk*

$$W(\mathbf{c}, \mu) = \int \min_{j=1,\ldots,k} \|x - c_j\|^2 \, \mathrm{d}\mu(x). \tag{2}$$

The optimal clustering risk is defined as

$$W^*(\mu) = \inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu) \ .$$

Since the early work of Hartigan [21, 22] and Pollard [35, 36, 37], the problem of k-means clustering and its algorithmic counterparts have been considered by many authors. Convergence properties of the empirical minimizer $\mathbf{c}_n$ of the clustering risk have been mostly studied in the case when $\mathcal{H} = \mathbb{R}^d$. Consistency of $\mathbf{c}_n$ was shown by Pollard [35, 37] and Abaya and Wise [1] who prove that in $\mathbb{R}^d$, the infimum in the definition of $W^*(\mu)$ is achieved and that $W(\mathbf{c}_n, \mu) \to W^*(\mu)$ almost surely (a.s.) as $n \to \infty$, whenever $\mathbb{E}\|X\|^2 < \infty$. Rates of convergence and nonasymptotic performance bounds have been considered by Pollard [36], Chou [11], Linder, Lugosi, and Zeger [31], Bartlett, Linder, and Lugosi [7], Linder [29, 30], Antos [3], and Antos, Györfi, and György [4]. For example, it is shown in [7] that if μ is such that $\mathbb{P}\{\|X\| \le 1\} = 1$, then

$$\mathbb{E}W(\mathbf{c}_n, \mu) - W^*(\mu) \le C \min\left(\sqrt{\frac{kd}{n}}\ , \sqrt{\frac{k^{1-2/d} d \log n}{n}}\right) \tag{3}$$

where C is a universal constant. On the other hand, there exists a constant c and μ with $\mathbb{P}\{\|X\| \le 1\} = 1$ such that

$$\mathbb{E}W(\mathbf{c}_n, \mu) - W^*(\mu) \ge c\sqrt{\frac{k^{1-4/d}}{n}}\ .$$

For further references, consult Graf and Luschgy [19] and Linder [30]. Note that the upper bounds mentioned above become useless when d is very large. In our setup, in which we allow X to take values in an infinite-dimensional Hilbert space, substantially different arguments are called for. In Section 2, we prove that when $\mathbb{P}\{\|X\| \le 1\}$, the expected excess clustering risk $\mathbb{E}W(\mathbf{c}_n, \mu) - W^*(\mu)$ is bounded by $Ck/\sqrt{n}$, where C is a universal constant. We also examine the case where X is not bounded. In order to do this, we replace the VC and covering number arguments by techniques based on Rademacher averages.

It is important to point out that minimizing the empirical clustering risk is a computationally hard problem as all known algorithms have a computational complexity exponential in the dimension of the space. In practice approximate solutions are needed, often leading to local optima. In this study we ignore this computational issue and assume that an (approximate) minimizer of the empirical clustering risk can be found. In Section 3 we discuss computational complexity from a different point of view: we propose to use Johnson-Lindenstrauss-type random projections as an effective tool for dimension reduction. This is independent of the particular algorithm used to minimize the empirical squared error. Our results reflect a tradeoff between accuracy and computational complexity (measured as the dimension of the space in which the clustering is performed) when one uses k-means clustering after random projection of the data to a low-dimensional space. We argue that random projections work better than other simplistic dimension reduction schemes. Proofs are postponed to Section 4.

2 Clustering performance in Hilbert spaces

Recall that the training data consists of n independent $\mathcal{H}$-valued random observations $X_1, \ldots, X_n$ with the same distribution as a generic random variable X with distribution μ. Throughout the paper, we suppose that $\mathbb{E}\|X\|^2 < \infty$.

Let $\delta_n \geq 0$. A collection $\mathbf{c}_n = (c_{n1}, \ldots, c_{nk})$ of vectors is called a δ_n-minimizer of the empirical clustering risk (1) over $\mathcal{H}^k$ if

$$W(\mathbf{c}_n, \mu_n) \leq W^*(\mu_n) + \delta_n \ ,$$

where $W^*(\mu_n) = \inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu_n)$. When $\delta_n = 0$, $\mathbf{c}_n$ is called an empirical clustering risk minimizer. (Note that the existence of an empirical risk minimizer is guaranteed by the fact that μ_n is supported on at most n points.) The following consistency result states that the clustering risk $W(\mathbf{c}_n, \mu)$ should be close to the optimal risk $W^*(\mu) = \inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu)$ as the size of the training data grows.

Proposition 2.1 *Assume that* $\mathbb{E}\|X\|^2 < \infty$. *Let* $\mathbf{c}_n$ *be a* δ_n*-minimizer of the empirical clustering risk. If* $\lim_{n\to\infty} \delta_n = 0$, *then*

(i) $\lim_{n \to \infty} W(\mathbf{c}_n, \mu) = W^*(\mu) \quad$ *a.s.,*

and

(*ii*) $\lim_{n\to\infty} \mathbb{E}W(\mathbf{c}_n,\mu) = W^*(\mu)$.

In the Euclidean case, that is, when $\mathcal{H}$ is isomorphic to some $\mathbb{R}^d$ ($d \geq 1$), statement (i) is due to Pollard [37] (see also Pollard [35, 36]) and Abaya and Wise [1]. The proof of the general case is essentially similar–for the sake of completeness, we sketch it in Section 4, where we also show that (ii) is a consequence of (i) and some properties of the L_2 Wasserstein distance (Rachev and Rüschendorf [38, 39]) between μ and μ_n.

Clearly, the consistency result of Proposition 2.1 does not provide any information on how many training samples are needed to ensure that the clustering risk of the δ_n-optimal empirical centers is close to the optimum. The starting point of our analysis is the following elementary inequality (see Devroye, Györfi and Lugosi [13, Chapter 8]):

$$\begin{aligned} &\mathbb{E}W(\mathbf{c}_n,\mu) - W^*(\mu) \\ &\quad = \mathbb{E}\left[(W(\mathbf{c}_n,\mu) - W(\mathbf{c}_n,\mu_n)) + (W(\mathbf{c}_n,\mu_n) - W^*(\mu))\right] \\ &\quad \leq \mathbb{E}\sup_{\mathbf{c}\in\mathcal{H}^k}\big(W(\mathbf{c},\mu_n) - W(\mathbf{c},\mu)\big) + \sup_{\mathbf{c}\in\mathcal{H}^k}\mathbb{E}\big(W(\mathbf{c},\mu) - W(\mathbf{c},\mu_n)\big) + \delta_n \ . \end{aligned}$$

Roughly, this means that if we can guarantee that the uniform deviation

$$\mathbb{E}\sup_{\mathbf{c}\in\mathcal{H}^k}\big(W(\mathbf{c},\mu_n) - W(\mathbf{c},\mu)\big)$$

of estimated clustering risks from their true values is small, then the risk of the selected $\mathbf{c}_n$ is close to the best risk over all $\mathbf{c}$ in $\mathcal{H}^k$. In $\mathbb{R}^d$, this can be achieved by exploiting standard techniques from empirical process theory such as entropy methods or the Vapnik-Chervonenkis inequality [43]. However, in infinite-dimensional Hilbert spaces these techniques yield suboptimal bounds.

In the next theorem we exhibit a universal upper bound which is valid in any separable (possibly infinite-dimensional) Hilbert space. To achieve this goal, we use a measure of complexity of a function class, successfully employed in learning theory, known as the Rademacher averages (Bartlett, Boucheron, and Lugosi [6], Koltchinskii [26]; see also Bartlett and Mendelson [8], Bartlett [5], and Ambroladze, Parrado-Hernandez, and Shawe-Taylor [2]). Contrary to the VC techniques used to derive (3), the structural properties of Rademacher averages (see [8]) make it a suitable tool to derive a dimension-free bound. For any $R \geq 0$, let $\mathcal{P}(R)$ denote the set of probability distributions on $\mathcal{H}$ supported on $\mathcal{F}_R$, the closed ball of radius R centered at the origin. In other words, $\mu \in \mathcal{P}(R)$ is equivalent to

$$\mathbb{P}\big\{\|X\| \leq R\big\} = 1.$$

The main result of this section is the following:

Theorem 2.1 *Assume that $\mu \in \mathcal{P}(R)$. For any δ_n-minimizer $\mathbf{c}_n$ of the empirical clustering risk, we have*

$$\mathbb{E}W(\mathbf{c}_n, \mu) - W^*(\mu) \leq \frac{8kR\sqrt{\mathbb{E}\|X\|^2} + 4kR^2}{\sqrt{n}} + \delta_n \ ,$$

and, consequently,

$$\mathbb{E}W(\mathbf{c}_n, \mu) - W^*(\mu) \leq \frac{12kR^2}{\sqrt{n}} + \delta_n \ .$$

Remark. (DEPENDENCE ON k.) As we mentioned in the introduction, in $\mathbb{R}^d$ the expected excess risk may be bounded by a constant multiple of $\sqrt{kd/n}$. Even though the bound of Theorem 2.1 gets rid of the dependence on the dimension, this comes at the price of a worse dependence on the number k of clusters. The linear dependence on k is a consequence of our proof technique and we do not know whether it is possible to prove a bound of the order of $\sqrt{k/n}$. This would match the stated lower bound (when d is large).

Corollary 2.1 *Suppose that $\mu \in \mathcal{P}(R)$. Then, for any $x > 0$, with probability at least $1 - e^{-x}$,*

$$W(\mathbf{c}_n, \mu) - W^*(\mu) \leq \frac{12kR^2 + 4R^2\sqrt{2x}}{\sqrt{n}} + \delta_n \ .$$

The requirement $\mathbb{P}\{\|X\| \leq R\} = 1$ is standard in the clustering and data compression literature, where it is often called the peak power constraint. As stated by the next theorem, this requirement can be removed at the price of some technical complications.

Theorem 2.2 *Assume that $\mathbb{E}\|X\|^2 < \infty$. For any $x > 0$, there exist positive constants $C(\mu)$ and $N_0 = N_0(\mu, k, x)$ such that, for all $n \geq N_0$, with probability at least $1 - 2e^{-x}$,*

$$W(\mathbf{c}_n, \mu) - W^*(\mu) \leq C(\mu)\,\frac{k + \sqrt{x}}{\sqrt{n}} + \delta_n \ .$$

Remark. (FAST RATES.) In the finite-dimensional problem, there are some results showing that the convergence rate can be improved to $O(1/n)$ under certain assumptions on the distribution. Based on a result of Pollard [36] showing that for sources with continuous densities satisfying certain regularity properties, including the uniqueness of the optimal cluster centers, the suitably scaled difference of the optimal and empirically optimal centers has asymptotically multidimensional normal distribution, Chou [11] pointed out that for such distributions the expected excess risk decreases as $O(1/n)$. Further results were obtained by Antos, Györfi, and György [4] who prove that for any fixed distribution supported on a given finite set the convergence rate is $O(1/n)$, and provide for more general (finite-dimensional) distributions conditions implying an $O(\log n/n)$ rate of convergence. As pointed out by the authors, these conditions are, in general, difficult to verify. Recent general results of Koltchinskii [27] on empirical risk minimization show that whenever the optimal clustering centers are unique, and the distribution has a bounded support, the expected excess risk converges to zero at a rate faster than $n^{-1/2}$.

3 Random projections

In practice, handling high or infinite-dimensional data requires some dimension reduction techniques. A common practice to reduce dimension is by projecting the observations onto a lower-dimensional subspace that captures as much as possible the variation of the data. The most widely used methods achieving this goal are factorial methods, such as Principal Component Analysis (see, e.g., Mardia, Kent, and Bibby [33]) and its functional versions (see Ramsay and Silverman[40]). Unfortunately, most factorial methods in high-dimensional spaces are computationally expensive, with no guarantee that the distances between the original and projected observations are well preserved. In this section we argue that random projections to lower-dimensional subspaces are particularly well suited for clustering purposes.

In the random projection method, the original high-dimensional observations are projected onto a lower-dimensional space using a suitably scaled random matrix with independent, normally distributed entries. Random projections have been found to be a computationally efficient, yet sufficiently accurate method for dimensionality reduction. Promising experimental results are reported in Bingham and Mannila [9]. The key idea of random mapping arises from the Johnson-Lindenstrauss lemma [24], which states that any n point set in a Euclidean space can be embedded in a Euclidean space of dimension $O(\log n/\varepsilon^2)$ without distorting the distances between any pair of points by more than a factor of $1 \pm \varepsilon$, for any $\varepsilon \in (0,1)$. The original proof of Johnson and Lindenstrauss was simplified by Frankl and Maehara [16, 17], and further worked out using probabilistic techniques by Dasgupta and Gupta [12].

The Johnson-Lindenstrauss lemma is usually stated in the Euclidean setting, that is, when $\mathcal{H} \simeq \mathbb{R}^d$ ($d \geq 1$). The general case requires some simple adaptations, detailed below. Recently, this lemma has found several applications, including Lipschitz embeddings of graphs into normed spaces (Linial, London, and Rabinovich [32]) and searching for approximate nearest neighbors (see Kleinberg [25], Indyk and Motwani [23]).

To describe the random projections we suppose, without loss of generality, that $\mathcal{H}$ is infinite-dimensional. Since $\mathcal{H}$ is assumed to be separable, we may identify it with the space ℓ^2 of all sequences $x = (x_\alpha)_{\alpha \geq 1}$ such that $\sum_{\alpha=1}^{\infty} x_\alpha^2 < \infty$. Each data point X_i ($i = 1, \ldots, n$) is now represented by a vector in ℓ^2, denoted, with a slight abuse of notation, by $X_i = (X_i^{(1)}, X_i^{(2)}, \ldots)$.

Let s be a positive integer, and let $(N_{1\alpha})_{\alpha \geq 1}, \ldots, (N_{s\alpha})_{\alpha \geq 1}$ be s independent sequences of independent centered normal random variables with variance $1/s$. For each $D \geq 1$, set

$$X_{i,j}^D = \sum_{\alpha=1}^{D} N_{j\alpha} X_i^{(\alpha)}, \quad j = 1, \ldots, s,$$

or, in matrix form,

$$\begin{pmatrix} X_{i,1}^D \\ \vdots \\ X_{i,s}^D \end{pmatrix} = \begin{pmatrix} N_{11} & \ldots & N_{1D} \\ \vdots & \vdots & \vdots \\ N_{s1} & \ldots & N_{sD} \end{pmatrix} \begin{pmatrix} X_i^{(1)} \\ \vdots \\ X_i^{(D)} \end{pmatrix}.$$

Conditioned on $X_1, \ldots, X_n$, for fixed i and j, the sequence $(X_{i,j}^D)_{D \geq 1}$ is a sum of independent centered random variables, and therefore it is a martingale. Moreover, denoting by $\mathbb{E}_N$ expectation taken with respect to the normal random variables (conditioned on $X_1, \ldots, X_n$),

$$\mathbb{E}_N \left(X_{i,j}^D\right)^2 = \sum_{\alpha=1}^{D} \frac{\left(X_i^{(\alpha)}\right)^2}{s} \leq \frac{\|X_i\|^2}{s}.$$

Thus, the sequence $(X_{i,j}^D)_{D\geq 1}$ is a martingale bounded in L_2. Consequently, it converges almost surely and in L_2 to some random variable $X_{i,j}$ (see, e.g., Williams [45]). Moreover, there exists a random variable $Z_{i,j}$ in L_2 such that $|X_{i,j}^D| \leq Z_{i,j}$. It follows by dominated convergence that

$$\lim_{D\to\infty} \mathbb{E}_N \big(X_{i,j}^D\big)^2 \to \mathbb{E}_N X_{i,j}^2 = \frac{\|X_i\|^2}{s} \ .$$

The vector $\overline{X}_i = (X_{i,1}, \ldots, X_{i,s}) \in \mathbb{R}^s$ may be regarded as a *random projection* of X_i to an s-dimensional subspace. Clearly, for fixed X_i, each component $X_{i,j}$ is a normal random variable with mean 0 and variance $\|X_i\|^2/s$. Therefore, $\mathbb{E}_N\|\overline{X}_i\|^2 = \|X_i\|^2$ and (with X_i fixed) $s\|\overline{X}_i\|^2/\|X_i\|^2$ has χ^2 distribution with s degrees of freedom. Similarly, for any $i \neq i' \in \{1, \ldots, n\}$,

$$\mathbb{E}_N\|\overline{X}_i - \overline{X}_{i'}\|^2 = \|X_i - X_{i'}\|^2$$

and $s\|\overline{X}_i - \overline{X}_{i'}\|^2/\|X_i - X_{i'}\|^2$ has χ^2 distribution with s degrees of freedom. Now by a simple Chernoff bound (Chernoff [10]) for the χ^2 distribution, we have

$$\mathbb{P}_N\left\{\frac{\|\overline{X}_i - \overline{X}_{i'}\|^2}{\|X_i - X_{i'}\|^2} - 1 > \varepsilon\right\} \leq \exp\left[\frac{s}{2}\big(-\varepsilon + \ln(1+\varepsilon)\big)\right] \ ,$$

and, similarly,

$$\mathbb{P}_N\left\{\frac{\|\overline{X}_i - \overline{X}_{i'}\|^2}{\|X_i - X_{i'}\|^2} - 1 < -\varepsilon\right\} \leq \exp\left[\frac{s}{2}\big(\varepsilon + \ln(1-\varepsilon)\big)\right] \ .$$

By the union bound we obtain the following:

Theorem 3.1 (Johnson-Lindenstrauss lemma) *Let $\mathcal{H}$ be a separable Hilbert space. For any $\varepsilon, \delta \in (0,1)$ and any positive integer n, let s be a positive integer such that*

$$s \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log \frac{n}{\sqrt{\delta}} \ .$$

Define a linear map $f : \mathcal{H} \to \mathbb{R}^s$ as a random projection described above. Then, for any set $\mathcal{D}$ of n points in $\mathcal{H}$, with probability at least $1-\delta$, for all $(u,v) \in D^2$,

$$(1-\varepsilon)\|u-v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2 \ .$$

Thus, random projections approximately preserve pairwise distances in the data set, and therefore are particularly well suited for the purposes of k-means clustering.

Let $s = \lceil 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log(n/\sqrt{\delta}) \rceil$ and let $\overline{X}_1, \ldots, \overline{X}_n \in \mathbb{R}^s$ be the randomly projected data points $X_1, \ldots, X_n$. We propose the following algorithm:

Determine a collection $\overline{\mathbf{c}}_n = (\overline{c}_{n1}, \ldots, \overline{c}_{nk}) \in (\mathbb{R}^s)^k$ of cluster centers which minimizes the empirical clustering risk in $\mathbb{R}^s$ based on the projected data $\overline{X}_1, \ldots, \overline{X}_n$, and let $\overline{S}_{n1}, \ldots, \overline{S}_{nk} \subset \mathbb{R}^s$ be the associated Voronoi cells. Define the cluster centers in $\mathcal{H}$ by

$$\widehat{c}_{nj} = \frac{\sum_{i=1}^n X_i \mathbb{I}_{\{\overline{X}_i \in \overline{S}_{nj}\}}}{\sum_{i=1}^n \mathbb{I}_{\{\overline{X}_i \in \overline{S}_{nj}\}}}, \quad j = 1, \ldots, k,$$

and denote by $\widehat{\mathbf{c}}_n$ the collection of these k centers. The cluster centers then determine the associated Voronoi partition of $\mathcal{H}$ into k cells. The following result ensures that replacing the empirically optimal cluster centers $\mathbf{c}_n$ by $\widehat{\mathbf{c}}_n$ does not harm too much.

Theorem 3.2 *Fix the sample $\mathcal{D} = \{X_1, \ldots, X_n\}$. For any $\varepsilon, \delta \in (0,1)$, let the positive integer s and the random projection be as in Theorem 3.1 above. Then, with probability at least $1-\delta$,*

$$W(\widehat{\mathbf{c}}_n, \mu_n) \leq \frac{1+\varepsilon}{1-\varepsilon} W(\mathbf{c}_n, \mu_n) \ .$$

We may combine this with results of the previous sections to establish performance bounds for the clustering algorithm based on random projections.

Corollary 3.1 *Assume that $\mu \in \mathcal{P}(R)$, and let $\varepsilon \in (0, 1/2)$ and $\delta \in (0,1)$. Define $s = \lceil 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log(n/\sqrt{\delta}) \rceil$, and consider the clustering centers $\widehat{\mathbf{c}}_n$ found by the clustering algorithm based on random projections described above. Then, for any $x > 0$, with probability at least $1 - e^{-x}$ (with respect to the random sample), with probability at least $1-\delta$ (with respect to the random projections),*

$$W(\widehat{\mathbf{c}}_n, \mu) - W^*(\mu) \leq \frac{24kR^2 + 12R^2\sqrt{2x}}{\sqrt{n}} + 4\varepsilon R^2 \ .$$

Proof. According to Corollary 2.1, with probability at least $1 - e^{-x}$ (with respect to the random sample),

$$W(\mathbf{c}_n, \mu) - W^*(\mu) \leq \frac{12kR^2 + 4R^2\sqrt{2x}}{\sqrt{n}} .$$

Thus,

$$W(\widehat{\mathbf{c}}_n, \mu) - W^*(\mu) \leq W(\widehat{\mathbf{c}}_n, \mu) - W(\mathbf{c}_n, \mu) + \frac{12kR^2 + 4R^2\sqrt{2x}}{\sqrt{n}} .$$

Let us decompose the term $W(\widehat{\mathbf{c}}_n, \mu) - W(\mathbf{c}_n, \mu)$ as follows:

$$\begin{aligned} & W(\widehat{\mathbf{c}}_n, \mu) - W(\mathbf{c}_n, \mu) \\ & \quad = W(\widehat{\mathbf{c}}_n, \mu) - W(\widehat{\mathbf{c}}_n, \mu_n) + W(\widehat{\mathbf{c}}_n, \mu_n) - W(\mathbf{c}_n, \mu_n) \\ & \qquad + W(\mathbf{c}_n, \mu_n) - W(\mathbf{c}_n, \mu) . \end{aligned} \tag{4}$$

Theorem 3.2 allows one to upper bound the second term: with probability at least $1 - \delta$ (with respect to the random projections),

$$W(\widehat{\mathbf{c}}_n, \mu_n) - W(\mathbf{c}_n, \mu_n) \leq 4\varepsilon R^2 .$$

With respect to the first term in (4), we note that

$$\begin{aligned} W(\widehat{\mathbf{c}}_n, \mu) - W(\widehat{\mathbf{c}}_n, \mu_n) &\leq \sup_{\mathbf{c} \in \mathcal{H}^k} \big(W(\mathbf{c}, \mu) - W(\mathbf{c}, \mu_n) \big) \\ &\leq \mathbb{E} \sup_{\mathbf{c} \in \mathcal{H}^k} \big(W(\mathbf{c}, \mu) - W(\mathbf{c}, \mu_n) \big) + 4R^2 \sqrt{\frac{2x}{n}} , \end{aligned}$$

where the last inequality arises from a standard application of the bounded differences concentration inequality. Bounding the third term in (4) using the same principle and applying Lemma 4.3 leads to the conclusion. $\square$

Remark. This corollary shows a tradeoff between performance and computational complexity. If one projects onto a space of dimension $O(\log n/\varepsilon^2)$, the price to pay is an excess clustering risk of the order of ε. With other, simplistic, dimension reduction techniques no such bounds can be proven. Consider, for example, a commonly used dimension reduction technique that simply keeps the first m_n components of X (in an orthonormal representation of X), where m_n is some pre-specified positive integer, possibly growing with the sample size. Then it is easy to see that, even though almost sure convergence of the clustering risk to the optimum may be achieved, convergence may be arbitrarily slow. For concreteness, assume that X takes its values in ℓ^2. Let $k = 1$ and suppose that μ is concentrated on just one point, $c = (c_1, c_2, \ldots)$, where the c_α's are nonnegative numbers with $\sum_\alpha c_\alpha^2 < \infty$. Then the optimal cluster center is clearly c, but by truncating at the first m_n components the best one can do is to take $c' = (c_1, \ldots, c_{m_n}, 0, 0, \ldots)$, giving a clustering risk equal to $\sum_{\alpha > m_n} c_\alpha^2$. No matter how fast m_n grows, the clustering risk may converge to zero at an arbitrarily slow rate.

Remark. In the corollary above we assumed, for simplicity, that $\mu \in \mathcal{P}(R)$. In this case $W(\mathbf{c}_n, \mu_n) \leq R^2$ with probability one, and by Theorem 3.2, $\hat{\mathbf{c}}_n$ is a $4\varepsilon R^2$-minimizer of the empirical clustering risk, and Theorem 2.1 implies the corollary. However, it is easy to generalize the statement, since, as it is clear from the proof of Theorem 2.2, if $\mathbb{E}\|X\|^2 < \infty$, then $W(\mathbf{c}_n, \mu_n)$ is bounded, eventually, almost surely, by a constant, so Theorem 2.2 implies an analog statement with the appropriate trivial modifications.

Remark. (COMPUTATIONAL MODEL.) There is no standard computational model to handle Hilbert-space valued data. In the algorithm described above we assumed implicitly that the random projections can be calculated easily. This may not be unrealistic if an orthonormal representation of the X_i's is available. Instead of discussing such details, we simply assume the existence of a computational oracle that computes a random projection at a unit cost. In this paper we have ignored some other important issues of computational complexity. It is well known that finding the empirically optimal cluster centers is, in general, NP hard. In practice, approximate solutions have been used to avoid prohibitive complexity. Of course, dimension reduction techniques are useful to lower computational complexity, but in this paper we do not pursue this issue further, and just investigate theoretical properties of minimizers of the empirical clustering risk in the randomly projected subspace.

4 Proofs

4.1 Sketch of proof of Proposition 2.1

The following sketch is based on arguments of Pollard [35] (see also Theorem 2 in Linder [31]). Note that in this section we only use the fact that $\mathcal{H}$ is a separable and complete vector space.

The basic idea is that for large n the empirical distribution μ_n is a good estimate of μ, so the optimal clustering for μ_n should provide a good approximation to the optimal clustering for μ. Recall that the L_2 Wasserstein distance (Rachev and Rüschendorf [38, 39]) between two probability measures μ_1 and μ_2 on $\mathcal{H}$, with finite second moment, is defined as

$$\gamma(\mu_1, \mu_2) = \inf_{\nu \in \mathcal{M}(\mu_1, \mu_2)} \Big[\int \|x - y\|^2 \, \mathrm{d}\nu(x, y) \Big]^{1/2},$$

where $\mathcal{M}(\mu_1, \mu_2)$ is the set of all laws on $\mathcal{H} \times \mathcal{H}$ with marginals μ_1 and μ_2. Equivalently,

$$\gamma(\mu_1, \mu_2) = \inf_{X \sim \mu_1, Y \sim \mu_2} \big(\mathbb{E}\|X - Y\|^2 \big)^{1/2},$$

where the infimum is taken over all joint distributions of two random $\mathcal{H}$-valued random vectors X and Y such that X has distribution μ_1 and Y has distribution μ_2. It may be proven (Rachev and Rüschendorf [38]) that γ is a metric on the space of probability distributions on $\mathcal{H}$ with finite second moment, and that the infimum in the definition of $\gamma(\mu_1, \mu_2)$ is attained.

The following inequality (see Linder [30, Lemma 3]) shows that if two distributions μ_1 and μ_2 are close in γ metric, then their clustering error are also similar:

$$\sup_{\mathbf{c} \in \mathcal{H}^k} \big| W(\mathbf{c}, \mu_1)^{1/2} - W(\mathbf{c}, \mu_2)^{1/2} \big| \leq \gamma(\mu_1, \mu_2) \ . \tag{5}$$

Lemma 4.1 below relates the clustering risk $W(\mathbf{c}_n, \mu)$ of a δ_n-minimizer of the empirical clustering risk to the optimal risk $\inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu)$ in terms of the γ distance between the source distribution μ and the empirical distribution μ_n.

Lemma 4.1 *Let $\mathbf{c}_n$ be a δ_n-minimizer of the empirical clustering risk. Then*

$$W(\mathbf{c}_n, \mu)^{1/2} - \big[\inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu) \big]^{1/2} \leq 2\gamma(\mu, \mu_n) + \sqrt{\delta_n}.$$

Proof of Lemma 4.1. Let $\varepsilon > 0$ be arbitrary, and let $\mathbf{c}^*$ be any element of $\mathcal{H}^k$ satisfying

$$\inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu) \leq W(\mathbf{c}^*, \mu) < \inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu) + \varepsilon \ .$$

For any $t \in \mathbb{R}$, we set $(t)_+ = \max(t, 0)$. Then

$$
\begin{aligned}
& W(\mathbf{c}_n, \mu)^{1/2} - \big[\inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu)\big]^{1/2} \\
& \quad \leq W(\mathbf{c}_n, \mu)^{1/2} - \big[W(\mathbf{c}^*, \mu) - \varepsilon\big]_+^{1/2} \\
& \quad \leq W(\mathbf{c}_n, \mu)^{1/2} - W(\mathbf{c}^*, \mu)^{1/2} + \sqrt{\varepsilon} \\
& \quad = W(\mathbf{c}_n, \mu)^{1/2} - W(\mathbf{c}_n, \mu_n)^{1/2} + W(\mathbf{c}_n, \mu_n)^{1/2} - W(\mathbf{c}^*, \mu)^{1/2} + \sqrt{\varepsilon} \\
& \quad \leq W(\mathbf{c}_n, \mu)^{1/2} - W(\mathbf{c}_n, \mu_n)^{1/2} + W(\mathbf{c}^*, \mu_n)^{1/2} - W(\mathbf{c}^*, \mu)^{1/2} + \sqrt{\varepsilon} + \sqrt{\delta_n} \\
& \qquad \text{(by definition of } \mathbf{c}_n) \\
& \quad \leq 2\gamma(\mu, \mu_n) + \sqrt{\varepsilon} + \sqrt{\delta_n}\ ,
\end{aligned}
$$

where the last inequality follows from (5). ■

The two statements of Proposition 2.1 are immediate consequences of Lemma 4.1 and the following lemma.

Lemma 4.2 (i) $\lim_{n\to\infty} \gamma(\mu, \mu_n) = 0 \quad a.s. \qquad (ii)$ $\lim_{n\to\infty} \mathbb{E}\gamma^2(\mu, \mu_n) = 0$.

Proof of Lemma 4.2. Statement (i) is proved in detail in Linder [30, Theorem 2], and is based on the fact that the empirical measure μ_n converges to μ almost surely (see also Dudley [15, Chapter 11]).

Statement (ii) is less standard. To prove it, we denote by $\mathcal{M}(\mu, \mu_n)$ the (random) set of all laws on $\mathcal{H} \times \mathcal{H}$ with marginals μ and μ_n. By definition, the squared L_2 Wasserstein distance between μ and μ_n reads

$$
\gamma^2(\mu, \mu_n) = \inf_{\nu \in \mathcal{M}(\mu, \mu_n)} \int \|x - y\|^2 \, \mathrm{d}\nu(x, y)\ .
$$

Let C be an arbitrary nonnegative constant, and let $\mathcal{A}$ be the subset of $\mathcal{H}\times\mathcal{H}$ defined by

$$
\mathcal{A} = \big\{(x, y) \in \mathcal{H} \times \mathcal{H} : \max(\|x\|, \|y\|) \leq C\big\}\ .
$$

We may write, for any $\nu \in \mathcal{M}(\mu, \mu_n)$,

$$
\begin{aligned}
& \int \|x - y\|^2 \, \mathrm{d}\nu(x, y) \\
& \quad = \int_{\mathcal{A}} \|x - y\|^2 \, \mathrm{d}\nu(x, y) + \int_{\mathcal{A}^c} \|x - y\|^2 \, \mathrm{d}\nu(x, y)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \int_{\mathcal{A}} \|x-y\|^2 \,\mathrm{d}\nu(x,y) + 2\int_{\mathcal{A}^c} \|x\|^2 \,\mathrm{d}\nu(x,y) + 2\int_{\mathcal{A}^c} \|y\|^2 \,\mathrm{d}\nu(x,y) \\
&\quad (\text{since } \|x-y\|^2 \leq 2\|x\|^2 + 2\|y\|^2) \\
&\leq \int_{\mathcal{A}} \|x-y\|^2 \,\mathrm{d}\nu(x,y) \\
&\quad + 2\int \|x\|^2 \mathbb{I}_{\{\|x\|>C\}} \,\mathrm{d}\mu(x) + 2\int \|x\|^2 \mathbb{I}_{\{\|x\|\leq C, \|y\|>C\}} \,\mathrm{d}\nu(x,y) \\
&\quad + 2\int \|y\|^2 \mathbb{I}_{\{\|y\|>C\}} \,\mathrm{d}\mu_n(y) + 2\int \|y\|^2 \mathbb{I}_{\{\|x\|>C, \|y\|\leq C\}} \,\mathrm{d}\nu(x,y) \\
&\leq \int_{\mathcal{A}} \|x-y\|^2 \,\mathrm{d}\nu(x,y) \\
&\quad + 2\int \|x\|^2 \mathbb{I}_{\{\|x\|>C\}} \,\mathrm{d}\mu(x) + 2C^2 \mu_n\big\{\|y\| > C\big\} \\
&\quad + 2\int \|y\|^2 \mathbb{I}_{\{\|y\|>C\}} \,\mathrm{d}\mu_n(y) + 2C^2 \mu\big\{\|x\| > C\big\}.
\end{aligned}
$$

Consequently, by Markov's inequality,

$$
\begin{aligned}
&\int \|x-y\|^2 \,\mathrm{d}\nu(x,y) \\
&\quad \leq \int_{\mathcal{A}} \|x-y\|^2 \,\mathrm{d}\nu(x,y) \\
&\qquad + 2\int \|x\|^2 \mathbb{I}_{\{\|x\|>C\}} \,\mathrm{d}\mu(x) + 2\int \|y\|^2 \mathbb{I}_{\{\|y\|>C\}} \,\mathrm{d}\mu_n(y) \\
&\qquad + 2\int \|y\|^2 \mathbb{I}_{\{\|y\|>C\}} \,\mathrm{d}\mu_n(y) + 2\int \|x\|^2 \mathbb{I}_{\{\|x\|>C\}} \,\mathrm{d}\mu(x).
\end{aligned}
$$

Thus, taking the infimum over $\mathcal{M}(\mu, \mu_n)$ on both sides and taking expectations with respect to the X_i's, we deduce that

$$
\begin{aligned}
&\mathbb{E}\gamma^2(\mu, \mu_n) \\
&\quad \leq \mathbb{E} \inf_{\nu \in \mathcal{M}(\mu,\mu_n)} \int_{\mathcal{A}} \|x-y\|^2 \,\mathrm{d}\nu(x,y) + 8\int \|x\|^2 \mathbb{I}_{\{\|x\|>C\}} \,\mathrm{d}\mu(x).
\end{aligned}
$$

For a fixed $C \geq 0$, the first term tends to 0 as $n \to \infty$ according to statement (i) and the Lebesgue dominated convergence theorem. Since $\int \|x\|^2 \,\mathrm{d}\mu(x) < \infty$, the second term of the right-hand side vanishes as $C \to \infty$, and this concludes the proof of Lemma 4.2. ■

4.2 Proof of Theorem 2.1

An important consequence of the assumption $\mu \in \mathcal{P}(R)$ is that it is sufficient for our purpose to consider only cluster centers in the (closed and convex) ball $\mathcal{F}_R$ of the Hilbert space $\mathcal{H}$, since otherwise projecting any center that is not in $\mathcal{F}_R$ to the surface of $\mathcal{F}_R$ clearly reduces the clustering risk. Since $\mu \in \mathcal{P}(R)$, we also have $\mu_n \in \mathcal{P}(R)$ a.s., and, similarly, we only need to search for empirical centers living in $\mathcal{F}_R$.

Note that, for any $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{F}_R^k$, the risk $W(\mathbf{c}, \mu)$ defined in (2) may be rewritten as

$$W(\mathbf{c}, \mu) = \mathbb{E}\|X\|^2 + \mathbb{E}\left\{ \min_{j=1,\ldots,k} \Big[-2\langle X, c_j\rangle + \|c_j\|^2 \Big] \right\} .$$

Minimizing $W(\mathbf{c}, \mu)$ is therefore equivalent to minimizing the functional

$$\overline{W}(\mathbf{c}, \mu) = \mathbb{E}\left\{ \min_{j=1,\ldots,k} \Big[-2\langle X, c_j\rangle + \|c_j\|^2 \Big] \right\}$$

over all $\mathbf{c} \in \mathcal{F}_R^k$. Similarly, minimizing the empirical risk (1) is the same as minimizing

$$\overline{W}(\mathbf{c}, \mu_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \Big[-2\langle X, c_j\rangle + \|c_j\|^2 \Big].$$

Moreover, for any δ_n-minimizer $\mathbf{c}_n$ of the empirical risk,

$$W(\mathbf{c}_n, \mu) - \inf_{\mathbf{c}\in\mathcal{F}_R^k} W(\mathbf{c}, \mu) = \overline{W}(\mathbf{c}_n, \mu) - \inf_{\mathbf{c}\in\mathcal{F}_R^k} \overline{W}(\mathbf{c}, \mu), \tag{6}$$

and

$$\begin{aligned} &\mathbb{E}\overline{W}(\mathbf{c}_n, \mu) - \inf_{\mathbf{c}\in\mathcal{F}_R^k} \overline{W}(\mathbf{c}, \mu) \\ &\quad \le \mathbb{E} \sup_{\mathbf{c}\in\mathcal{F}_R^k} \big(\overline{W}(\mathbf{c}, \mu_n) - \overline{W}(\mathbf{c}, \mu)\big) + \sup_{\mathbf{c}\in\mathcal{F}_R^k} \mathbb{E}\big(\overline{W}(\mathbf{c}, \mu) - \overline{W}(\mathbf{c}, \mu_n)\big) + \delta_n. \end{aligned} \tag{7}$$

We are thus interested in upper bounds for the maximal deviation

$$\mathbb{E} \sup_{\mathbf{c}\in\mathcal{F}_R^k} \big(\overline{W}(\mathbf{c}, \mu_n) - \overline{W}(\mathbf{c}, \mu)\big) .$$

Note that the second term on the right-hand side of (7) is much easier and can trivially be bounded by the upper bound we obtain for the first term below.

Let $\sigma_1, \ldots, \sigma_n$ be n independent Rademacher random variables, that is $\{\pm 1\}$-valued independent random variables such that $\mathbb{P}\{\sigma_i = -1\} = \mathbb{P}\{\sigma_i = +1\} = 1/2$, independent of the X_i's. Let $\mathcal{G}$ be a class of real-valued functions defined on the Hilbert space $\mathcal{H}$. Then the Rademacher averages of $\mathcal{G}$ are defined by

$$R_n(\mathcal{G}) = \mathbb{E} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(X_i) \ .$$

The proof of Theorem 2.1 relies on the following lemma.

Lemma 4.3 *For every c in $\mathcal{F}_R$, let ℓ_c be the real-valued map defined by*

$$\ell_c(x) = -2\langle x, c\rangle + \|c\|^2, \quad x \in \mathcal{H} \ .$$

Then the following three statements hold:

(i)

$$\mathbb{E} \sup_{\mathbf{c} \in \mathcal{F}_R^k} \left(\overline{W}(\mathbf{c}, \mu_n) - \overline{W}(\mathbf{c}, \mu) \right) \leq 2\mathbb{E} \sup_{\mathbf{c} \in \mathcal{F}_R^k} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big[\min_{j=1,\ldots,k} \ell_{c_j}(X_i) \big].$$

(ii)

$$\mathbb{E} \sup_{\mathbf{c} \in \mathcal{F}_R^k} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big[\min_{j=1,\ldots,k} \ell_{c_j}(X_i) \big] \leq 2k \Big[\mathbb{E} \sup_{c \in \mathcal{F}_R} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle X_i, c\rangle + \frac{R^2}{2\sqrt{n}} \Big].$$

(iii)

$$\mathbb{E} \sup_{c \in \mathcal{F}_R} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle X_i, c\rangle \leq R \sqrt{\frac{\mathbb{E}\|X\|^2}{n}}.$$

Proof of Lemma 4.3.

(i) Let $X_1', \ldots, X_n'$ be an independent copy of $X_1, \ldots, X_n$, independent of the σ_i's. Then, by a standard symmetrization argument, we may write

$$\begin{aligned}
&\mathbb{E} \sup_{\mathbf{c}\in\mathcal{F}_R^k} \big(\overline{W}(\mathbf{c},\mu_n) - \overline{W}(\mathbf{c},\mu)\big) \\
&\quad \le \mathbb{E} \sup_{\mathbf{c}\in\mathcal{F}_R^k} \frac{1}{n}\sum_{i=1}^n \sigma_i \big[\min_{j=1,\ldots,k} \ell_{c_j}(X_i) - \min_{j=1,\ldots,k} \ell_{c_j}(X_i')\big] \\
&\quad \le \mathbb{E} \sup_{\mathbf{c}\in\mathcal{F}_R^k} \frac{1}{n}\sum_{i=1}^n \sigma_i \big[\min_{j=1,\ldots,k} \ell_{c_j}(X_i)\big] + \mathbb{E} \sup_{\mathbf{c}\in\mathcal{F}_R^k} \frac{1}{n}\sum_{i=1}^n (-\sigma_i) \big[\min_{j=1,\ldots,k} \ell_{c_j}(X_i')\big] \\
&\quad = 2\mathbb{E} \sup_{\mathbf{c}\in\mathcal{F}_R^k} \frac{1}{n}\sum_{i=1}^n \sigma_i \big[\min_{j=1,\ldots,k} \ell_{c_j}(X_i)\big].
\end{aligned}$$

(ii) To prove statement (ii), we will make use of the following properties of the Rademacher averages. Property 1 is a consequence of the contraction principle, due to Ledoux and Talagrand [28]. (Note that our definition of a Rademacher average does not involve absolute values, contrary to a perhaps more usual usage. This allows us to save a factor of 2 in the contraction principle.)

1. $R_n\big(|\mathcal{G}_1|\big) \le R_n(\mathcal{G}_1)$, where $|\mathcal{G}_1| = \big\{|g_1| : g_1 \in \mathcal{G}_1\big\}$.
2. $R_n(\mathcal{G}_1 \oplus \mathcal{G}_2) \le R_n(\mathcal{G}_1) + R_n(\mathcal{G}_2)$, where $\mathcal{G}_1 \oplus \mathcal{G}_2 = \{g_1 + g_2 : (g_1, g_2) \in \mathcal{G}_1 \times \mathcal{G}_2\}$.

The proof proceeds by induction on k. For $k = 1$, we have

$$\begin{aligned}
\mathbb{E} \sup_{c\in\mathcal{F}_R} \frac{1}{n}\sum_{i=1}^n \sigma_i \ell_c(X_i) \quad &= \mathbb{E} \sup_{c\in\mathcal{F}_R} \frac{1}{n}\sum_{i=1}^n \sigma_i \big[-2\langle X_i, c\rangle + \|c\|^2\big] \\
&\le 2\mathbb{E} \sup_{c\in\mathcal{F}_R} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle X_i, c\rangle + \mathbb{E} \sup_{c\in\mathcal{F}_R} \frac{\|c\|^2}{n}\sum_{i=1}^n \sigma_i \\
&\le 2\mathbb{E} \sup_{c\in\mathcal{F}_R} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle X_i, c\rangle + \frac{R^2}{n}\mathbb{E}\Big|\sum_{i=1}^n \sigma_i\Big| \\
&\quad \text{(since } \sup_{c\in\mathcal{F}_R} \|c\|^2 \le R^2\text{)}
\end{aligned}$$

$$\leq 2\mathbb{E}\sup_{c\in\mathcal{F}_R}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle X_i,c\rangle+\frac{R^2}{\sqrt{n}}$$
(by the Cauchy-Schwarz inequality).

For $k=2$, we obtain

$$\mathbb{E}\sup_{(c_1,c_2)\in\mathcal{F}_R^2}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\big[\min_{j=1,2}\ell_{c_j}(X_i)\big]$$
$$=\mathbb{E}\sup_{(c_1,c_2)\in\mathcal{F}_R^2}\frac{1}{2n}\sum_{i=1}^{n}\sigma_i\big[\ell_{c_1}(X_i)+\ell_{c_2}(X_i)-\big|\ell_{c_1}(X_i)-\ell_{c_2}(X_i)\big|\big]$$
(using $\min(a,b)=(a+b)/2-|a-b|/2$)
$$\leq 2\mathbb{E}\sup_{c\in\mathcal{F}_R}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell_c(X_i)$$
(by properties 1 and 2 above)
$$\leq 4\Big[\mathbb{E}\sup_{c\in\mathcal{F}_R}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle X_i,c\rangle+\frac{R^2}{2\sqrt{n}}\Big].$$

The recurrence rule is straightforward, using the arguments presented for $k=1$, $k=2$, and the fact that for any $(z_1,\ldots,z_k)\in\mathbb{R}^k$,

$$\min(z_1,\ldots,z_k)=\min\big(\min(z_1,\ldots,z_{\lfloor k/2\rfloor}),\min(z_{\lfloor k/2\rfloor+1},\ldots,z_k)\big)\ .$$

(iii)

$$\mathbb{E}\sup_{c\in\mathcal{F}_R}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle X_i,c\rangle=\mathbb{E}\sup_{c\in\mathcal{F}_R}\frac{1}{n}\Big\langle\sum_{i=1}^{n}\sigma_iX_i,c\Big\rangle$$
$$=\frac{R}{n}\mathbb{E}\Big\|\sum_{i=1}^{n}\sigma_iX_i\Big\|$$
$$\leq\frac{R}{n}\sqrt{\mathbb{E}\Big\|\sum_{i=1}^{n}\sigma_iX_i\Big\|^2}$$
(by the Cauchy-Schwarz inequality)
$$=R\sqrt{\frac{\mathbb{E}\|X\|^2}{n}}.$$

Theorem 2.1 is a consequence of inequality (6), inequality (7), and Lemma 4.3 $(i) - (iii)$. ■

4.3 Proof of Corollary 2.1

The proof is immediate from Theorem 2.1 and a standard application of the bounded differences concentration inequality (see, e.g., McDiarmid [34]).

4.4 Proof of Theorem 2.2

We start with the following lemma, which is a part of Theorem 1 in Linder [30]:

Lemma 4.4 *There exists a positive constant M, depending on μ, such that*

$$\inf_{\mathbf{c}\in\mathcal{H}^k} W(\mathbf{c},\mu) = \inf_{\mathbf{c}\in\mathcal{F}_M^k} W(\mathbf{c},\mu) \quad \textit{and} \quad \inf_{\mathbf{c}\in(\mathcal{F}_M^k)^c} W(\mathbf{c},\mu) > \inf_{\mathbf{c}\in\mathcal{F}_M^k} W(\mathbf{c},\mu)\ .$$

Let M be the constant of Lemma 4.4. Recall that $\mathbf{c}_n$ is a δ_n-minimizer of the empirical clustering risk over $\mathcal{H}^k$. If $\mathbf{c}_n \in \mathcal{F}_M^k$, we let $\tilde{\mathbf{c}}_n = \mathbf{c}_n$, otherwise we define $\tilde{\mathbf{c}}_n$ as any δ_n-minimizer of the empirical clustering risk over $\mathcal{F}_M^k$. We have

$$\begin{aligned}
&W(\mathbf{c}_n,\mu) - \inf_{\mathbf{c}\in\mathcal{H}^k} W(\mathbf{c},\mu)\\
&\quad = W(\mathbf{c}_n,\mu) - W(\tilde{\mathbf{c}}_n,\mu) + W(\tilde{\mathbf{c}}_n,\mu) - \inf_{\mathbf{c}\in\mathcal{H}^k} W(\mathbf{c},\mu)\\
&\quad = W(\mathbf{c}_n,\mu) - W(\tilde{\mathbf{c}}_n,\mu) + W(\tilde{\mathbf{c}}_n,\mu) - \inf_{\mathbf{c}\in\mathcal{F}_M^k} W(\mathbf{c},\mu),
\end{aligned}$$

where the last equality arises from Lemma 4.4. Denote by $(\Omega,\mathcal{A},\mathbb{P})$ the probability space on which the sequence of random variables $X_1, X_2, \ldots$ is defined, and fix $x > 0$. According to Corollary 2.1, there exists a subset Ω_1 of Ω of probability larger than $1 - e^{-x}$ such that, on Ω_1,

$$W(\tilde{\mathbf{c}}_n,\mu) - \inf_{\mathbf{c}\in\mathcal{F}_M^k} W(\mathbf{c},\mu) \le \frac{12kM^2 + 4M^2\sqrt{2x}}{\sqrt{n}} + \delta_n\ .$$

Define

$$D(M) = \inf W(\mathbf{c},\mu),$$

where the infimum is taken over all $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{H}^k$ such that, for at least one j, $\|c_j\| > M$. Clearly, by Lemma 4.4,

$$D(M) > \inf_{\mathbf{c} \in \mathcal{H}^k} W(\mathbf{c}, \mu).$$

Therefore, using Proposition 2.1, we deduce that a.s., for n large enough,

$$W(\mathbf{c}_n, \mu) < D(M),$$

which in turn implies that, a.s., for n large enough, each component of $\mathbf{c}_n$ is bounded above by M. We have thus proven that for each $\omega \in \Omega$, there exists $N = N(\omega)$ such that, for all $n \geq N$, $W(\mathbf{c}_n, \mu) = W(\tilde{\mathbf{c}}_n, \mu)$.

Note that the rank N may depend on ω. To circumvent this difficulty, for each $N \geq 1$, define the event

$$\Omega_N = \{\omega \in \Omega : W(\mathbf{c}_n, \mu) = W(\tilde{\mathbf{c}}_n, \mu) \quad \text{for all } n \geq N\}.$$

Clearly, $\mathbb{P}(\Omega \setminus \Omega_N) \downarrow 0$ as $N \to \infty$. Choose N_0 such that $\mathbb{P}(\Omega \setminus \Omega_{N_0}) \leq e^{-x}$. Then $\mathbb{P}(\Omega_{N_0}) > 1 - e^{-x}$, and, for all $n \geq N_0$, $W(\mathbf{c}_n, \mu) = W(\tilde{\mathbf{c}}_n, \mu)$ uniformly on Ω_{N_0}. Considering the event $\Omega_1 \cap \Omega_{N_0}$ leads to the desired result.

4.5 Proof of Theorem 3.2

Recall that we denote by $\overline{\mathbf{c}}_n = (\overline{c}_{n1}, \ldots, \overline{c}_{nk})$ the empirical clustering centers associated with the s-dimensional observations $\overline{X}_1, \ldots, \overline{X}_n$. Each $\overline{c}_{nj}$ is the mean of those $\overline{X}_i$'s in the Voronoi cell $\overline{S}_{nj}$, that is,

$$\overline{c}_{nj} = \frac{\sum_{i=1}^n \overline{X}_i \mathbb{I}_{\{\overline{X}_i \in \overline{S}_{nj}\}}}{\sum_{i=1}^n \mathbb{I}_{\{\overline{X}_i \in \overline{S}_{nj}\}}}, \quad j = 1, \ldots, k\ .$$

Define

$$\overline{\alpha}_j = \sum_{i=1}^n \mathbb{I}_{\{\overline{X}_i \in \overline{S}_{nj}\}}\ .$$

Since no confusion is possible, we continue to write μ_n for the empirical measure associated with the projected data $\overline{X}_1, \ldots, \overline{X}_n$. Recalling that each

$\overline{c}_{nj}$ is the mean of the $\overline{X}_i$'s falling in $\overline{S}_{nj}$, we obtain

$$\begin{aligned}
W(\overline{\mathbf{c}}_n, \mu_n) &= \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\overline{X}_i - \overline{c}_{nj}\|^2 \\
&= \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} \|\overline{X}_i - \overline{c}_{nj}\|^2 \mathbb{I}_{\{\overline{X}_i \in \overline{S}_{nj}\}} \\
&= \sum_{j=1}^{k} \frac{1}{2n\overline{\alpha}_j} \sum_{i_1,i_2=1}^{n} \|\overline{X}_{i_1} - \overline{X}_{i_2}\|^2 \mathbb{I}_{\{(\overline{X}_{i_1},\overline{X}_{i_2}) \in \overline{S}^2_{nj}\}}.
\end{aligned}$$

Invoking the optimality of the k-means clustering procedure (see Lemma 1 in Linder [30]), we obtain

$$W(\overline{\mathbf{c}}_n, \mu_n) \leq \sum_{j=1}^{k} \frac{1}{2n\beta_j} \sum_{i_1,i_2=1}^{n} \|\overline{X}_{i_1} - \overline{X}_{i_2}\|^2 \mathbb{I}_{\{(X_{i_1},X_{i_2}) \in S^2_{nj}\}} \ ,$$

where the S_{nj}'s are the Voronoi cells associated with $\mathbf{c}_n = (c_{n1}, \ldots, c_{nk})$, and

$$\beta_j = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in S_{nj}\}}.$$

Consequently, by Theorem 3.1, with probability at least $1-\delta$,

$$\begin{aligned}
W(\overline{\mathbf{c}}_n, \mu_n) &\leq (1+\varepsilon) \sum_{j=1}^{k} \frac{1}{2n\beta_j} \sum_{i_1,i_2=1}^{n} \|X_{i_1} - X_{i_2}\|^2 \mathbb{I}_{\{(X_{i_1},X_{i_2}) \in S^2_{nj}\}} \\
&= (1+\varepsilon) W(\mathbf{c}_n, \mu_n).
\end{aligned}$$

Similarly,

$$(1-\varepsilon) W(\widehat{\mathbf{c}}_n, \mu_n) \leq W(\overline{\mathbf{c}}_n, \mu_n)$$

as desired.

Acknowledgments. The authors thank two referees for valuable comments and suggestions.

References

[1] Abaya, E.A. and Wise, G.L. (1984). Convergence of vector quantizers with applications to optimal quantization, *SIAM Journal on Applied Mathematics*, Vol. 44, pp. 183–189.

[2] Ambroladze, A., Parrado-Hernandez, E. and Shawe-Taylor, J. (2007). Complexity of pattern classes and the Lipschitz property, *Theoretical Computer Science*, Vol. 382, pp. 232–246.

[3] Antos, A. (2005). Improved minimax bounds on the test and training distortion of empirically designed vector quantizers, *IEEE Transactions on Information Theory*, Vol. 51, pp. 4022–4032.

[4] Antos, A., Györfi, L. and György, A. (2005). Improved convergence rates in empirical vector quantizer design, *IEEE Transactions on Information Theory*, Vol. 51, pp. 4013–4022.

[5] Bartlett, P.L. (2003). Prediction algorithms: complexity, concentration and convexity, in *Proceedings of the 13th IFAC Symposium on System Identification*, pp. 1507–1517.

[6] Bartlett, P.L., Boucheron, S. and Lugosi, G. (2001), Model selection and error estimation, *Machine Learning*, Vol. 48, pp. 85–113.

[7] Bartlett, P.L., Linder, T. and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design, *IEEE Transactions on Information Theory*, Vol. 44, pp. 1802–1813.

[8] Bartlett, P.L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results, *Journal of Machine Learning Research*, Vol. 3, pp. 463–482.

[9] Bingham, E. and Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data, in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, pp. 245–250.

[10] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *The Annals of Mathematical Statistics*, Vol. 23, pp. 493–507.

[11] Chou, P.A. (1994). The distortion of vector quantizers trained on n vectors decreases to the optimum at $O_P(1/n)$, *IEEE Transactions on Information Theory*, Vol. 8, pp. 457–457.

[12] Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss, *Random Structures and Algorithms*, Vol. 22, pp. 60–65.

[13] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York.

[14] Duda, R.O., Hart, P.E. and Stork, D.G. (2000). *Pattern Classification*, Second Edition, Wiley-Interscience, New York.

[15] Dudley, R.M. (2002). *Real Analysis and Probability*, Second Edition, Cambridge University Press, Cambridge.

[16] Frankl, P. and Maehara, H. (1988). The Johnson-Lindenstrauss lemma and the sphericity of some graphs, *Journal of Combinatorial Theory, Series B*, Vol. 44, pp. 355–362.

[17] Frankl, P. and Maehara, H. (1990). Some geometric applications of the beta distribution, *Annals of the Institute of Statistical Mathematics*, Vol. 42, pp. 463–474.

[18] Gersho, A. and Gray, R.M. (1992). *Vector Quantization and Signal Compression*, Kluwer Academic Press, Boston.

[19] Graf, S. and Luschgy, H. (2000). *Foundations of Quantization for Probability Distributions*, Springer, Lecture Notes in Mathematics, 1730, Berlin.

[20] Gray, R.M. and Neuhoff, D.L. (1998). Quantization, *IEEE Transactions on Information Theory*, Vol. 44, pp. 2325–2384.

[21] Hartigan, J.A. (1975). *Clustering Algorithms*, Wiley, New York.

[22] Hartigan, J.A. (1978). Asymptotic distributions for clustering criteria, *The Annals of Statistics*, Vol. 6, pp. 117–131.

[23] Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality, *Proceedings of the 30th Symposium on Theory of Computing*, pp. 604–613.

[24] Johnson, W. and Lindenstrauss, J. (1984). Extensions of Lipschitz maps into a Hilbert space, *Contemporary Mathematics*, Vol. 26, pp. 189–206.

[25] Kleinberg, J.M. (1997). Two algorithms for nearest-neighbor search in high dimensions, in *29th Annual ACM Symposium on Theory of Computing*, pp. 599–608.

[26] Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization, *IEEE Transactions on Information Theory*, Vol. 47, pp. 1902–1914.

[27] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization, *The Annals of Statistics*, Vol. 34, pp. 2593–2656.

[28] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*, Springer-Verlag, Berlin.

[29] Linder, T. (2000). On the training distortion of vector quantizers, *IEEE Transactions on Information Theory*, Vol. 46, pp. 1617–1623.

[30] Linder, T. (2001). *Learning-Theoretic Methods in Vector Quantization*, Lecture Notes for the Advanced School on the Principles of Nonparametric Learning, Udine, Italy, July 9-13.

[31] Linder, T., Lugosi, G. and Zeger, K. (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding, *IEEE Transactions on Information Theory*, Vol. 40, pp. 1728–1740.

[32] Linial, N., London, E. and Rabinovich, Y. (1995). The geometry of graphs and some of its algorithmic applications, *Combinatorica*, Vol. 15, pp. 215–245.

[33] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*, Academic Press, London.

[34] McDiarmid, C. (1989). On the method of bounded differences, in *Surveys in Combinatorics 1989*, pp. 148–188, Cambridge University Press, Cambridge.

[35] Pollard, D. (1981). Strong consistency of k-means clustering, *The Annals of Statistics*, Vol. 9, pp. 135–140.

[36] Pollard, D. (1982). A central limit theorem for k-means clustering, *The Annals of Probability*, Vol. 10, pp. 919–926.

[37] Pollard, D. (1982). Quantization and the method of k-means, *IEEE Transactions on Information Theory*, Vol. 28, pp. 199–205.

[38] Rachev, S.T. and Rüschendorf, L. (1998). *Mass Transportation Problems, Volume I: Theory*, Springer-Verlag, New York.

[39] Rachev, S.T. and Rüschendorf, L. (1998). *Mass Transportation Problems, Volume II: Applications*, Springer-Verlag, New York.

[40] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*, Springer, New York.

[41] Schölkopf, B. and Smola, A.J. (2002). *Learning with Kernels*, The MIT Press, Cambridge.

[42] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge.

[43] Vapnik, V.N. (1998). *Statistical Learning Theory*, John Wiley & Sons, New York.

[44] Vempala, S. and Wang, G. (2002). A spectral algorithm for learning mixtures of distributions, in *Proceedings of the 43rd IEEE Foundations of Computer Science (FOCS '02).*

[45] Williams, D. (1991). *Probability with Martingales*, Cambridge University Press, Cambridge.