Uniform error bounds for K-dimensional coding schemes in Hilbert spaces

Andreas Maurer[1] and Massimiliano Pontil[2]

[1] Adalbertstrasse 55
D-80799 München, Germany
andreasmaurer@compuserve.com
[2] Dept. of Computer Science
University College London
Malet Pl., WC1E, London, UK
m.pontil@cs.ucl.ac.uk

Abstract. We give a bound on the expected reconstruction error for a general coding method where data in a Hilbert space are represented by finite dimensional coding vectors. The result can be specialized to K-means clustering, nonnegative matrix factorization and the sparse coding techniques introduced by Olshausen and Field.

1 Introduction

We consider the generalization performance of a general class of K-dimensional coding schemes for data x drawn from a distribution μ on the unit ball of a Hilbert space H. These schemes have the form

$$\begin{aligned}\hat{x} &= T\hat{y}\\ \hat{y} &= \arg\min_{y\in A}\left(\|x-Ty\|^2+g\left(y\right)\right),\end{aligned}$$

where $A\subseteq\mathbb{R}^K$ is some set of *codes* (which we can always assume to span $\mathbb{R}^K$) and $g:\mathbb{R}^K\to\mathbb{R}_+$ is some *regularizing function* used to encourage or discourage the use of certain codes, but g may also be chosen zero. The pair (A,g) defines the particular *coding scheme*.

$T:\mathbb{R}^K\to H$ is a linear map, which defines a particular *implementation* of the coding scheme. It embeds the set A of codes in H and yields the set $T\left(A\right)$ of exactly codable patterns. The quantity

$$f_T\left(x\right)=\min_{y\in A}\left(\|x-Ty\|^2+g\left(y\right)\right)$$

is the (regularized) reconstruction error.

Given a coding scheme (A,g) and a finite number of independent observations $x_1,...,x_m\in H$, a common sense approach searches for an implementation T_{opt} which is optimal on average over the observed points, that is

$$T_{\text{opt}}=\arg\min_{T\in\mathcal{C}}\frac{1}{m}\sum_{i=1}^{m}f_T\left(x_i\right), \tag{1}$$

where $\mathcal{C}$ denotes some class of linear embeddings $T : \mathbb{R}^K \to H$. As we shall see, this framework is general enough to include principal component analysis, K-means clustering, non-negative matrix factorization [9] and the sparse coding schemes as proposed in [12].

To give a justification of this approach (which can be regarded as empirical risk minimization) we require that the class of sets $\{T(A) : T \in \mathcal{C}\}$ is uniformly bounded, or, equivalently, that the quantity

$$\|\mathcal{C}\|_A = \sup_{T\in\mathcal{C}} \|T\|_A = \sup_{T\in\mathcal{C}} \sup_{y\in A} \|Ty\|$$

is finite. We then have the following high probability bound on the expected reconstruction error, uniformly valid for all $T \in \mathcal{C}$.

Theorem 1. *Assume that $K > 1$, $\|\mathcal{C}\|_A \geq 1$, that the functions f_T for $T \in \mathcal{C}$, when restricted to the unit ball of H, have range contained in $[0, b]$, and that the measure μ is supported on the unit ball of H. Fix $\delta > 0$.*

Then with probability at least $1 - \delta$ in the observed data $\mathbf{x} \sim \mu^m$ we have for every $T \in \mathcal{C}$ that

$$\mathbb{E}_{x\sim\mu} f_T(x) - \frac{1}{m}\sum_{i=1}^{m} f_T(x_i) \leq \frac{K}{\sqrt{m}}\left(20\|\mathcal{C}\|_A + \frac{b}{2}\sqrt{\ln\left(16m\|\mathcal{C}\|_A^2\right)}\right) + b\sqrt{\frac{\ln 1/\delta}{2m}}.$$

It may be argued that finite sample bounds are not as relevant for unsupervised learning as for supervised learning, because of the lower cost of unlabeled data. Convergence of the empirical to the expected objective seems nevertheless a minimal requirement for the justification of any unsupervised learning method. If $\|\mathcal{C}\|_A < \infty$ and $b < \infty$ our result immediately implies convergence in probability, uniform in all possible implementations of the respective coding scheme. We are not aware of other comparable results for nonnegative matrix factorization [9] or the sparse coding techniques as in [12].

Before providing a proof of Theorem 1 we illustrate its implications in some specific cases of interest.

2 Examples of coding schemes

Several coding schemes can be expressed in our framework. We briefly describe these methods and how our result applies.

2.1 Principal component analysis

This classical method (PCA) seeks the K-dimensional orthogonal projection which maximizes the projected variance and then uses this projection to encode future data. Let T_P be an isometry which maps $\mathbb{R}^K$ to the range of a projection P. Since

$$\|Px\|^2 = \|x\|^2 - \min_{y\in\mathbb{R}^K} \|x - T_P y\|^2,$$

finding P to maximize the true or empirical expectation of $\|Px\|^2$ is equivalent to finding T to minimize the corresponding expectation of $\min_{y\in\mathbb{R}^K}\|x-Ty\|^2$. If we use the projection P to encode a given $x\in H$ then $Px=T_P\hat{y}$ where $\hat{y}\in\mathbb{R}^K$ is the minimizer $\|x-T_Py\|^2$. We see that PCA is described by our framework upon the identifications $A=\mathbb{R}^K$, $g\equiv 0$ where $\mathcal{C}$ is restricted to the class of isometries $T:\mathbb{R}^K\to H$. Given $T\in\mathcal{C}$ and $x\in H$ the reconstruction error is

$$f_T(x)=\min_{y\in\mathbb{R}^K}\|x-Ty\|^2.$$

If the data are constrained to be in the unit ball of H, as we generally assume, then it is easily seen that we can take A to be the unit ball of $\mathbb{R}^K$ without changing any of the encodings. We can therefore apply our result with $\|\mathcal{C}\|_A=1$ and $b=1$. This is besides the point however, because in the simple case of PCA much better bounds are available ([13], [17]). In fact we will prove a bound of order $\sqrt{K/m}$ in the course of the proof of Theorem 1 (see Lemma 4 below). In [17] local Rademacher averages are used to give faster rates under certain circumstances.

An objection to PCA is, that generic codes have K nonzero components, while for practical and theoretical reasons sparse codes with much less than K nonzero components are preferable.

2.2 K-means clustering or vector quantization

Here $A=\{e_1,...,e_K\}$, where the e_k form an orthonormal basis of $\mathbb{R}^K$ and $g\equiv 0$. An implementation T now defines a set of centers $\{Te_1,...,Te_K\}$, the reconstruction error is $\min_{k=1}^K\|x-Te_k\|^2$ and a data point x is coded by the e_k such that Te_k is nearest to x. The algorithm (1) becomes

$$T_{\text{opt}}=\arg\min_{T\in\mathcal{C}}\frac{1}{m}\sum_{i=1}^m\min_{k=1}^K\|x-Te_k\|^2.$$

It is reasonable to assume that every center Te_k has at most unit norm, so that $\|\mathcal{C}\|_A=1$. Since all data points are in the unit ball we have $\|x-Te_k\|^2\le 4$ so we can set $b=4$ and the bound on the estimation error becomes

$$\left(20+2\sqrt{\ln(16m)}\right)\frac{K}{\sqrt{m}}+\sqrt{\frac{8\ln(1/\delta)}{m}}.$$

The order of this bound matches up to $\sqrt{\ln m}$ the order given in [3] or [14]. To illustrate our method we will also prove the bound

$$\sqrt{18\pi}\frac{K}{\sqrt{m}}+\sqrt{\frac{8\ln(1/\delta)}{m}}$$

(Theorem 5), which is slightly better than those in [3] or [14]. There is a lower bound of order $\sqrt{K/m}$ in [2], and it is unknown which of the two bounds (upper or lower) is tight.

In K-means clustering every code has only one nonzero component, so that sparsity is enforced in a maximal way. On the other hand this results in a weaker approximation capability of the coding scheme.

2.3 Nonnegative matrix factorization

Here A is the cone $A = \left\{\sum_{k=1}^{K} \lambda_k e_k : \lambda_i \geq 0\right\}$ and $g \equiv 0$. A chosen embedding T generates a cone $T(A) \subset H$ onto which incoming data is projected. In the original formulation by Lee and Seung [9] it is postulated that both the data and the vectors Te_k be contained in the positive orthant of some finite dimensional space, but we can drop most of these restrictions, keeping only the requirement that $\langle Te_k, Te_l \rangle \geq 0$ for $1 \leq k, l \leq K$.

No coding will change if we require that $\|Te_k\| = 1$ for all $1 \leq k \leq K$ by a suitable normalization and we can restrict A to its intersection with the unit ball in $\mathbb{R}^K$ (see Lemma 2 below) and set $\|\mathcal{C}\|_A = \sqrt{K}$. From Theorem 1 we obtain the bound

$$\frac{K}{\sqrt{m}}\left(20\sqrt{K} + \frac{1}{2}\sqrt{\ln(16mK)}\right) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

on the estimation error. We do not know of any other generalization bounds for this coding scheme.

Nonnegative matrix factorization appears to encourage sparsity, but cases have been reported where sparsity was not observed [10]. In fact this undesirable behaviour should be generic for exactly codable data. Various authors have therefore proposed additional constraints ([10], [6]). It is clear that additional constraints on $\mathcal{C}$ can only improve generalization and that the passage from A to a subset can only improve our bounds.

2.4 Sparse coding of Olshausen and Field

In the original formulation [12] $A = \mathbb{R}^K$ but g is one of the functions $g(y) = -\lambda \sum_i e^{-y_i^2}$, $g(y) = \lambda \sum_i \ln\left(1 + y_i^2\right)$ or $g(y) = \lambda \sum_i |y_i|$ and $\lambda > 0$ is a regularization parameter which controls how strongly sparsity is to be encouraged. To see how our result applies, we focus on the last and most conventional regularizer $g(y) = \lambda \|y\|_1$. If $\hat{y}$ is a minimizer for $\|x - Ty\|^2 + \lambda \|y\|_1$ with $\|x\| \leq 1$ then

$$\begin{aligned}\lambda \|\hat{y}\| \leq \lambda \|\hat{y}\|_1 &\leq \|x - T\hat{y}\|^2 + \lambda \|\hat{y}\|_1 \\ &\leq \|x - T0\|^2 + \lambda \|0\|_1 = \|x\|^2 \leq 1,\end{aligned}$$

so $\|\hat{y}\| \leq \lambda^{-1}$, which shows that we can equivalently set A to be the ball of radius λ^{-1} in the definition of this coding scheme. If we have a uniform bound $\|\mathcal{C}\|_\infty$ on the spectral norms of the operators $T \in \mathcal{C}$ we get $\|\mathcal{C}\|_A \leq \lambda^{-1} \|\mathcal{C}\|_\infty$. By the same argument as above all f_T have range contained in $[0, 1]$, so the Theorem can be applied with $b = 1$ to yield the bound

$$\frac{K}{\sqrt{m}}\left(\frac{20\|\mathcal{C}\|_\infty}{\lambda} + \frac{1}{2}\sqrt{\ln\left(16m\lambda^{-2}\|\mathcal{C}\|_\infty^2\right)}\right) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

on the estimation error. It is interesting to observe that increasing the regularization parameter λ, both increases sparsity and improves estimation. With similar but more complicated methods the Theorem can also be applied to the other regularizers.

The method of Olshausen and Field [12] approximates with a compromise of geometric proximity and sparsity and our result asserts that the observed value of this compromise generalizes to unseen data if enough data have been observed.

3 Proofs

We first introduce some notation, conventions and auxiliary results. Then we set about to prove our main result.

3.1 Notation, definitions and auxiliary results

Throughout H denotes a Hilbert space. The term *norm* and the notation $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ always refer to the euclidean norm and inner product on $\mathbb{R}^K$ or on H. Other norms are characterized by subscripts. If H_1 and H_2 are any Hilbert spaces $\mathcal{L}(H_1, H_2)$ denotes the vector space of bounded linear transformations from H_1 to H_2. If $H_1 = H_2$ we just write $\mathcal{L}(H_1) = \mathcal{L}(H_1, H_1)$. With $\mathcal{U}(H_1, H_2)$ we denote the set of isometries in $\mathcal{L}(H_1, H_2)$, that is maps U satisfying $\|Ux\| = \|x\|$ for all $x \in H_1$.

We use $\mathcal{L}_2(H)$ for the set of Hilbert-Schmidt operators on H, which becomes itself a Hilbert space with the inner product $\langle T, S\rangle_2 = \mathrm{tr}(T^*S)$ and the corresponding (Frobenius-) norm $\|\cdot\|_2$.

For $x \in H$ the operator Q_x is defined by $Q_x z = \langle z, x\rangle$. For any $T \in \mathcal{L}_2(H)$ the identity

$$\langle T^*T, Q_x\rangle_2 = \|Tx\|^2 \tag{2}$$

is easily verified.

Suppose that $A \subseteq \mathbb{R}^K$ spans $\mathbb{R}^K$, that H' is any Hilbert space (which could also be $\mathbb{R}^K$). It is easily verified that the quantity

$$\|T\|_A = \sup_{y \in A} \|Ty\|$$

defines a norm on $\mathcal{L}\left(\mathbb{R}^K, H'\right)$.

We use the following well known result on covering numbers (e.g. Proposition 5 in [4]).

Proposition 1. *Let B be a ball of radius r in an N-dimensional Banach space and $\epsilon > 0$. There exists a subset $B_\epsilon \subset B$ such that $|B_\epsilon| \leq (4r/\epsilon)^N$ and $\forall z \in B, \exists z' \in B_\epsilon$ with $d(z, z') \leq \epsilon$, where d is the metric of the Banach space.*

The following concentration inequality, known as the bounded difference inequality [11], goes back to the work of Hoeffding [5].

Theorem 2. *Let μ_i be a probability measure on a space Ω_i, for $i = 1, ..., m$. Let $\Omega = \prod_{i=1}^{m} \Omega_i$ and $\mu = \otimes_{i=1}^{m} \mu_i$ be the product space and product measure respectively. Suppose the function $\Psi : \Omega \to \mathbb{R}$ satisfies*

$$|\Psi(\mathbf{x}) - \Psi(\mathbf{x}')| \leq c_i$$

whenever $\mathbf{x}$ and $\mathbf{x}'$ differ only in the i-th coordinate. Then

$$\Pr_{\mathbf{x}\sim\mu} \{\Psi(\mathbf{x}) - \mathbb{E}_{\mathbf{x}'\sim\mu}\Psi(\mathbf{x}') \geq t\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{m} c_i^2}\right).$$

Throughout σ_i will denote a sequence of mutually independent random variables, uniformly distributed on $\{-1, 1\}$ and γ_i, γ_{ij} will be (multipled indexed) sequences of mutually independent Gaussian random variables, with zero mean and unit standard deviation.

If $\mathcal{F}$ is a class of real functions on a space $\mathcal{X}$ and μ a probability measure on $\mathcal{X}$ then for $m \in \mathbb{N}$ the Rademacher and Gaussian complexities of $\mathcal{F}$ w.r.t. μ are defined ([8],[1]) as

$$\mathcal{R}_m(\mathcal{F}, \mu) = \frac{2}{m}\mathbb{E}_{\mathbf{x}\sim\mu^m}\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i f(x_i),$$

$$\Gamma_m(\mathcal{F}, \mu) = \frac{2}{m}\mathbb{E}_{\mathbf{x}\sim\mu^m}\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \gamma_i f(x_i)$$

repectively.

Appropriately scaled Gaussian complexities can be substituted for Rademacher complexities, by virtue of the next Lemma. For a proof see, for example, [8, p. 97].

Lemma 1. *For $A \subseteq \mathbb{R}^k$ we have $\mathcal{R}(A) \leq \sqrt{\pi/2}\ \Gamma(A)$.*

The next result is known as Slepian's lemma ([15], [8]).

Theorem 3. *Let Ω and Ξ be mean zero, separable Gaussian processes indexed by a common set $\mathcal{S}$, such that*

$$\mathbb{E}(\Omega_{s_1} - \Omega_{s_2})^2 \leq \mathbb{E}(\Xi_{s_1} - \Xi_{s_2})^2 \text{ for all } s_1, s_2 \in \mathcal{S}.$$

Then

$$\mathbb{E}\sup_{s\in\mathcal{S}} \Omega_s \leq \mathbb{E}\sup_{s\in\mathcal{S}} \Xi_s.$$

The following result, which generalizes Theorem 8 in [1], plays a central role in our proof.

Theorem 4. *Let $\{\mathcal{F}_n : 1 \leq n \leq N\}$ be a finite collection of $[0, b]$-valued function classes on a space $\mathcal{X}$, and μ a probability measure on $\mathcal{X}$. Then $\forall \delta \in (0, 1)$ we have with probability at least $1 - \delta$ that*

$$\max_{n\leq N} \sup_{f\in\mathcal{F}_n} \left[\mathbb{E}_{x\sim\mu} f(x) - \frac{1}{m}\sum_{i=1}^{m} f(x_i)\right] \leq \max_{n\leq N} \mathcal{R}_m(\mathcal{F}_n, \mu) + b\sqrt{\frac{\ln N + \ln(1/\delta)}{2m}}.$$

Proof. Denote with Ψ_n the function on $\mathcal{X}^m$ defined by

$$\Psi_n(\mathbf{x}) = \sup_{f \in \mathcal{F}_n} \left[\mathbb{E}_{x \sim \mu} f(x) - \frac{1}{m} \sum_{i=1}^{m} f(x_i) \right], \ \mathbf{x} \in \mathcal{X}^m.$$

By standard symmetrization (see [16]) we have $\mathbb{E}_{\mathbf{x} \sim \mu^m} \Psi_n(\mathbf{x}) \leq \mathcal{R}_m(\mathcal{F}_n, \mu) \leq \max_{n \leq N} \mathcal{R}_m(\mathcal{F}_n, \mu)$. Modifying one of the x_i can change the value of any $\Psi_n(\mathbf{x})$ by at most b/m, so that by a union bound and the bounded difference inequality (Theorem 2)

$$\Pr\left\{ \max_{n \leq N} \Psi_n > \max_{n \leq N} \mathcal{R}_m(\mathcal{F}_n, \mu) + t \right\} \leq \sum_n \Pr\{\Psi_n > \mathbb{E}\Psi_n + t\} \leq N e^{-2m(t/b)^2}.$$

Solving $\delta = N e^{-2m(t/b)^2}$ for t gives the result. □

The following lemma was used in Section 2.3.

Lemma 2. *Suppose* $\|x\| \leq 1$, $\|c_k\| = 1$, $\langle c_k, c_l \rangle \geq 0$, $y \in \mathbb{R}^K$, $y_i \geq 0$. *If* y *minimizes*

$$h(y) = \left\| x - \sum_{k=1}^{K} y_k c_k \right\|^2,$$

then $\|y\| \leq 1$.

Proof. Assume that y is a minimzer of h and $\|y\| > 1$.Then

$$\left\| \sum_{k=1}^{K} y_k c_k \right\|^2 = \|y\|^2 + \sum_{k \neq l} y_k y_l \langle c_k, c_l \rangle > 1.$$

Let the real function f be defined by $f(t) = h(ty)$. Then

$$\begin{aligned} f'(1) &= 2 \left(\left\| \sum_{k=1}^{K} y_k c_k \right\|^2 - \left\langle x, \sum_{k=1}^{K} y_k c_k \right\rangle \right) \\ &\geq 2 \left(\left\| \sum_{k=1}^{K} y_k c_k \right\|^2 - \left\| \sum_{k=1}^{K} y_k c_k \right\| \right) \\ &= 2 \left(\left\| \sum_{k=1}^{K} y_k c_k \right\| - 1 \right) \left\| \sum_{k=1}^{K} y_k c_k \right\| \\ &> 0. \end{aligned}$$

So f cannot have a minimum at 1, whence y cannot be a minimizer of h. □

3.2 Proof of the main results

We now fix a spanning set $A \subseteq \mathbb{R}^K$ and a "regularizer" $g : A \to \mathbb{R}_+$. Recall that, for $T \in \mathcal{L}\left(\mathbb{R}^K, H\right)$, we had introduced the notation

$$f_T(x) = \inf_{y \in A} \left(\|x - Ty\|^2 + g(y) \right), x \in H.$$

Our principal object of study is the function class

$$\mathcal{F} = \left\{ x \mapsto \inf_{y \in A} \left(\|x - Ty\|^2 + g(y) \right) : T \in \mathcal{C} \right\} = \left\{ f_T : T \in \mathcal{C} \right\},$$

restricted to the unit ball in H, when $\mathcal{C} \subset \mathcal{L}\left(\mathbb{R}^K, H\right)$ is some fixed set of candidate implementations of our coding scheme.

To illustrate our method we first consider the somewhat simpler special case of K-means clustering, corresponding to the choices $A = \{e_1, ..., e_K\}$, $g \equiv 0$ and $\mathcal{C} = \{T : \|T\|_A \leq 1\}$, equivalent to the requirement that $\|Te_k\| \leq 1$ for all $T \in \mathcal{C}$ and all $k \in \{1, ..., K\}$. As already noted in Section 2.2 the vectors Te_k define the cluster centers.

Theorem 5. *For every $\delta > 0$ with probability greater $1-\delta$ in the sample $\mathbf{x} \sim \mu^m$ we have for all $T \in \mathcal{C}$*

$$\mathbb{E}_{x \sim \mu} \min_{k=1}^{K} \|x - Te_k\|^2 \leq \frac{1}{m} \sum_{i=1}^{m} \min_{k=1}^{K} \|x_i - Te_k\|^2 + K\sqrt{\frac{18\pi}{m}} + \sqrt{\frac{8 \ln(1/\delta)}{m}}.$$

Proof. According to [1] we need to bound the Rademacher complexity of the function class $\mathcal{F}$. By Lemma 1 it suffices to bound the corresponding Gaussian complexity, which we shall do using Slepian's Lemma (Theorem 3). We have

$$\mathcal{R}(\mathcal{F}, \mu) \leq \sqrt{\frac{\pi}{2}} \Gamma(\mathcal{F}, \mu) = \sqrt{\frac{\pi}{2}} \frac{2}{m} \mathbb{E}_{\mathbf{x} \sim \mu^m} \mathbb{E}_\gamma \sup_{T \in \mathcal{C}} \sum_{i=1}^{m} \gamma_i \min_{k=1}^{K} \|x_i - Te_k\|^2. \quad (3)$$

Now we fix a sample $\mathbf{x}$ and define Gaussian processes Ω and Ξ indexed by $\mathcal{C}$

$$\Omega_T = \sum_{i=1}^{m} \gamma_i \min_{k=1}^{K} \|x_i - Te_k\|^2 \text{ and } \Xi_T = \sum_{i=1}^{m} \sum_{k=1}^{K} \gamma_{ik} \|x_i - Te_k\|^2.$$

Using orthonormality of the γ_i and γ_{ik} we obtain for $T_1, T_2 \in \mathcal{C}$

$$\begin{aligned}
\mathbb{E}(\Omega_{T_1} - \Omega_{T_2})^2 &= \sum_{i=1}^{m} \left(\min_k \|x_i - T_1 e_k\|^2 - \min_k \|x_i - T_2 e_k\|^2 \right)^2 \\
&\leq \sum_{i=1}^{m} \max_k \left(\|x_i - T_1 e_k\|^2 - \|x_i - T_2 e_k\|^2 \right)^2 \\
&\leq \sum_{i=1}^{m} \sum_{k=1}^{K} \left(\|x_i - T_1 e_k\|^2 - \|x_i - T_2 e_k\|^2 \right)^2 \qquad (*) \\
&= \mathbb{E}(\Xi_{T_1} - \Xi_{T_2})^2.
\end{aligned}$$

By Slepian's Lemma, the triangle inequality, Schwarz' and Jensen's inequalities

$$\begin{aligned}
&\mathbb{E}_\gamma \sup_{T\in\mathcal{C}} \sum_{i=1}^m \gamma_i \min_{k=1}^K \|x_i - Te_k\|^2 \\
&= \mathbb{E}_\gamma \sup_{T\in\mathcal{C}} \Omega_T \\
&\leq \mathbb{E}_\gamma \sup_{T\in\mathcal{C}} \Xi_T \text{ (Slepian)} \\
&= \mathbb{E}_\gamma \sup_{T\in\mathcal{C}} \sum_{i=1}^m \sum_{k=1}^K \gamma_{ik} \|x_i - Te_k\|^2 \\
&\leq 2K\mathbb{E}_\gamma \left\| \sum_{i=1}^m \gamma_i x_i \right\| + K\mathbb{E}_\gamma \left| \sum_{i=1}^m \gamma_i \right| \text{ (triangle and Schwarz)} \\
&\leq 3K\sqrt{m} \text{ (Jensen).}
\end{aligned}$$

Substitution in (3) yields $\mathcal{R}(\mathcal{F},\mu) \leq K\sqrt{18\pi/m}$, which, using Theorem 4 with $N = 1$ and $b = 4$ implies the result. □

It is tempting to use the same technique in the general case. Unfortunately an essential step in the application of Slepian's Lemma, marked (*) above, is impossible if A is infinite, so that a more devious path has to be chosen.

The idea is the following: Every implementing map $T \in \mathcal{C}$ can be factored as $T = U \circ S$, where S is a $K \times K$ matrix, $S \in \mathcal{L}\left(\mathbb{R}^K\right)$, and U is an isometry, $U \in \mathcal{U}(\mathbb{R}^K, H)$. Suitably bounded $K \times K$ matrices form a compact, finite dimensional set, the complexity of which can be controlled using covering numbers, while the complexity arising from the set of isometries can be controlled with Rademacher and Gaussian averages. Theorem 4 then combines these complexity estimates.

For fixed $S \in \mathcal{L}\left(\mathbb{R}^K\right)$ we denote

$$\mathcal{G}_S = \left\{ f_{US} : U \in \mathcal{U}\left(\mathbb{R}^K, H\right) \right\}.$$

Recall the notation $\|\mathcal{C}\|_A = \sup_{T\in\mathcal{C}} \|T\|_A = \sup_{T\in\mathcal{C}} \sup_{y\in A} \|Ty\|$. With $\mathcal{S}$ we denote the set of $K \times K$-matrices

$$\mathcal{S} = \left\{ S \in \mathcal{L}\left(\mathbb{R}^K\right) : \|S\|_A \leq \|\mathcal{C}\|_A \right\}.$$

Lemma 3. *Assume $\|\mathcal{C}\|_A \geq 1$, that the functions in $\mathcal{F}$, when restricted to the unit ball of H, have range contained in $[0,b]$, and that the measure μ is supported on the unit ball of H. Then with probability at least $1-\delta$ for all $T \in \mathcal{C}$*

$$\begin{aligned}
&\mathbb{E}_{x\sim\mu} f_T(x) - \frac{1}{m}\sum_{i=1}^m f_T(x_i) \\
&\qquad \leq \sup_{S\in\mathcal{S}} \mathcal{R}_m(\mathcal{G}_S,\mu) + \frac{bK}{2}\sqrt{\frac{\ln\left(16m\|\mathcal{C}\|_A^2\right)}{m}} + \frac{8\|\mathcal{C}\|_A}{\sqrt{m}} + b\sqrt{\frac{\ln(1/\delta)}{2m}}.
\end{aligned}$$

Proof. Fix $\epsilon > 0$. The set $\mathcal{S}$ is the ball of radius $\|\mathcal{C}\|_A$ in the K^2-dimensional Banach space $\left(\mathcal{L}\left(\mathbb{R}^K\right), \|.\|_A\right)$ so by Proposition 1 we can find a subset $\mathcal{S}_\epsilon \subset \mathcal{S}$, of cardinality $|\mathcal{S}_\epsilon| \leq \left(4\|\mathcal{C}\|_A/\epsilon\right)^{K^2}$ such that every member of $\mathcal{S}$ can be approximated by a member of $\mathcal{S}_\epsilon$ up to distance ϵ in the norm $\|.\|_A$.

We claim that for all $T \in \mathcal{C}$ there exist $U \in \mathcal{U}(\mathbb{R}^K, H)$ and $S_\epsilon \in \mathcal{S}_\epsilon$ such that

$$|f_T(x) - f_{US_\epsilon}(x)| < 4\|\mathcal{C}\|_A \epsilon,$$

for all x in the unit ball of H. To see this write $T = US$ with $U \in \mathcal{U}(\mathbb{R}^K, H)$ and $S \in \mathcal{L}(\mathbb{R}^K)$. Then, since U is an isometry, we have

$$\|S\|_A = \sup_{y \in A} \|Sy\| = \sup_{y \in A} \|Ty\| = \|T\|_A \leq \|\mathcal{C}\|_A$$

so that $S \in \mathcal{S}$. We can therefore choose $S_\epsilon \in \mathcal{S}_\epsilon$ such that $\|S_\epsilon - S\|_A < \epsilon$. Then for $x \in H$, with $\|x\| \leq 1$, we have

$$\begin{aligned}
|f_T(x) - f_{US_\epsilon}(x)| &= \inf_{y \in A}\left(\|x - USy\|^2 + g(y)\right) - \inf_{y \in A}\left(\|x - US_\epsilon y\|^2 + g(y)\right) \\
&\leq \sup_{y \in A}\left(\|x - USy\|^2 - \|x - US_\epsilon y\|^2\right) \\
&= \sup_{y \in A} \langle US_\epsilon y - USy, 2x - (USy + US_\epsilon y)\rangle \\
&\leq (2 + 2\|\mathcal{C}\|_A) \sup_{y \in A} \|(S_\epsilon - S)y\| \leq 4\|\mathcal{C}\|_A \epsilon.
\end{aligned}$$

Apply Theorem 4 to the finite collection of function classes $\{\mathcal{G}_S : S \in \mathcal{S}_\epsilon\}$ to see that with probability at least $1 - \delta$

$$\begin{aligned}
&\sup_{T \in \mathcal{C}} \mathbb{E}_{x \sim \mu} f_T(x) - \frac{1}{m}\sum_{i=1}^m f_T(x_i) \\
&\leq \max_{S \in \mathcal{S}_\epsilon} \sup_{U \in \mathcal{U}(\mathbb{R}^K, H)} \mathbb{E}_{x \sim \mu} f_{US}(x) - \frac{1}{m}\sum_{i=1}^m f_{US}(x_i) + 8\|\mathcal{C}\|_A \epsilon \\
&\leq \max_{S \in \mathcal{S}_\epsilon} \mathcal{R}_m(\mathcal{G}_S, \mu) + b\sqrt{\frac{\ln|\mathcal{S}_\epsilon| + \ln(1/\delta)}{2m}} + 8\|\mathcal{C}\|_A \epsilon \\
&\leq \sup_{S \in \mathcal{S}} \mathcal{R}_m(\mathcal{G}_S, \mu) + \frac{bK}{2}\sqrt{\frac{\ln\left(16m\|\mathcal{C}\|_A^2\right)}{m}} + \frac{8\|\mathcal{C}\|_A}{\sqrt{m}} + b\sqrt{\frac{\ln(1/\delta)}{2m}},
\end{aligned}$$

where the last line follows from the known bound on $|\mathcal{S}_\epsilon|$, subadditivity of the square root and the choice $\epsilon = 1/\sqrt{m}$. □

To complete the proof of Theorem 1 we now fix some $S \in \mathcal{S}$ and focus on the corresponding function class $\mathcal{G}_S$. Observe that for an isometry $U \in \mathcal{U}(\mathbb{R}^K, H)$ the

operator U^*U is the identity on $\mathbb{R}^K$ and that UU^* is the orthogonal projection onto the range of U. We therefore have, for $x \in H$,

$$\begin{aligned}\inf_{y\in A}\|x-USy\|^2 &= \|x-UU^*x\|^2 + \inf_{y\in A}\|UU^*x-USy\|^2 \\ &= \|x\|^2 - \|UU^*x\|^2 + \inf_{y\in A}\|U^*x-Sy\|^2\end{aligned}$$

so that $\mathcal{G}_S = \mathcal{D} + \mathcal{E}_S$, where

$$\begin{aligned}\mathcal{D} &= \left\{x \mapsto \|x\|^2 - \|UU^*x\|^2 : U \in \mathcal{U}\left(\mathbb{R}^K, H\right)\right\} \\ \mathcal{E}_S &= \left\{x \mapsto \inf_{y\in A}\|U^*x-Sy\|^2 + g(y) : U \in \mathcal{U}\left(\mathbb{R}^K, H\right)\right\}.\end{aligned}$$

We will bound the Rademacher complexities of these two function classes in turn.

Observe that the function class $\mathcal{D}$ is the class of reconstruction errors of PCA, so the next lemma and an application of Theorem 4 with $N = 1$ and $b = 1$ also give a generalization bound for PCA of order $\sqrt{K/m}$.

Lemma 4. $\mathcal{R}(\mathcal{D}, \mu) \leq 2\sqrt{K/m}$.

Proof. For $z \in H$ define the outer product operator Q_z by $Q_z x = \langle x, z\rangle z$. With $\langle ., .\rangle_2$ and $\|.\|_2$ denoting the Hilbert-Schmidt inner product and norm respectively we have for $\|x_i\| \leq 1$

$$\begin{aligned}\mathbb{E}_\sigma \sup_{f\in\mathcal{D}}\sum_{i=1}^m \sigma_i f(x_i) &= \mathbb{E}_\sigma \sup_{U\in\mathcal{U}}\sum_{i=1}^m \sigma_i\left(\|x_i\|^2 - \|UU^*x_i\|^2\right) \\ &= \mathbb{E}_\sigma \sup_{U\in\mathcal{U}}\left\langle\sum_{i=1}^m \sigma_i Q_{x_i}, UU^*\right\rangle_2 \\ &\leq \mathbb{E}_\sigma\left\|\sum_{i=1}^m \sigma_i Q_{x_i}\right\|_2 \sup_{U\in\mathcal{U}}\|UU^*\|_2 \\ &\leq \sqrt{mK},\end{aligned}$$

since the Hilbert-Schmidt norm of a K-dimensional projection is $\sqrt{K}$. The result follows upon multiplication with $2/m$ and taking the expectation in μ^m. □

Lemma 5. *For any $S \in \mathcal{L}\left(\mathbb{R}^K\right)$ we have*

$$\mathcal{R}(\mathcal{E}_S, \mu) \leq \frac{4\left(1 + \|S\|_A\right)K}{\sqrt{m}}\sqrt{\frac{\pi}{2}}.$$

Proof. Let $\|x_i\| \leq 1$ and define Gaussian processes Ω_U and Ξ_U indexed by $\mathcal{U}(\mathbb{R}^K, H)$

$$\Omega_U = \sum_{i=1}^{m} \gamma_i \inf_{y \in A} \left(\|U^* x_i - Sy\|^2 + g(y) \right)$$

$$\Xi_U = 2\left(1 + \|S\|_A\right) \sum_{k=1}^{K} \sum_{i=1}^{m} \gamma_{ik} \langle x_i, U e_k \rangle ,$$

where the e_k are the canonical basis of $\mathbb{R}^K$. For $U_1, U_2 \in \mathcal{U}(\mathbb{R}^K, H)$ we have

$$\begin{aligned}
\mathbb{E}\left(\Omega_{U_1} - \Omega_{U_2}\right)^2 &\leq \sum_{i=1}^{m} \sup_{y \in A} \langle U_1^* x_i - U_2^* x_i, U_1^* x_i + U_2^* x_i - 2Sy \rangle^2 \\
&\leq \sum_{i=1}^{m} \|U_1^* x_i - U_2^* x_i\|^2 \sup_{y \in A} \|U_1^* x_i + U_2^* x_i - 2Sy\|^2 \\
&\leq 4\left(1 + \|S\|_A\right)^2 \sum_{i=1}^{m} \sum_{k=1}^{K} \left(\langle x_i, U_1 e_k \rangle - \langle x_i, U_2 e_k \rangle \right)^2 \\
&= \mathbb{E}\left(\Xi_{U_1} - \Xi_{U_2}\right)^2 .
\end{aligned}$$

It follows from Lemma 1 and Slepians lemma (Theorem 3) that

$$\mathcal{R}_m\left(\mathcal{E}_S, \mu\right) \leq \mathbb{E}_{\mathbf{x} \sim \mu^m} \frac{2}{m} \sqrt{\frac{\pi}{2}} \mathbb{E}_{\boldsymbol{\gamma}} \sup_U \Xi_U ,$$

so the result follows from the following inequalities, using Schwarz' and Jensens inequality, the orthonormality of the γ_{ik} and the fact that $\|x_i\| \leq 1$ on the support of μ.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\gamma}} \sup_U \Xi_U &= 2\left(1 + \|S\|_A\right) \mathbb{E} \sup_U \sum_{k=1}^{K} \left\langle \sum_{i=1}^{m} \gamma_{ik} x_i, U e_k \right\rangle \\
&\leq 2\left(1 + \|S\|_A\right) \sum_{k=1}^{K} \mathbb{E} \left\| \sum_{i=1}^{m} \gamma_{ik} x_i \right\| \\
&\leq 2\left(1 + \|S\|_A\right) K \sqrt{m}.
\end{aligned}$$

□

Using the subadditivity of the Rademacher complexity, the last two results give for $K > 1$ and $\|\mathcal{C}\|_A \geq 1$

$$\begin{aligned}
\sup_{S \in \mathcal{S}} \mathcal{R}_m\left(\mathcal{G}_S, \mu\right) &\leq \mathcal{R}_m\left(\mathcal{D}, \mu\right) + \sup_{S \in \mathcal{S}} \mathcal{R}_m\left(\mathcal{E}_S, \mu\right) \\
&\leq \frac{1}{\sqrt{m}} \left(2\sqrt{K} + 8K \|\mathcal{C}\|_A \sqrt{\frac{\pi}{2}} \right) \leq \frac{12K \|\mathcal{C}\|_A}{\sqrt{m}},
\end{aligned}$$

and substitution in Lemma 3 gives Theorem 1.

References

1. P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3: 463–482, 2002.
2. P. Bartlett, T. Linder, G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44: 1802–1813, 1998.
3. G. Biau, L. Devroye, G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790, 2008.
4. F. Cucker and S. Smale. On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, 39 (1):1–49, 2001.
5. W. Hoeffding. Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association*, 58:13–30, 1963.
6. P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.
7. V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers, *The Annals of Statistics*, 30(1): 1–50, 2002.
8. M. Ledoux, M. Talagrand. *Probability in Banach Spaces*, Springer, 1991.
9. D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791, 1999.
10. S. Z. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized parts-based representations. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Vol. I, pages 207–212, Hawaii, USA, 2001.
11. C. McDiarmid. *Concentration*, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, p195-248, Springer, Berlin, 1998.
12. B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.
13. J. Shawe-Taylor, C. K. I. Williams, N. Cristianini, J. S. Kandola. On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory* 51(7): 2510–2522, 2005.
14. O. Wigelius, A. Ambroladze, J. Shawe-Taylor. Statistical analysis of clustering with applications. Preprint, 2007.
15. D. Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.*, 41: 463–501, 1962.
16. A.W. van der Vaart and J.A. Wallner. *Weak Convergence and Empirical Processes*, Springer Verlag, 1996.
17. L. Zwald, L., O. Bousquet, and G. Blanchart. Statistical properties of kernel principal component analysis. *Machine Learning* 66(2-3): 259–294, 2006.