Dimensionality reduction in Hilbert spaces

Maxim Raginsky

October 29, 2013

Dimensionality reduction is a generic name for any procedure that takes a complicated object living in a high-dimensional (or possibly even infinite-dimensional) space and approximates it in some sense by a finite-dimensional vector. We are interested in a particular class of dimensionality reduction methods. Consider a data source that generates vectors in some Hilbert space $\mathscr{H}$, which is either infinite-dimensional or has a finite but extremely large dimension (think $\mathbb{R}^d$ with the usual Euclidean norm, where d is huge). We will assume that the vectors of interest lie in the unit ball of $\mathscr{H}$,

$$B(\mathscr{H}) \triangleq \Big\{x \in \mathscr{H} : \|x\| \leq 1\Big\},$$

where $\|x\| = \sqrt{\langle x, x\rangle}$ is the norm on $\mathscr{H}$. We wish to represent each $x \in B(\mathscr{H})$ by a vector $\widehat{y} \in \mathbb{R}^k$ for some fixed k (if $\mathscr{H}$ is d-dimensional, then of course we must have $d \gg k$). For instance, k may represent some storage limitation, such as a device that can store no more than k real numbers (or, more realistically, k double-precision floating-point numbers, which for all practical purposes can be thought of as real numbers). The mapping $x \mapsto \widehat{y}$ can be thought of as an *encoding* rule. In addition, given $\widehat{y} \in \mathbb{R}^k$, we need a *decoding* rule that takes $\widehat{y}$ and outputs a vector $\widehat{x} \in \mathscr{H}$ that will serve as an approximation of x. In general, the cascade of mappings

$$x \xrightarrow{\text{encoding}} \widehat{y} \xrightarrow{\text{decoding}} \widehat{x}$$

will be lossy, i.e., $x \neq \widehat{x}$. So, the goal is to ensure that the squared norm error $\|x - \widehat{x}\|^2$ is as small as possible. In this lecture, we will see how Rademacher complexity techniques can be used to characterize the performance of a particular fairly broad class of dimensionality reduction schemes in Hilbert spaces. Our exposition here is based on a beautiful recent paper of Maurer and Pontil [MP10].

We will consider a particular type of dimensionality reduction schemes, where the encoder is a (nonlinear) projection, whereas the decoder is a linear operator from $\mathbb{R}^k$ into $\mathscr{H}$ (the Appendix contains some basic facts pertaining to linear operators between Hilbert spaces). To specify such a scheme, we fix a pair (Y, T) consisting of a closed set $\mathsf{Y} \subseteq \mathbb{R}^k$ and a linear operator $T : \mathbb{R}^k \to \mathscr{H}$. We call Y the *codebook* and use the encoding rule

$$\widehat{y} = \arg\min_{y \in \mathsf{Y}} \|x - Ty\|^2. \tag{1}$$

Unless Y is a closed subspace of $\mathbb{R}^k$, this encoding map will be nonlinear. The decoding, on the other hand, is linear: $\widehat{x} = T\widehat{y}$. With these definitions, the reconstruction error is given by

$$\|x - \widehat{x}\|^2 = f_T(x) \triangleq \min_{y \in \mathsf{Y}} \|x - Ty\|^2.$$

Now suppose that the input to our dimensionality reduction scheme is a *random vector* $X \in B(\mathscr{H})$ with some unknown distribution P. Then we measure the performance of the coding scheme (Y, T) by its expected reconstruction error

$$L(T) \triangleq \mathbb{E}_P[f_T(X)] \equiv \mathbb{E}_P\left[\min_{y\in\mathsf{Y}} \|X - Ty\|^2\right]$$

(note that, even though the reconstruction error depends on the codebook Y, we do not explicitly indicate this dependence, since the choice of Y will be fixed by a particular application). Now let $\mathscr{T}$ be some fixed class of admissible linear decoding maps $T : \mathbb{R}^k \to \mathscr{H}$. So, if we knew P, we could find the best decoder $\tilde{T} \in \mathscr{T}$ that achieves

$$L^*(\mathscr{T}) \triangleq \inf_{T\in\mathscr{T}} L(T)$$

(assuming, of course, that the infimum exists and is achieved by at least one $T \in \mathscr{T}$).

By now, you know the drill: We don't know P, but we have access to a large set of samples $X_1, \ldots, X_n$ drawn i.i.d. from P. So we attempt to learn $\tilde{T}$ via ERM:

$$\begin{aligned}\widehat{T}_n &\triangleq \operatorname*{arg\,min}_{T\in\mathscr{T}} \frac{1}{n}\sum_{i=1}^n f_T(X_i) \\ &= \operatorname*{arg\,min}_{T\in\mathscr{T}} \frac{1}{n}\sum_{i=1}^n \min_{y\in\mathsf{Y}} \|X_i - Ty\|^2.\end{aligned}$$

Our goal is to establish the following result:

Theorem 1. *Assume that* Y *is a closed subset of the unit ball* $B_2^k = \{y \in \mathbb{R}^k : \|y\|_2 \le 1\}$*, and that every* $T \in \mathscr{T}$ *satisfies*

$$\begin{aligned}\|Te_j\| &\le \alpha, \qquad 1 \le j \le k \\ \|T\|_{\mathsf{Y}} &\triangleq \sup_{y\in\mathsf{Y}, y\neq 0} \|Ty\| \le \alpha\end{aligned}$$

for some finite $\alpha \ge 1$*. Then*

$$L(\widehat{T}_n) \le L^*(\mathscr{T}) + \frac{60\alpha^2 k^2}{\sqrt{n}} + 4\alpha^2\sqrt{\frac{2\log(1/\delta)}{n}} \tag{2}$$

with probability at lest $1-\delta$*. In the special case when* $\mathsf{Y} = \{e_1, \ldots, e_k\}$*, the standard basis in* $\mathbb{R}^k$*, the event*

$$L(\widehat{T}_n) \le L^*(\mathscr{T}) + \frac{40\alpha^2 k}{\sqrt{n}} + 4\alpha^2\sqrt{\frac{2\log(1/\delta)}{n}} \tag{3}$$

holds with probability at least $1-\delta$*.*

Remark 1. The above result is slightly weaker than the one from [MP10]; as a consequence, the constants in Eqs. (2) and (3) are slightly worse than they could otherwise be.

1 Examples

Before we get down to business and prove the theorem, let's look at a few examples.

1.1 Principal component analysis (PCA)

The objective of PCA is, given k, construct a projection Π onto a k-dimensional closed subspace of $\mathscr{H}$ to maximize the average "energy content" of the projected vector:

$$\begin{aligned} &\text{maximize } \mathbb{E}\|\Pi X\|^2 \\ &\text{subject to } \dim \Pi(\mathscr{H}) = k \\ &\qquad\qquad \Pi^2 = \Pi \end{aligned} \tag{4}$$

For any $x \in \mathscr{H}$,

$$\|\Pi x\|^2 = \|x\|^2 - \|(I-\Pi)x\|^2, \tag{5}$$

where I is the identity operator on $\mathscr{H}$. To prove (5), expand the right-hand side:

$$\begin{aligned} \|x\|^2 - \|(I-\Pi)x\|^2 &= \|x\|^2 - \|x - \Pi x\|^2 \\ &= 2\langle x, \Pi x\rangle - \|\Pi x\|^2 \\ &= \|\Pi x\|^2, \end{aligned}$$

where the last step is by the properties of projections. Thus,

$$\begin{aligned} \|\Pi x\|^2 &= \|x\|^2 - \|x - \Pi x\|^2 \\ &= \|x\|^2 - \min_{x' \in \mathscr{K}} \|x - x'\|^2, \end{aligned} \tag{6}$$

where $\mathscr{K}$ is the range of Π (the closure of the linear span of all vectors of the form $\Pi x, x \in \mathscr{H}$). Moreover, any projection operator $\Pi : \mathscr{H} \to \mathscr{K}$ with $\dim(\mathscr{H}) = k$ can be factored as TT^*, where $T : \mathbb{R}^k \to \mathscr{H}$ is an isometry (see Appendix for definitions and the proof of this fact). Using this fact, we can write

$$\mathscr{K} = \Pi(\mathscr{H}) = \left\{ Ty : y \in \mathbb{R}^k \right\}.$$

Using this in (6), we get

$$\|\Pi x\|^2 = \|x\|^2 - \min_{y \in \mathbb{R}^k} \|x - Ty\|^2.$$

Hence, solving the optimization problem (4) is equivalent to finding the best linear decoding map $\tilde{T}$ for the pair $(\mathsf{Y}, \mathscr{T})$, where $\mathsf{Y} = \mathbb{R}^k$ and $\mathscr{T}$ is the collection of all isometries $T : \mathbb{R}^k \to \mathscr{H}$. Moreover, if we recall our assumption that $X \in B(\mathscr{H})$ with probability one, then we see that there is no loss of generality if we take

$$\mathsf{Y} = B_2^k \triangleq \left\{ y \in \mathbb{R}^k : \|y\|_2 \le 1 \right\},$$

i.e., the unit ball in $(\mathbb{R}^k, \|\cdot\|_2)$. This follows from the fact that $\|\Pi x\| \le \|x\|$ for any projection Π, so, in particular, for $\Pi = TT^*$ the encoding $\widehat{y}$ in (1) can be written as $\widehat{y} = T^* x$, and

$$\|\widehat{y}\|_2 = \|T\widehat{y}\| = \|TT^* x\| = \|\Pi x\| \le \|x\| \le 1.$$

Thus, Theorem 1 applies with $\alpha = 1$. That said, there are much tighter bounds for PCA that rely on deeper structural results pertaining to finite-dimensional subspaces of Hilbert spaces, but that is beside the point. The key idea here is that we can already get nice bounds using the tools already at our fingertips.

1.2 Vector quantization or k-means clustering

Vector quantization (or k-means clustering) is a procedure that takes a vector $x \in \mathscr{H}$ and maps it to its nearest neighbor in a finite set $\mathscr{C} = \{\xi_1, \dots, \xi_k\} \subset \mathscr{H}$, where k is a given positive integer:

$$\widehat{x} = \operatorname*{arg\,min}_{\xi \in \mathscr{C}} \|x - \xi\|^2.$$

If X is random with distribution P, then the optimal k-point quantizer is given a size-k set $\tilde{\mathscr{C}} = \{\tilde{\xi}_1, \dots, \tilde{\xi}_k\}$ that minimizes the reconstruction error

$$\mathbb{E}_P \left[\min_{\xi \in \mathscr{C}} \|X - \xi\|^2 \right]$$

over all $\mathscr{C} \subset \mathscr{H}$ with $|\mathscr{C}| = k$. We can cast the problem of finding $\tilde{\mathscr{C}}$ in our framework by taking $\mathsf{Y} = \{e_1, \dots, e_k\}$ (the standard basis in $\mathbb{R}^k$) and letting $\mathscr{T}$ be the set of all linear operators $T : \mathbb{R}^k \to \mathscr{H}$. It is easy to see that any $\mathscr{C} \subset \mathscr{H}$ with $|\mathscr{C}| = k$ can be obtained as an image of the standard basis $\{e_1, \dots, e_k\}$ under some linear operator $T : \mathbb{R}^k \to \mathscr{H}$. Indeed, for any $\mathscr{C} = \{\xi_1, \dots, \xi_k\}$, we can just *define* a linear operator $T : \mathbb{R}^k \to \mathscr{H}$ by

$$Te_j \triangleq \xi_j, \qquad 1 \le j \le k$$

and then extending it to all of $\mathbb{R}^k$ by linearity:

$$T\left(\sum_{j=1}^k y_j e_j \right) = \sum_{j=1}^k y_j Te_j = \sum_{j=1}^k y_j \xi_j.$$

So, another way to interpret the objective of vector quantization is as follows: given a distribution P supported on $B(\mathscr{H})$, we seek a k-element set $\mathscr{C} = \{\xi_1, \dots, \xi_k\} \subset \mathscr{H}$, such that the random vector $X \sim P$ can be well-approximated on average by linear combinations of the form

$$\sum_{j=1}^k y_j \xi_j,$$

where the vector of coefficients $y = (y_1, \dots, y_k)$ can have only one nonzero component, which is furthermore required to be equal to 1. In fact, there is no loss of generality in assuming that $\mathscr{C} \subset B(\mathscr{H})$ as well. This is a consequence of the fact that, for any $x \in B(\mathscr{H})$ and any $x' \in \mathscr{H}$, we can always find some $x'' \in B(\mathscr{H})$ such that

$$\|x - x''\| \le \|x - x'\|.$$

Indeed, it suffices to take $x'' = \operatorname{arg\,min}_{z \in B(\mathscr{H})} \|x' - z\|^2$, and it is not hard to show that $x'' = x'/\|x'\|$.

Thus, Theorem 1 applies with $\alpha = 1$. Moreover, the excess risk grows *linearly* with dimension k, cf. Eq. (3). It is not known whether this linear dependence on k is optimal — there are $\Omega(\sqrt{k/n})$ lower bounds for vector quantization, but it is still an open question whether these lower bounds are tight [MP10].

1.3 Nonnegative matrix factorization

Consider approximating the random vector $X \sim P$, where P is supported on the unit ball $B(\mathscr{H})$, by linear combinations of the form

$$\sum_{j=1}^{k} y_j \xi_j,$$

where the real vector $y = (y_1, \ldots, y_k)$ is constrained to lie in the nonnegative orthant

$$\mathbb{R}^k_+ \triangleq \left\{ y = (y_1, \ldots, y_k) \in \mathbb{R}^k : y_j \geq 0, 1 \leq j \leq k \right\},$$

while the unit vectors $\xi_1, \ldots, \xi_k \in B(\mathscr{H})$ are constrained by the positivity condition

$$\langle \xi_j, \xi_\ell \rangle_{\mathscr{H}} \geq 0, \qquad 1 \leq j, \ell \leq k.$$

This is a generalization of the *nonnegative matrix factorization* (NMF) problem, originally posed by Lee and Seung [LS99].

To cast NMF in our framework, let $\mathsf{Y} = \mathbb{R}^k_+$, and let $\mathscr{T}$ be the set of all linear operators $T : \mathbb{R}^k \to \mathscr{H}$ such that (i) $\|Te_j\| = 1$ for all $1 \leq j \leq k$ and (ii) $\langle Te_j, Te_\ell \rangle \geq 0$ for all $1 \leq j, \ell \leq k$. Then the choice of T is equivalent to the choice of $\xi_1, \ldots, \xi_k \in B(\mathscr{H})$, as above. Moreover, it can be shown that, for any $x \in B(\mathscr{H})$ and any $T \in \mathscr{T}$, the minimum of $\|x - Ty\|^2$ over all $y \in \mathbb{R}^k_+$ is achieved at some $\widehat{y} \in \mathbb{R}^k_+$ with $\|\widehat{y}\|_2 \leq 1$. Thus, there is no loss of generality if we take $\mathsf{Y} = \mathbb{R}^k_+ \cap B^k_2$. In this case, the conditions of Theorem 1 are satisfied with $\alpha = 1$.

1.4 Sparse coding

Take Y to be the ℓ_1 unit ball

$$B^k_1 \triangleq \left\{ y = (y_1, \ldots, y_k) \in \mathbb{R}^k : \|y\|_1 = \sum_{j=1}^{k} |y_j| \leq 1 \right\},$$

and let $\mathscr{T}$ be the collection of all linear operators $T : \mathbb{R}^k \to \mathscr{H}$ with $\|Te_j\| \leq 1$ for all $1 \leq j \leq k$. In this case, the dimensionality reduction problem is to approximate a random $X \in B(\mathscr{H})$ by a linear combination of the form

$$\sum_{j=1}^{k} y_j \xi_j,$$

where $y = (y_1, \ldots, y_k) \in \mathbb{R}^k$ satisfies the constraint $\|y\|_1 \leq 1$, while the vectors $\xi_1, \ldots, \xi_k$ belong to the unit ball $B(\mathscr{H})$. Then for any $y = \sum_{j=1}^{k} y_j e_j \in \mathsf{Y}$ we have

$$\begin{aligned} \|Ty\| &= \left\| \sum_{j=1}^{k} y_j Te_j \right\| \\ &\leq \sum_{j=1}^{k} |y_j| \|Te_j\| \\ &\leq \|y\|_1 \cdot \max_{1 \leq j \leq k} \|Te_j\| \\ &\leq 1, \end{aligned}$$

where the third line is by Hölder's inequality. Then the conditions of Theorem 1 are satisfied with $\alpha = 1$.

2 Proof of Theorem 1

Now we turn to the proof of Theorem 1. The format of the proof is the familiar one: if we consider the empirical reconstruction error

$$\begin{aligned} L_n(T) &\triangleq \frac{1}{n}\sum_{i=1}^n f_T(X_i) \\ &= \frac{1}{n}\sum_{i=1}^n \min_{y\in\mathsf{Y}} \|X_i - Ty\|^2 \end{aligned}$$

for every $T \in \mathscr{T}$ and define the uniform deviation

$$\Delta_n(X^n) \triangleq \sup_{T\in\mathscr{T}} |L_n(T) - L(T)|, \tag{7}$$

then

$$L(\widehat{T}_n) \le L^*(\mathscr{T}) + 2\Delta_n(X^n).$$

Now, for any $x \in B(\mathscr{H})$, any $y \in \mathsf{Y}$, and any $T \in \mathscr{T}$, we have

$$0 \le \|x - Ty\|^2 \le 2\|x\|^2 + 2\|Ty\|^2 \le 4\alpha^2.$$

Thus, the uniform deviation $\Delta_n(X^n)$ has bounded differences with $c_1 = \ldots = c_n = 4\alpha^2/n$, so by McDiarmid's inequality,

$$L(\widehat{T}_n) \le L^*(\mathscr{T}) + 2\mathbb{E}\Delta_n(X^n) + 4\alpha^2\sqrt{\frac{2\log(1/\delta)}{n}}, \tag{8}$$

with probability at least $1-\delta$. By the usual symmetrization argument, we obtain the bound $\mathbb{E}\Delta_n(X^n) \le 2\mathbb{E}R_n(\mathscr{F}(X^n))$, where $\mathscr{F}$ is the class of functions f_T for all $T \in \mathscr{T}$. Now, the whole affair hinges on getting a good upper bound on the Rademacher averages $R_n(\mathscr{F}(X^n))$. We will do this in several steps, and we need to introduce some additional machinery along the way.

2.1 Gaussian averages

Let $\gamma_1, \ldots, \gamma_n$ be i.i.d. standard normal random variables. In analogy to the Rademacher average of a bounded set $\mathscr{A} \subset \mathbb{R}^n$, we can define the *Gaussian average* of $\mathscr{A}$ [BM02] as

$$G_n(\mathscr{A}) \triangleq \mathbb{E}_{\gamma^n} \sup_{a\in\mathscr{A}} \left| \frac{1}{n}\sum_{i=1}^n \gamma_i a_i \right|.$$

Lemma 1 (Gaussian averages vs. Rademacher averages)**.**

$$R_n(\mathscr{A}) \le \sqrt{\frac{\pi}{2}} G_n(\mathscr{A}). \tag{9}$$

Proof. Let $\sigma^n = (\sigma_1, \ldots, \sigma_n)$ be an n-tuple of i.i.d. Rademacher random variables independent of γ^n. Since each γ_i is a symmetric random variable, it has the same distribution as $\sigma_i|\gamma_i|$. Therefore,

$$\begin{aligned}
G_n(\mathscr{A}) &= \frac{1}{n}\mathbb{E}_{\gamma^n}\sup_{a\in\mathscr{A}}\left|\sum_{i=1}^n \gamma_i a_i\right| \\
&= \frac{1}{n}\mathbb{E}_{\sigma^n}\mathbb{E}_{\gamma^n}\sup_{a\in\mathscr{A}}\left|\sum_{i=1}^n \sigma_i|\gamma_i| a_i\right| \\
&\geq \frac{1}{n}\mathbb{E}_{\sigma^n}\sup_{a\in\mathscr{A}}\left|\sum_{i=1}^n \sigma_i a_i \mathbb{E}_{\gamma_i}|\gamma_i|\right| \\
&= \mathbb{E}|\gamma_1| \cdot \frac{1}{n}\mathbb{E}_{\sigma^n}\sup_{a\in\mathscr{A}}\left|\sum_{i=1}^n \sigma_i a_i\right| \\
&= \mathbb{E}|\gamma_1| R_n(\mathscr{A}),
\end{aligned}$$

where the second step is by convexity, while in the last step we have used the fact that $\gamma_1, \ldots, \gamma_n$ are i.i.d. random variables. Now, if γ is a standard normal random variable, then

$$\begin{aligned}
\mathbb{E}|\gamma| &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |t| e^{-t^2/2} dt \\
&= \frac{1}{\sqrt{2\pi}}\int_0^{\infty} t e^{-t^2/2} dt - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} t e^{-t^2/2} dt \\
&= \sqrt{\frac{2}{\pi}}\int_0^{\infty} t e^{-t^2/2} dt \\
&= \sqrt{\frac{2}{\pi}}.
\end{aligned}$$

Rearranging, we get (9). □

Gaussian averages are often easier to work with than Rademacher averages. The reason for this is that, for any n real constants $a_1, \ldots, a_n$, the sum $W_a \triangleq a_1\gamma_1 + \ldots + a_n\gamma_n$ is a Gaussian random variable with mean 0 and variance $a_1^2 + \ldots + a_n^2$. Moreover, for any finite collection of vectors $a^{(1)}, \ldots, a^{(m)} \in \mathscr{A}$, the random variables $W_{a^{(1)}}, \ldots, W_{a^{(m)}}$ are jointly Gaussian. Thus, the collection of random variables $(W_a)_{a\in\mathscr{A}}$ is a zero-mean *Gaussian process*, where we say that a collection of real-valued random variables $(W_a)_{a\in\mathscr{A}}$ is a Gaussian process if all finite linear combinations of the W_a's are Gaussian random variables. In particular, we can compute covariances: for any $a, a' \in \mathscr{A}$,

$$\begin{aligned}
\mathbb{E}[W_a W_{a'}] &= \mathbb{E}\left[\sum_{i=1}^n\sum_{j=1}^n \gamma_i\gamma_j a_i a'_j\right] \\
&= \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[\gamma_i\gamma_j] a_i a'_j \\
&= \sum_{i=1}^n a_i a'_i \\
&= \langle a, a'\rangle
\end{aligned}$$

and things like

$$\begin{aligned}\mathbb{E}[(W_a - W_{a'})^2] &= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n \gamma_i \gamma_j (a_i - a'_i)(a_j - a'_j)\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[\gamma_i \gamma_j\right](a_i - a'_i)(a_j - a'_j) \\ &= \sum_{i=1}^n (a_i - a'_i)^2 \\ &= \|a - a'\|^2.\end{aligned}$$

The latter quantities are handy because of a very useful result called *Slepian's lemma* [Sle62, LT91]:

Lemma 2. *Let $(W_a)_{a\in\mathscr{A}}$ and $(V_a)_{a\in\mathscr{A}}$ be two zero-mean Gaussian processes with some index set $\mathscr{A}$ (not necessarily a subset of $\mathbb{R}^n$), such that*

$$\mathbb{E}[(W_a - W_{a'})^2] \le \mathbb{E}[(V_a - V_{a'})^2], \qquad \forall a, a' \in \mathscr{A}. \tag{10}$$

Then

$$\mathbb{E}\sup_{a\in\mathscr{A}} W_a \le \mathbb{E}\sup_{a\in\mathscr{A}} V_a. \tag{11}$$

Slepian's lemma is typically used to obtain upper bounds on the expected supremum of one Gaussian process in terms of another, which is hopefully easier to handle. The only wrinkle is that we can't apply Slepian's lemma to the problem of estimating the Gaussian average $G_n(\mathscr{A})$ because of the absolute value. However, if all $a \in \mathscr{A}$ are uniformly bounded in norm, the absolute value makes little difference:

Lemma 3. *Let $\mathscr{A} \subset \mathbb{R}^n$ be a set of vectors uniformly bounded in norm, i.e., there exists some $L < \infty$ such that $\|a\| \le L$ for all $a \in \mathscr{A}$. Let*

$$\tilde{G}_n(\mathscr{A}) \triangleq \frac{1}{n}\mathbb{E}\left[\sup_{a\in\mathscr{A}} \sum_{i=1}^n \gamma_i a_i\right]. \tag{12}$$

Then

$$\tilde{G}_n(\mathscr{A}) \le G_n(\mathscr{A}) \le 2\tilde{G}_n(\mathscr{A}) + \sqrt{\frac{2}{\pi}}\frac{L}{n}. \tag{13}$$

Proof. The first inequality in (13) is obvious. For the second inequality, pick an arbitrary $a' \in \mathscr{A}$, let $W_a = \sum_{i=1}^n \gamma_i a_i$ for any $a \in \mathscr{A}$, and write

$$\begin{aligned}G_n(\mathscr{A}) &= \frac{1}{n}\mathbb{E}\left[\sup_{a\in\mathscr{A}} |W_a|\right] \\ &\le \frac{1}{n}\mathbb{E}\left[\sup_{a\in\mathscr{A}} |W_a - W_{a'}|\right] + \frac{1}{n}\mathbb{E}|W_{a'}|.\end{aligned}$$

Since a' was arbitrary, this gives

$$\begin{aligned}G_n(\mathscr{A}) &\le \sup_{a'\in\mathscr{A}}\left\{\frac{1}{n}\mathbb{E}\left[\sup_{a\in\mathscr{A}} |W_a - W_{a'}|\right] + \frac{1}{n}\mathbb{E}|W_{a'}|.\right\} \\ &\le \frac{1}{n}\mathbb{E}\left[\sup_{a,a'\in\mathscr{A}} |W_a - W_{a'}|\right] + \frac{1}{n}\sup_{a'\in\mathscr{A}}\mathbb{E}|W_{a'}|.\end{aligned} \tag{14}$$

For any two a, a', the random variable $W_a - W_{a'}$ is symmetric, so

$$\mathbb{E}\left[\sup_{a,a' \in \mathscr{A}} |W_a - W_{a'}|\right] = 2\mathbb{E}\left[\sup_{a \in \mathscr{A}} W_a\right].$$

Moreover, for any $a' \in \mathscr{A}$, $W_{a'}$ is Gaussian with zero mean and variance $\|a'\|^2 \le L^2$. Thus,

$$\sup_{a' \in \mathscr{A}} \mathbb{E}|W_{a'}| \le L\mathbb{E}|\gamma| = \sqrt{\frac{2}{\pi}} L.$$

Using the two above formulas in (14), we get the second inequality in (13), and the lemma is proved. □

Armed with this lemma, we can work with the quantity $\tilde{G}_n(\mathscr{A})$ instead of the Gaussian average $G_n(\mathscr{A})$. The advantage is that now we can rely on tools like Slepian's lemma.

2.2 Bounding the Rademacher average

Now everything hinges on bounding the Gaussian average $G_n(\mathscr{F}(x^n))$ for a fixed sample $x^n = (x_1, \ldots, x_n)$, which in turn will give us a bound on the Rademacher average $R_n(\mathscr{F}(x^n))$, by Lemmas 1 and 3. Let $(\gamma_i)_{1 \le i \le n}$, $(\gamma_{ij})_{1 \le i \le n, 1 \le j \le k}$, and $(\gamma_{ij\ell})_{1 \le i \le n, 1 \le j, \ell \le k}$ be mutually independent sequences of i.i.d. standard Gaussian random variables. Define the following zero-mean Gaussian processes, indexed by $T \in \mathscr{T}$:

$$\begin{aligned}
W_T &\triangleq \sum_{i=1}^{n} \gamma_i f_T(x_i), \\
V_T &\triangleq \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \langle x_i, Te_j \rangle, \\
U_T &\triangleq \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{\ell=1}^{k} \gamma_{ij\ell} \langle Te_j, Te_\ell \rangle, \\
\Upsilon_T &\triangleq \sqrt{8} V_T + \sqrt{2} U_T.
\end{aligned}$$

By definition,

$$\begin{aligned}
G_n(\mathscr{F}(x^n)) &= \mathbb{E} \sup_{T \in \mathscr{T}} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma_i f_T(x_i) \right| \\
&= \frac{1}{n} \mathbb{E} \sup_{T \in \mathscr{T}} |W_T|,
\end{aligned}$$

and we define $\tilde{G}_n(\mathscr{F}(x^n))$ similarly. We will use Slepian's lemma to upper-bound $\tilde{G}_n(\mathscr{F}(x^n))$ in terms of expected suprema of $(V_T)_{T\in\mathscr{T}}$ and $(U_T)_{T\in\mathscr{T}}$. To that end, we start with

$$\begin{aligned}
\mathbb{E}\left[(W_T - W_{T'})^2\right] &= \sum_{i=1}^n \left(f_T(x_i) - f_{T'}(x_i)\right)^2 \\
&= \sum_{i=1}^n \left(\min_{y\in\mathsf{Y}} \|x_i - Ty\|^2 - \min_{y\in\mathsf{Y}} \|x_i - T'y\|^2\right)^2 \\
&\le \sum_{i=1}^n \left(\max_{y\in\mathsf{Y}} \left|\|x_i - Ty\|^2 - \|x_i - T'y\|^2\right|\right)^2 \\
&= \sum_{i=1}^n \left(\max_{y\in\mathsf{Y}} \left|2\langle x_i, Ty - T'y\rangle + \|Ty\|^2 - \|T'y\|^2\right|\right)^2 \\
&\le 8\sum_{i=1}^n \max_{y\in\mathsf{Y}} \left|\langle x_i, Ty - T'y\rangle\right|^2 + 2\sum_{i=1}^n \max_{y\in\mathsf{Y}} \left(\|Ty\|^2 - \|T'y\|^2\right)^2, \qquad (15)
\end{aligned}$$

where in the third line we have used properties of inner products, and the last line is by the inequality $(a+b)^2 \le 2a^2 + 2b^2$. Now, for each i,

$$\begin{aligned}
\max_{y\in\mathsf{Y}} \left|\langle x_i, Ty - T'y\rangle\right| &= \max_{y\in\mathsf{Y}} \left|\sum_{j=1}^k y_j \langle x_i, Te_j - T'e_j\rangle\right|^2 \\
&\le \max_{y\in\mathsf{Y}} \|y\|_2^2 \sum_{j=1}^k \left|\langle x_i, Te_j - T'e_j\rangle\right|^2 \\
&\le \sum_{j=1}^k \left|\langle x_i, Te_j - T'e_j\rangle\right|^2,
\end{aligned}$$

where in the second step we have used Cauchy–Schwarz. Summing over $1 \le i \le n$, we see that

$$\begin{aligned}
\sum_{i=1}^n \max_{y\in\mathsf{Y}} \left|\langle x_i, Ty - T'y\rangle\right| &\le \sum_{i=1}^n \sum_{j=1}^k \left|\langle x_i, Te_j - T'e_j\rangle\right|^2 \\
&= \mathbb{E}\left[(V_T - V_{T'})^2\right]. \qquad (16)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\max_{y\in\mathsf{Y}} \left(\|Ty\|^2 - \|T'y\|^2\right)^2 &= \max_{y\in\mathsf{Y}} \left(\sum_{j=1}^k \sum_{\ell=1}^k y_j y_\ell \langle Te_j, Te_\ell\rangle - \langle T'e_j, T'e_\ell\rangle\right)^2 \\
&\le \max_{y\in\mathsf{Y}} \sum_{j=1}^k \sum_{\ell=1}^k y_j^2 y_\ell^2 \cdot \sum_{j=1}^k \sum_{\ell=1}^k \left(\langle Te_j, Te_\ell\rangle - \langle T'e_j, T'e_\ell\rangle\right)^2 \\
&= \max_{y\in\mathsf{Y}} \|y\|_2^4 \cdot \sum_{j=1}^k \sum_{\ell=1}^k \left(\langle Te_j, Te_\ell\rangle - \langle T'e_j, T'e_\ell\rangle\right)^2 \\
&\le \sum_{j=1}^k \sum_{\ell=1}^k \left(\langle Te_j, Te_\ell\rangle - \langle T'e_j, T'e_\ell\rangle\right)^2.
\end{aligned}$$

Therefore,

$$\sum_{i=1}^n \max_{y\in\mathsf{Y}} \left(\|Ty\|^2 - \|T'y\|^2\right) \le \mathbb{E}\left[(U_T - U_{T'})^2\right]. \qquad (17)$$

Using (16) and (17) in (15), we have

$$\begin{aligned}\mathbb{E}\left[(W_T - W_{T'})^2\right] &\le 8\,\mathbb{E}\left[(V_T - V_{T'})^2\right] + 2\,\mathbb{E}\left[(U_T - U_{T'})^2\right] \\ &= \mathbb{E}\left[(\Upsilon_T - \Upsilon_{T'})^2\right].\end{aligned}$$

We can therefore apply Slepian's lemma (Lemma 2) to $(W_T)_{T\in\mathcal{T}}$ and $(\Upsilon_T)_{T\in\mathcal{T}}$ to write

$$\begin{aligned}\tilde{G}_n(\mathcal{F}(x^n)) &= \frac{1}{n}\mathbb{E}\sup_{T\in\mathcal{T}} W_T \\ &\le \frac{1}{n}\mathbb{E}\sup_{T\in\mathcal{T}} \Upsilon_T \\ &\le \frac{\sqrt{8}}{n}\,\mathbb{E}\sup_{T\in\mathcal{T}} V_T + \frac{\sqrt{2}}{n}\,\mathbb{E}\sup_{T\in\mathcal{T}} U_T. \end{aligned} \tag{18}$$

We now upper-bound the expected suprema of V_T and U_T. For the former,

$$\begin{aligned}\mathbb{E}\sup_{T\in\mathcal{T}} V_T &= \mathbb{E}\sup_{T\in\mathcal{T}} \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}\langle x_i, Te_j\rangle \\ &= \mathbb{E}\sup_{T\in\mathcal{T}} \sum_{j=1}^{k}\left\langle \sum_{i=1}^{n}\gamma_{ij}x_i, Te_j\right\rangle && \text{(linearity)} \\ &\le \mathbb{E}\sup_{T\in\mathcal{T}} \sum_{j=1}^{k}\left\|\sum_{i=1}^{n}\gamma_{ij}x_i\right\| \|Te_j\| && \text{(Cauchy–Schwarz)} \\ &\le \mathbb{E}\sum_{j=1}^{k}\left\|\sum_{i=1}^{n}\gamma_{ij}x_i\right\| \sup_{T\in\mathcal{T}}\|Te_j\| \\ &\le \alpha\sum_{j=1}^{k}\mathbb{E}\left\|\sum_{i=1}^{n}\gamma_{ij}x_i\right\| && \text{(assumption on } \|T\|) \\ &\le \alpha\sum_{j=1}^{k}\mathbb{E}\sqrt{\sum_{i=1}^{n}\sum_{i'=1}^{n}\gamma_{ij}\gamma_{i'j}\langle x_i, x_i'\rangle} && \text{(linearity)} \\ &\le \alpha\sum_{j=1}^{k}\sqrt{\sum_{i=1}^{n}\sum_{i'=1}^{n}\mathbb{E}\left[\gamma_{ij}\gamma_{i'j}\right]\langle x_i, x_i'\rangle} && \text{(Jensen)} \\ &= \alpha\sum_{j=1}^{k}\sqrt{\sum_{i=1}^{n}\|x_i\|^2} && \text{(properties of i.i.d. Gaussians)} \\ &\le \alpha k\sqrt{n}. && (x_i \in B(\mathcal{H}) \text{ for all } i)\end{aligned}$$

Similarly, for the latter,

$$\begin{aligned}
\mathbb{E}\sup_{T\in\mathscr{T}} U_T &= \mathbb{E}\sup_{T\in\mathscr{T}} \sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\ell=1}^{k}\gamma_{ij\ell}\langle Te_j, Te_\ell\rangle \\
&\le \sum_{j=1}^{k}\sum_{\ell=1}^{k}\mathbb{E}\sup_{T\in\mathscr{T}}\sum_{i=1}^{n}\gamma_{ij\ell}\langle Te_j, Te_\ell\rangle \\
&\le \sum_{j=1}^{k}\sum_{\ell=1}^{k}\mathbb{E}\left|\sum_{i=1}^{n}\gamma_{ij\ell}\right|\sup_{T\in\mathscr{T}}\|Te_j\|\|Te_\ell\| \\
&\le \alpha^2k^2\sqrt{\frac{2n}{\pi}}.
\end{aligned}$$

Substituting these bounds into (18), we have

$$\tilde{G}_n(\mathscr{F}(x^n)) \le \frac{1}{n}\left(\alpha k\sqrt{8n} + \alpha^2k^2\frac{2\sqrt{n}}{\sqrt{\pi}}\right) \le \frac{5\alpha^2k^2}{\sqrt{n}}.$$

Thus, applying Lemmas 1 and 3, we have

$$\begin{aligned}
R_n(\mathscr{F}(x^n)) &\le \sqrt{\frac{\pi}{2}}G_n(\mathscr{F}(x^n)) \\
&\le \sqrt{\frac{\pi}{2}}\left[2\tilde{G}_n(\mathscr{F}(x^n)) + \sqrt{\frac{2}{\pi}}\frac{\max_{T\in\mathscr{T}}\sqrt{\sum_{i=1}^{n}|f_T(x_i)|^2}}{n}\right] \\
&\le \sqrt{\frac{\pi}{2}}\left[\frac{10\alpha^2k^2}{\sqrt{n}} + \sqrt{\frac{2}{\pi}}\frac{2\alpha}{\sqrt{n}}\right] \\
&\le \frac{15\alpha^2k^2}{\sqrt{n}}
\end{aligned}$$

Recalling (8), we see that the event (2) holds with probability at least $1-\delta$.

For the special case of k-means clustering, i.e., when $\mathsf{Y} = \{e_1,\dots,e_k\}$, we follow a slightly different strategy. Define a zero-mean Gaussian process

$$\Xi_T \triangleq \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\|x_i - Te_j\|^2, \qquad T\in\mathscr{T}.$$

Then

$$\begin{aligned}
\mathbb{E}\left[(W_T - W_{T'})^2\right] &= \sum_{i=1}^{n}\left(\min_{1\le j\le k}\|x_i - Te_j\|^2 - \min_{1\le j\le k}\|_i - T'e_j\|^2\right)^2 \\
&\sum_{i=1}^{n}\max_{1\le j\le k}\left(\|x_i - Te_j\|^2 - \|x_i - T'e_j\|^2\right)^2 \\
&\le \sum_{i=1}^{n}\sum_{j=1}^{k}\left(\|x_i - Te_j\|^2 - \|x_i - T'e_j\|^2\right)^2 \\
&= \mathbb{E}\left[(\Xi_T - \Xi_{T'})^2\right].
\end{aligned}$$

For the process (Ξ_T), we have

$$\begin{aligned}
\mathbb{E}\sup_{T\in\mathscr{T}}\Xi_T &= \mathbb{E}\sup_{T\in\mathscr{T}}\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\|x_i-Te_j\|^2 \\
&= \mathbb{E}\sup_{T\in\mathscr{T}}\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\left\{\|x_i\|^2-2\langle x_i,Te_j\rangle+\|Te_j\|^2\right\} \\
&\le \sum_{i=1}^{k}\mathbb{E}\sup_{T\in\mathscr{T}}\left\{2\sum_{i=1}^{n}\gamma_{ij}|\langle x_i,Te_j\rangle|+\sum_{i=1}^{n}\gamma_{ij}\|Te_j\|^2\right\} \\
&\le 3k\alpha^2\sqrt{n},
\end{aligned}$$

where the methods used to obtain this bound are similar to what we did for (V_T) and (U_T). Using Lemmas 1–3, we have

$$\begin{aligned}
R_n(\mathscr{F}(x^n)) &\le \sqrt{\frac{\pi}{2}}G_n(\mathscr{F}(x^n)) \\
&\le \sqrt{\frac{\pi}{2}}\left[2\tilde{G}_n(\mathscr{F}(x^n))+\sqrt{\frac{2}{\pi}}\frac{\max_{T\in\mathscr{T}}\sqrt{\sum_{i=1}^{n}|f_T(x_i)|^2}}{n}\right] \\
&\le \sqrt{\frac{\pi}{2}}\left[\frac{6\alpha^2k}{\sqrt{n}}+\sqrt{\frac{2}{\pi}}\frac{2\alpha}{\sqrt{n}}\right] \\
&\le \frac{10\alpha^2k}{\sqrt{n}}.
\end{aligned}$$

Again, recalling (8), we see that the event (3) occurs with probability at least $1-\delta$. The proof of Theorem 1 is complete.

A Linear operators between Hilbert spaces

We assume, for simplicity, that all Hilbert spaces $\mathscr{H}$ of interest are *separable*. By definition, a Hilbert space $\mathscr{H}$ is separable if it has a countable dense subset: there exists a countable set $\{h_1,h_2,\ldots\}\subset\mathscr{H}$, such that for any $h\in\mathscr{H}$ and any $\varepsilon>0$ there exists some $j\in\mathbb{N}$, for which $\|h-h_j\|_{\mathscr{H}}<\varepsilon$. Any separable Hilbert space $\mathscr{H}$ has a countable complete and orthonormal basis, i.e., a countable set $\{\varphi_1,\varphi_2,\ldots\}\subset\mathscr{H}$ with the following properties:

1. **Orthonormality** — $\langle\varphi_i,\varphi_j\rangle_{\mathscr{H}}=\delta_{ij}$;
2. **Completeness** — if there exists some $h\in\mathscr{H}$ which is orthogonal to all φ_j's, i.e., $\langle h,\varphi_j\rangle=0$ for all j, then $h=0$.

As a consequence, any $h\in\mathscr{H}$ can be uniquely represented as an infinite linear combination

$$h=\sum_{j=1}^{\infty}c_j\varphi_j,\qquad\text{where } c_j=\langle h,\varphi_j\rangle_{\mathscr{H}},$$

where the infinite series converges in norm, i.e., for any $\varepsilon > 0$ there exists some $n \in \mathbb{N}$, such that

$$\left\| \varphi - \sum_{j=1}^{n} c_j \varphi_j \right\|_{\mathcal{H}} < \varepsilon.$$

Moreover, $\|h\|_{\mathcal{H}}^2 = \sum_{j=1}^{\infty} |c_j|^2$.

Let $\mathcal{H}$ and $\mathcal{K}$ be two Hilbert spaces. A *linear operator* from $\mathcal{H}$ into $\mathcal{K}$ is a mapping $T : \mathcal{H} \to \mathcal{K}$, such that (i) $T(\alpha h + \alpha' h') = \alpha T h + \alpha' T h'$ for any two $h, h' \in \mathcal{H}$ and $\alpha, \alpha' \in \mathbb{R}$. A linear operator $T : \mathcal{H} \to \mathcal{K}$ is *bounded* if

$$\|T\|_{\mathcal{H} \to \mathcal{K}} \triangleq \sup_{h \in \mathcal{H}, h \neq 0} \frac{\|Th\|_{\mathcal{K}}}{\|h\|_{\mathcal{H}}} < \infty.$$

We will denote the space of all bounded linear operators $T : \mathcal{H} \to \mathcal{K}$ by $\mathcal{L}(\mathcal{H}, \mathcal{K})$. When $\mathcal{H} = \mathcal{K}$, we will write $\mathcal{L}(\mathcal{H})$ instead. For any operator $T \in \mathcal{L}(\mathcal{H}, \mathcal{K})$, we have the *adjoint operator* $T^* \in \mathcal{L}(\mathcal{K}, \mathcal{H})$, which is characterized by

$$\langle g, Th \rangle_{\mathcal{K}} = \langle T^* g, h \rangle_{\mathcal{H}}, \qquad \forall g \in \mathcal{K}, h \in \mathcal{H}.$$

If $T \in \mathcal{L}(\mathcal{H})$ has the property that $T = T^*$, we say that T is *self-adjoint*.

Some examples:

- The *identity operator* on $\mathcal{H}$, denoted by $I_{\mathcal{H}}$, maps each $h \in \mathcal{H}$ to itself. $I_{\mathcal{H}}$ is a self-adjoint operator with $\|I_{\mathcal{H}}\| \equiv \|I_{\mathcal{H}}\|_{\mathcal{H} \to \mathcal{H}} = 1$. We will often omit the index $\mathcal{H}$ and just write I.

- A *projection* is an operator $\Pi \in \mathcal{L}(\mathcal{H})$ satisfying $\Pi^2 = \Pi$, i.e., $\Pi(\Pi h) = \Pi h$ for any $h \in \mathcal{H}$. This is a bounded operator with $\|\Pi\| = 1$. Any projection is self-adjoint.

- An *isometry* is an operator $T \in \mathcal{L}(\mathcal{H}, \mathcal{K})$, such that $\|Th\|_{\mathcal{K}} = \|h\|_{\mathcal{H}}$ for all $h \in \mathcal{H}$, i.e., T preserves norms. If T is an isometry, then $T^* T = I_{\mathcal{H}}$, while $TT^* \in \mathcal{L}(\mathcal{K})$ is a projection. This is easy to see:

  $$(TT^*)(TT^*) = T(T^* T)T^* = TT^*.$$

  If $T \in \mathcal{L}(\mathcal{H})$ and $T^* \in \mathcal{L}(\mathcal{H})$ are both isometries, then T is called a *unitary operator*.

- If $\Pi \in \mathcal{L}(\mathcal{H})$ is a projection whose range $\mathcal{K} \subseteq \mathcal{H}$ is a closed k-dimensional subspace, then there exists an isometry $T \in \mathcal{L}(\mathbb{R}^k, \mathcal{K})$, such that $\Pi = TT^*$. Here, $\mathbb{R}^k$ is a Hilbert space with the usual $\|\cdot\|_2$ norm. To see this, let $\{\psi_1, \ldots, \psi_k\} \subset \mathcal{H}$ be an orthonormal basis of $\mathcal{K}$, and complete it to a countable basis $\{\psi_1, \psi_2, \ldots, \psi_k, \psi_{k+1}, \psi_{k+2}, \ldots\}$ for the entire $\mathcal{H}$. Here, the elements of $\{\psi_j\}_{j=k+1}^{\infty}$ are mutually orthonormal and orthogonal to $\{\psi_j\}_{j=1}^{k}$. Any $h \in \mathcal{H}$ has a unique representation

  $$h = \sum_{j=1}^{\infty} \alpha_j \psi_j$$

  for some real coefficients $\alpha_1, \alpha_2, \ldots$. With this, we can write out the action of Π explicitly as

  $$\Pi h = \sum_{j=1}^{k} \alpha_j \psi_j.$$

Now consider the map $T : \mathbb{R}^k \to \mathscr{K}$ that takes

$$\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k \longmapsto \sum_{j=1}^{k} \alpha_j \psi_j.$$

It is easy to see that T is an isometry. Indeed,

$$\|T\alpha\|_{\mathscr{H}} = \left\| \sum_{j=1}^{k} \alpha_j \psi_j \right\|_{\mathscr{H}} = \sqrt{\sum_{j=1}^{k} \alpha_j^2} = \|\alpha\|_2.$$

The adjoint of T is easily computed: for any $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k$ and any $h' = \sum_{j=1}^{\infty} \alpha'_j \psi_j \in \mathscr{H}$,

$$\begin{aligned}
\langle h', T\alpha \rangle_{\mathscr{H}} &= \langle \Pi h', T\alpha \rangle_{\mathscr{H}} \\
&= \left\langle \sum_{j=1}^{k} \alpha'_j \psi_j, \sum_{j=1}^{k} \alpha_j \psi_j \right\rangle \\
&= \sum_{j=1}^{k} \alpha'_j \alpha_j \\
&= \langle T^* h', \alpha \rangle.
\end{aligned}$$

Since this must hold for arbitrary $\alpha \in \mathbb{R}^k$ and $h' \in \mathscr{H}'$, we must have $T^* h' = T^* \left(\sum_{j=1}^{\infty} \alpha'_j \psi_j \right) = (\alpha'_1, \ldots, \alpha'_j)$. Now let's compute $T^* h$ for any $h = \sum_j \alpha_j \psi_j$:

$$\begin{aligned}
T T^* h &= T(T^* h) \\
&= T\left(T^* \left(\sum_{j=1}^{\infty} \alpha_j \psi_j \right) \right) \\
&= T\left((\alpha_1, \ldots, \alpha_k) \right) \\
&= \sum_{j=1}^{k} \alpha_j \psi_j \\
&= \Pi h.
\end{aligned}$$

References

[BM02] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[LS99] D. D. Lee and H. S. Seung. Learing the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

[MP10] A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, November 2010.

[Sle62] D. Slepian. The one-sided barrier problem for Gaussian noise. *Bell Systems Technical Journal*, 41:463–501, 1962.