Lower bounds on the distance to low dimensional random subspaces

Andreas Maurer

Adalbertstrasse 55
D-80799 München, Germany
andreasmaurer@compuserve.com

Abstract. We give a probabilistic lower bound on the distance from a fixed point to the span of a small iid sample of a bounded random vector with values in a Hilbert space. The bound is expressed in terms of the eigenvalues of the covariance operator associated with the random vector. Applied to linear regression with square loss it leads to nonparametric, small sample size lower error bounds for kernel algorithms.

1 Introduction

Let $\mathbf{X} = (X_1, ..., X_m)$ be a sequence of iid random variables with values in a Hilbert space H. With $[\mathbf{X}]$ we denote the linear span of the X_i, so that $[\mathbf{X}]$ is a random subspace. Given a vector $u \in H$ we are interested in lower bounds on the random variable

$$d(u, [\mathbf{X}]) = \min_{y \in [\mathbf{X}]} \|u - y\| .$$

Such lower bounds have to depend on some form of high-dimensionality of the distribution of the random variable $X = X_1$. Apart from a boundedness constraint on the random variable $\|X\|^2$ we will describe this distribution only through properties of its second order moments, more specifically in terms of the trace $\|C\|_1$ and the largest eigenvalue $\|C\|_\infty$ of the covariance operator C. If $\|C\|_1$ is large, but $\|C\|_\infty$ is small, then the dimensionality is large. This is illustrated when X is concentrated and uniformly distributed on the intersection of the unit sphere with an N-dimensional subspace in H (we will abbreviate this circumstance by saying that X is N-spherical). Then $\|C\|_1 = 1$, but $\|C\|_\infty = 1/N$. Our main result is the following:

Theorem 1. *Let b be such that $\|X\|^2 \leq b$ a.s. If*

$$4m\left(1 - b^{-1}\|C\|_1\right) + 16m^2 b^{-1}\|C\|_\infty < \delta/2$$

then with probability at least $1 - \delta$

$$d(u, [\mathbf{X}]) \geq \|u\| \sqrt{1 - \frac{4m\|C\|_\infty}{b\delta}}.$$

Substituting the values of the norms when X is N-spherical shows that the lower bound converges to $\|u\|$ for every $\delta > 0$ as $N \to \infty$. For finite N we obtain the bound

$$d\left(u, [\mathbf{X}]\right) \geq \|u\| \sqrt{\frac{N - 4m\delta^{-1}}{N}},$$

which resembles the lower bound obtained explicitely for the N-spherical case (see [8]), but the additional condition $m < \sqrt{N\delta/32}$ shows that the present result is considerably weaker, while it is much more general because it only depends on the sequence of eigenvalues of the covariance operator. We will prove Theorem 1 in Section 3.

Our result can be applied to linear regression with square loss. The vector u is to be thought of as a target function, while $\mathbf{X}$ is the input part of the training sample for some learning algorithm f. A large class of learning algorithms, among them all kernel-techniques, have the property that the hypotheses they generate have to lie in the subspace spanned by the input sample. A lower bound on the distance from u to $[\mathbf{X}]$ then implies a lower bound for the distance from this hypothesis to the target function and is therefore indicative of a lower error bound for the algorithm f. The expected square error is in fact equal to the square of the distance for an appropriately redefined inner product. These ideas are spelled out in Section 4 and lead to the following result:

Theorem 2. *If $\|X\| \leq 1$ a.s. and $4m\left(1 - \|C\|_2^2 / \|C\|_\infty\right) + 16m^2 \|C\|_\infty < \delta/2$, then for every $u \in H$ with probability at least $1 - \delta$*

$$\min_{w \in [\mathbf{X}]} \frac{\mathbb{E}\left(\langle w, X\rangle - \langle u, X\rangle\right)^2}{\mathbb{E}\left\langle u, X\right\rangle^2} \geq 1 - 4m \|C\|_\infty / \delta.$$

The $\mathbb{E}\left\langle u, X\right\rangle^2$ in the denominator normalizes the expected error of the trivial hypothesis $w = 0$ to 1. Here $\|C\|_2^2$ is the sum of the squared eigenvalues of C. If X is N-spherical we obtain a lower bound of the normalized error of $1 - 4m/\left(N\delta\right)$ under the condition that $m < \sqrt{N\delta/32}$.

Examples of nontrivial lower bounds are not only obtained for the N-spherical case. For our bounds the only relevant features of N-spherical random variables are the peak-power constraint $\|X\|^2 \leq 1$ and the fact that the (descending) sequence $\eta\left(N\right)$ of covariance-eigenvalues has the form

$$\eta_k\left(N\right) = \left\{ \begin{array}{c} 1/N \text{ if } 1 \leq k \leq N \\ 0 \quad \text{if} \quad N < k \end{array} \right. .$$

Any random variable X with $\|X\|^2 \leq 1$ where C has the same sequence of eigenvalues will lead to the same bound. An example would be the uniform distribution on the extreme points of the ℓ_1-unit ball in $\mathbb{R}^N$ embedded in H. It is also not surprising, and in fact easy to show, that nontrivial lower bounds follow for any random variable with $\|X\|^2 \leq 1$ and a covariance operator whose sequence of eigenvalues is a sufficiently small perturbation of $\eta(N)$ for sufficiently large N. This point will be discussed further in Section 5.

Bounds on the distances to random subspaces generated by Gaussian or N-spherical random variables have been studied in [2]. In these special cases exponential tail bounds can be derived which then allow a simple proof of the Johnson Lindenstrauss Lemma. In our case only second order information is available leading to much weaker tail bounds derived from Markov's inequality. An essential element of our proof is related to a bound on the condition number of the sample covariance matrix, a problem analyzed in detail by Edelman [3], but again for Gaussian variables. Exponential concentration of $d(u, [\mathbf{X}])$ in the N-spherical case is used in [8] to give small-sample lower error bounds for halfspace learning from the uniform distribution for any unitarily equivariant algorithm. Theorem 2 above appears to be the first such lower bound valid for a generic class of distributions and all target funtions.

2 Notation and definitions

Throughout this paper H will be a real separable Hilbert space, which may be finite of infinite dimensional. The inner product and norm on H are denoted $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$. If T is a bounded positive operator with trivial null-space then a new inner product $\langle \cdot, \cdot \rangle_T$ is defined on H by $\langle y, z \rangle_T = \langle Ty, z \rangle$. The corresponding norm is denoted $\|\cdot\|_T$.

For a linear operator T on H we use the following norms

$$\begin{aligned}
\|T\|_1 &= \sup \left\{ \sum_k \langle Te_k, e_k \rangle : (e) \text{ is an orthonormal basis of } H \right\} \\
\|T\|_2 &= \sup \left\{ \left(\sum_k \|Te_k\|^2 \right)^{1/2} : (e) \text{ is an orthonormal basis of } H \right\} \\
\|T\|_\infty &= \sup \{ \|Tx\| : x \in H, \|x\| \leq 1 \}.
\end{aligned}$$

If the quantities in the first two definitions are finite, then the supremum is attained and independent of the chosen basis. If T is a positive trace class operator then these norms coincide with the usual ℓ_p-norms on the sequences of eigenvalues of T (see [9] for more details on these norms).

These definitions also apply to the Hilbert space $\mathbb{R}^m$ instead of H and to the $m \times m$ matrices acting in $\mathbb{R}^m$. We will use the same notation in this case, the nature of the various arguments should be unambiguous in the respective context. For $m \times m$ matrices the norm $\|\cdot\|_2$ is often called the Frobenius norm, for operators on H it is called the Hilbert Schmidt norm.

There is a distinguished random variable X taking values in H. We will always assume a peak power-constraint, either $\|X\|^2 \leq b$ a.s. or more specifically

$\|X\|^2 \leq 1$. With $\mathbf{X} = (X_1, ..., X_m)$ we denote the vector obtained from m iid copies of X.

With the random variable X we associate an operator C, the covariance operator, on H, defined by

$$\langle Cy, z\rangle = \mathbb{E}\,\langle y, X\rangle\,\langle X, z\rangle\,, \forall y, z \in H.$$

Observe that the relationship between the random variable X and the covariance operator depends on the inner product, and for different inner products we obtain different covariance operators from the same random variable. If the inner product is given by a positive operator T, as $\langle\cdot,\cdot\rangle_T$ above, then we denote the corresponding covariance operator by C_T. We have, for any $y, z \in H$,

$$\langle C_T y, z\rangle_T = \mathbb{E}\,\langle y, X\rangle_T\,\langle X, z\rangle_T = \mathbb{E}\,\langle Ty, X\rangle\,\langle X, Tz\rangle = \langle CTy, Tz\rangle = \langle CTy, z\rangle_T\,,$$

so that $C_T = CT$.

The covariance operator is clearly symmetric and nonnegative definite. If (e) is an orthonormal basis of H then

$$\|C\|_1 = \sum_k \langle Ce_k, e_k\rangle = \mathbb{E}\,\langle e_k, X\rangle\,\langle X, e_k\rangle = \mathbb{E}\,\|X\|^2\,, \tag{1}$$

so by the peak-power constraint C is trace class. With (λ) we denote the sequence of eigenvalues of C in nonincreasing order $\lambda_1 \geq \lambda_2 \geq ...$. If the random variable X has the form $\psi(\omega)$ where $\psi : \Omega \to H$ is some feature map from a probability space (Ω, μ) to H then the sequence (λ) is identical to the sequence of eigenvalues of the integral operator K on $L_2(\mu)$ given by

$$(Kf)(x) = \int f(y)\,\langle\psi(y), \psi(x)\rangle\, d\mu(y)\,.$$

See [5] and [10] for details on the interpretation of the covariance operator in kernel methods.

For $i \neq j$, since X_i and X_j are independent, we get

$$\mathbb{E}\,\langle X_i, X_j\rangle^2 = \mathbb{E}\,\langle CX_j, X_j\rangle \leq \|C\|_\infty\,\mathbb{E}\,\|X\|^2\,. \tag{2}$$

The linear subspace spanned by $X_1, ..., X_m$ will be denoted by $[\mathbf{X}]$. With $P_{[\mathbf{X}]}$ we denote the orthogonal projection in H onto the subspace $[\mathbf{X}]$. From the definition of $d(u, [\mathbf{X}])$ in the introduction we have

$$d(u, [\mathbf{X}]) = \sqrt{\|u\|^2 - \left\|P_{[\mathbf{X}]}u\right\|^2}, \tag{3}$$

so a lower bound on $d(u, [\mathbf{X}])$ can be obtained from an upper bound on $\left\|P_{[\mathbf{X}]}u\right\|^2$.

With $\mathbf{X}$ we associate a random $m \times m$ matrix $G(\mathbf{X})$, the empirical Gramian (or kernel matrix), defined by

$$G_{ij}(\mathbf{X}) = \langle X_i, X_j\rangle\,.$$

To estimate the smallest eigenvalue of $G(\mathbf{X})$ we will use the following theorem [6, Theorem 4.2.2].

Theorem 3 (Rayleigh-Ritz). *If A is a symmetric real matrix, then the smallest eigenvalue $\sigma_{\min}$ of A satisfies*

$$\sigma_{\min}(A) = \min_{\|v\|=1} \langle Av, v\rangle .$$

3 Proof of Theorem 1

We begin with an explicit representation of $P_{[\mathbf{X}]}$ in terms of the inverse Gramian:

Lemma 1. *If $G(\mathbf{X})$ is invertible then for $y \in H$ we have*

$$P_{[\mathbf{X}]}y = \sum_{i,j=1}^{m} \langle y, X_i\rangle G_{ij}^{-1}(\mathbf{X}) X_j .$$

Proof. It is straightforward to verify that the operator T defined by the right hand side above satisfies $Ty = y$ for $y \in [\mathbf{X}]$ and $Ty = 0$ for $y \in [\mathbf{X}]^{\perp}$.

The next lemma is the key to our result and is related to an estimate of the condition number of $G(\mathbf{X})$.

Lemma 2. *If $\|X\|^2 < b$ a.s. then*

$$\Pr\left\{\left\|G^{-1}(\mathbf{X})\right\|_\infty > 2b^{-1}\right\} < 4m\left(1 - b^{-1}\|C\|_1\right) + 16m^2 b^{-1}\|C\|_\infty .$$

Proof. Let Ψ be the event that some $\|X_i\|^2$ is less than $(3/4)\,b$. By a union bound and Markov's inequality we get

$$\begin{aligned}\Pr\Psi &= \Pr\left\{\exists i : \|X_i\|^2 < (3/4)\,b\right\}\\ &\le m\Pr\left\{\|X\|^2 < (3/4)\,b\right\} = m\Pr\left\{b - \|X\|^2 > b/4\right\}\\ &\le \frac{4m\left(b - \mathbb{E}\|X\|^2\right)}{b} = 4m\left(1 - b^{-1}\|C\|_1\right).\end{aligned}$$

Now we define a random $m \times m$ matrix $B(\mathbf{X})$ by

$$B_{ij}(\mathbf{X}) = \begin{cases} 0 & \text{if } i = j \\ \langle X_i, X_j\rangle & \text{if } i \ne j \end{cases},$$

and let Φ be the event that the Frobenius norm of $B(\mathbf{X})$ is larger than $b/4$. By Markov's inequality and the inequality (2) we get

$$\begin{aligned}\Pr\Phi &= \Pr\left\{\|B(\mathbf{X})\|_2 > b/4\right\} = \Pr\left\{\|B(\mathbf{X})\|_2^2 > b^2/16\right\}\\ &\le 16b^{-2}\mathbb{E}\sum_{i\ne j}\langle X_i, X_j\rangle^2 \le 16b^{-2}m^2\|C\|_\infty \mathbb{E}\|X\|^2\\ &\le 16b^{-1}m^2\|C\|_\infty .\end{aligned}$$

Then $\Pr(\varPsi \cup \varPhi) \leq 4m\left(1 - b^{-1}\left\|C\right\|_1\right) + 16m^2 b^{-1}\left\|C\right\|_\infty$. But if neither $\varPsi$ nor $\varPhi$ happen we can apply the Rayleigh-Ritz theorem to get for the smallest eigenvalue $\sigma_{\min}\left(G\left(\mathbf{X}\right)\right)$

$$\begin{aligned}\sigma_{\min}\left(G\left(\mathbf{X}\right)\right) &= \min_{\|v\|=1}\left\langle G\left(\mathbf{X}\right)v, v\right\rangle \\ &= \min_{\|v\|=1}\left(\sum_{i=1}^m v_i^2\left\|X_i\right\|^2 + \sum_{i\neq j} v_i v_j\left\langle X_i, X_j\right\rangle\right) \\ &\geq \min_i\left\|X_i\right\|^2 - \left\|B\left(\mathbf{X}\right)\right\|_2 \\ &\geq \left(3/4\right)b - b/4 = b/2,\end{aligned}$$

so that $G\left(\mathbf{X}\right)$ is invertible and $\left\|G^{-1}\left(\mathbf{X}\right)\right\|_\infty \leq 2b^{-1}$.

Lemma 3. *Let $u \in H$. Define a vector $v\left(u, \mathbf{X}\right) \in \mathbb{R}^m$ by $v\left(u, \mathbf{X}\right)_i = \left\langle u, X_i\right\rangle$. Then*

$$\Pr\left\{\left\|v\left(u, \mathbf{X}\right)\right\|^2 \leq \delta^{-1}m\left\|C\right\|_\infty\left\|u\right\|^2\right\} \geq 1 - \delta.$$

Proof. This is just Markov's inequality

$$\begin{aligned}\Pr\left\{\sum_{i=1}^m\left\langle u, X_i\right\rangle^2 > t\right\} &\leq t^{-1}\mathbb{E}\sum_{i=1}^m\left\langle u, X_i\right\rangle^2 = t^{-1}m\left\langle Cu, u\right\rangle \\ &\leq t^{-1}m\left\|C\right\|_\infty\left\|u\right\|^2,\end{aligned}$$

equating the last expression to δ and solving for t.

Proof (of Theorem 1). By hypothesis $4m\left(1 - b^{-1}\left\|C\right\|_1\right) + 16m^2 b^{-1}\left\|C\right\|_\infty < \delta/2$,so by Lemma 2 $\Pr\left\{\left\|G^{-1}\left(\mathbf{X}\right)\right\|_\infty > 2/b\right\} < \delta/2$. From Lemma 3 we obtain

$$\Pr\left\{\left\|v\left(u, \mathbf{X}\right)\right\|^2 \leq 2\delta^{-1}m\left\|C\right\|_\infty\left\|u\right\|^2\right\} \geq 1 - \delta/2.$$

Combining these two results in a union bound we obtain $\Pr\varPsi \geq 1 - \delta$ where $\varPsi$ is the event

$$\varPsi = \left\{\left\|G^{-1}\left(\mathbf{X}\right)\right\|_\infty \leq 2/b\right\} \cap \left\{\left\|v\left(u, \mathbf{X}\right)\right\|^2 \leq 2\delta^{-1}m\left\|C\right\|_\infty\left\|u\right\|^2\right\}.$$

But in the event of $\varPsi$ we can use Lemma 1 to get with probability at least $1 - \delta$

$$\begin{aligned}\left\|P_{[\mathbf{X}]}u\right\|^2 &= \left\langle P_{[\mathbf{X}]}u, u\right\rangle \\ &= \sum_{i,j=1}^m\left\langle u, X_i\right\rangle G_{ij}^{-1}\left(\mathbf{X}\right)\left\langle X_j, u\right\rangle \\ &\leq \left\|G^{-1}\left(\mathbf{X}\right)\right\|_\infty\left\|v\left(u, \mathbf{X}\right)\right\|^2 \\ &\leq 4b^{-1}\delta^{-1}m\left\|C\right\|_\infty\left\|u\right\|^2.\end{aligned}$$

so from equation (3) we get

$$d\left(u, [\mathbf{X}]\right) = \sqrt{\left\|u\right\|^2 - \left\|P_{[\mathbf{X}]}u\right\|^2} \geq \left\|u\right\|\sqrt{1 - \frac{4m\left\|C\right\|_\infty}{b\delta}}.$$

4 Application to linear regression

For this section we assume $\|X\| \leq 1$ a.s. and that C is nonsingular.

We consider the following situation. Along with the vector $\mathbf{X} = (X_1, ..., X_m)$ a vector $\mathbf{Y} = (\langle u, X_1 \rangle, ..., \langle u, X_m \rangle) \in \mathbb{R}^m$ is observed. The vector $\mathbf{Y}$ depends on the data $\mathbf{X}$ and the target vector u and is supposed to give us a clue on the target vector. The pair $(\mathbf{X}, \mathbf{Y})$ is fed into an algorithm $f : H^m \times \mathbb{R}^m \rightarrow H$, which produces a hypothesis $w = f(\mathbf{X}, \mathbf{Y}) \in H$. The performance of this hypothesis is described by the expected square loss

$$\mathrm{err}_X(u, w) = \mathbb{E}(\langle u, X \rangle - \langle w, X \rangle)^2 .$$

The expected square loss on hypothesis w and target u is equal to

$$\begin{aligned} \mathrm{err}_X(u, w) &= \mathbb{E} \langle u - w, X \rangle \langle X, u - w \rangle = \langle C(u - w), u - w \rangle \\ &= \|u - w\|_C^2 , \end{aligned}$$

where the norm $\|\cdot\|_C$ on H has been induced by the new inner product $\langle \cdot, \cdot \rangle_C$ defined by

$$\langle w, z \rangle_C = \langle Cw, z \rangle .$$

Suppose now that the algorithm f is *unitarily equivariant* in the sense that for every $(\mathbf{x}, \mathbf{y}) \in H^m \times \mathbb{R}^m$ and for every unitary operator V on H we have

$$f(V\mathbf{x}, \mathbf{y}) = f((Vx_1, ..., Vx_m), \mathbf{y}) = Vf(\mathbf{x}, \mathbf{y}) .$$

This condition is satisfied in particular by kernel algorithms (see e.g. [1] for an introduction to kernel techniques). Taking V to be the identity on $[\mathbf{x}]$ and minus the identity on the orthogonal complement $[\mathbf{x}]^\perp$ shows at once that unitarily equivariant algorithms must satisfy $f(\mathbf{x}, \mathbf{y}) \in [\mathbf{x}]$, for all $(\mathbf{x}, \mathbf{y}) \in H^m \times \mathbb{R}^m$. This property is sometimes referred to as a *representer theorem.* Then for $u \in H$ we have

$$\begin{aligned} \mathrm{err}_X(u, f(\mathbf{X}, \mathbf{Y})) &= \|u - f(\mathbf{X}, \mathbf{Y})\|_C^2 \\ &\geq \min_{w \in [\mathbf{X}]} \|u - w\|_C^2 = d_C(u, [\mathbf{X}]) , \end{aligned}$$

where $d_C(u, [\mathbf{X}])$ is just the minimal distance from u to $[\mathbf{X}]$ in the metric derived from the inner product $\langle \cdot, \cdot \rangle_C$. We can therefore apply our theorem to obtain a lower error bound for all unitarily equivariant algorithms and for all target vectors u.

As an almost sure upper bound on $\|X\|_C^2$ we can take $b = \|C\|_\infty$, because

$$\|X\|_C^2 = \langle CX, X \rangle \leq \|C\|_\infty \|X\|^2 = \|C\|_\infty , \text{ a.s.}$$

As explained in Section 2 the covariance operator C_C of X relative to the metric $\langle ., . \rangle_C$ is given by $C_C = C^2$. In particular $\|C_C\|_\infty = \|C\|_\infty^2$ and $\|C_C\|_1 = \|C\|_2^2$. Substitution in Theorem 1 then gives the Theorem 2 stated in the introduction.

This lower bound is quite different from the classical lower bounds in statistical learning theory such as in [4] and [7]. The first of these results holds for every algorithm and a distribution mischievously designed to make the algorithm fail. The second holds for every algorithm and the more benign N-spherical distribution, but the target function is designed to make the algorithm fail. More similar to the present bound is the one in [8] which holds only for unitarily equivariant algorithms and N-spherical distributions, but it also holds for all target functions. Bounds of this type allow to rigorously reason against the use of unitarily equivariant algorithms and to argue in favour of multi-task, or transfer learning algorithms. The present bound is also of this type, but replaces the N-spherical distribution by the more generic class of distributions whose spectral properties are sufficiently similar to those of an N-spherical distribution.

5 The sequence of eigenvalues

In this section we give some crude estimates to exhibit some simple properties of the eigenvalue sequence (λ) sufficient for non-trivial lower bounds. We assume that $\|X\|^2 \leq 1$. a.s.

Let λ_k be the k-th eigenvalue of C in descending order. For $N \in \mathbb{N}$ we define a sequence $\eta(N)$

$$\eta_k(N) = \begin{cases} 1/N \text{ if } 1 \leq k \leq N \\ \;\;0 \quad \text{if} \quad N < k \end{cases},$$

so $\eta(N)$ would be the sequence of eigenvalues of the covariance operator corresponding to an N-spherical random variable. If the sequence λ is very close to $\eta(N)$ in the ℓ_1-distance, then it seems very intuitive that we can regard the underlying distribution as approximately N-dimensional. To make this more precise we denote

$$\epsilon_N = \|\eta(N) - \lambda\|_1 ,$$

and give some simple properties of the sequence ϵ_N.

Lemma 4. *For all $N > 4$ we have*
(i) $1 - \|C\|_1 \leq \epsilon_N$.
(ii) $N^{-1} - \epsilon_N \leq \|C\|_\infty \leq N^{-1} + \epsilon_N$
(iii) if $\epsilon_N < 1/2$ then $N^{-1} - \epsilon_N \leq \|C\|_2^2 \leq N^{-1} + \epsilon_N$

Proof. (i) $1 - \|C\|_1 = \|\eta(N)\|_1 - \|\lambda\|_1 \leq \|\eta(N) - \lambda\|_1 = \epsilon_N$.
(ii) $\left|\|C\|_\infty - N^{-1}\right| = \left|\|\lambda\|_\infty - \|\eta(N)\|_\infty\right| \leq \|\lambda - \eta(N)\|_\infty \leq \|\lambda - \eta(N)\|_1 = \epsilon_N$.
(iii) If $N \geq 4$ and $\epsilon_N < 1/2$ then $2N^{-1} + \epsilon_N < 1$, so using Hölder's inequality and (ii) we have

$$\begin{aligned}\left|\|C\|_2^2 - N^{-1}\right| &= \left|\sum_k \left(\lambda_k^2 - \eta_k^2(N)\right)\right| \leq \sum_k |\lambda_k + \eta_k(N)| \, |\lambda_k - \eta_k(N)| \\ &\leq \|\lambda + \eta(N)\|_\infty \|\lambda - \eta(N)\|_1 \leq \left(\|\lambda\|_\infty + \|\eta(N)\|_\infty\right)\epsilon_N \\ &\leq \left(2N^{-1} + \epsilon_N\right)\epsilon_N \leq \epsilon_N.\end{aligned}$$

Our final result states the lower bounds in terms of the sequence ϵ_N. For nontrivial lower bounds we then require that $N^{-1} + \epsilon_N \ll m^{-2}$.To obtain a nontrivial lower bound for regression with unitarily equivariant algorithms we require in addition that $\epsilon_N \ll N m^{-1}$.

Theorem 4. *Let $\|X\|^2 \leq 1$ a.s. $\delta \in (0,1)$ and $m \geq 1$, $N \geq 4$.*

(i) If $40m^2 \left(N^{-1} + \epsilon_N\right) < \delta$ then with probability at least $1 - \delta$

$$d\left(u, [\mathbf{X}]\right) \geq \|u\| \sqrt{1 - \frac{4m\left(N^{-1} + \epsilon_N\right)}{\delta}}.$$

(ii) If $8mN\epsilon_N + 16m^2 \left(N^{-1} + \epsilon_N\right) < \delta/2$ then, with probability at least $1-\delta$, the appropriately scaled expected square loss satisfies

$$\min_{w \in [\mathbf{X}]} \frac{\mathbb{E}\left(\langle u, X\rangle - \langle w, X\rangle\right)^2}{\mathbb{E}\left\langle u, X\right\rangle^2} \geq 1 - \frac{4m\left(N^{-1} + \epsilon_N\right)}{\delta}.$$

Proof. (i) We apply Theorem 1 with $b = 1$. If $40m^2 \left(N^{-1} + \epsilon_N\right) < \delta$ then by part (i) and (ii) of Lemma 4 we have

$$\begin{aligned} 4m\left(1 - \|C\|_1\right) + 16m^2 \|C\|_\infty &\leq 4m\epsilon_N + 16m^2 \left(N^{-1} + \epsilon_N\right) \\ &\leq 20m^2 \left(N^{-1} + \epsilon_N\right) \leq \delta/2, \end{aligned}$$

so, again by part (ii), with probability at least $1 - \delta$

$$d\left(u, [\mathbf{X}]\right) \geq \|u\| \sqrt{1 - \frac{4m\|C\|_\infty}{\delta}} \geq \|u\| \sqrt{1 - \frac{4m\left(N^{-1} + \epsilon_N\right)}{\delta}}.$$

(ii) We can use Theorem 2 since now $\|X\| = 1$ a.s. If $8mN\epsilon_N + 16m^2 \left(N^{-1} + \epsilon_N\right) < \delta/2$ then $\epsilon_N < 1/2$, so we can use all parts of Lemma 4.Then

$$\begin{aligned} & 4m\left(1 - \|C\|_2^2 / \|C\|_\infty\right) + 16m^2 \|C\|_\infty \\ & \leq 4m\left(1 - \frac{N^{-1} - \epsilon_N}{N^{-1} + \epsilon_N}\right) + 16m^2 \left(N^{-1} + \epsilon_N\right) \\ & = \frac{8mN\epsilon_N}{1 + N\epsilon_N} + 32m^2 N^{-1} < \delta/2, \end{aligned}$$

so the conclusion follows by Theorem 2.

References

1. N. Cristianini and J. Shawe-Taylor, Support Vector Machines, *Cambridge University Press*, 2000.
2. S. Dasgupta, A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22: 60–65, 2003.
3. A. Edelman, *Eigenvalues and Condition Numbers of Random Matrices*, Ph.D. dissertation, Math. Dept., Mass. Inst. Technol., 1989.
4. A. Ehrenfeucht, D. Haussler, M. Kearns, L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3): 247–251, 1989.
5. A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf. Measuring statistical dependence with Hilbert-Schmidt norms. Technical Report 140, Max-Planck-Institut für biologische Kybernetik, 2005.
6. R. A. Horn and C. R. Johnson. Matrix Analysis. Cambridge University Press, 1985.
7. P. M. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556-1559, 1995.
8. A. Maurer and M. Pontil, A uniform lower error bound for half-space learning, *tbp in ALT*, 2008.
9. M. Reed and B. Simon. *Functional Analysis*, part I of *Methods of Mathematical Physics, Academic Press*, 1980.
10. L. Zwald, O. Bousquet, and G. Blanchard, Statistical properties of kernel principal component analysis, Proceedings of COLT 2004, 2004.