Journal of Complexity 27 (2011) 489–499

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

Covering numbers of Gaussian reproducing kernel Hilbert spaces

Thomas Kühn

Mathematisches Institut, Universität Leipzig, Johannisgasse 26, D-04103 Leipzig, Germany

ARTICLE INFO

Article history:
Received 8 September 2010
Accepted 4 January 2011
Available online 22 January 2011

Keywords:
Covering numbers
Gaussian RKHS
Learning theory
Smooth Gaussian processes
Small deviations

ABSTRACT

Metric entropy quantities, like covering numbers or entropy numbers, and positive definite kernels play an important role in mathematical learning theory. Using smoothness properties of the Fourier transform of the kernels, Zhou [D.-X. Zhou, The covering number in learning theory, J. Complexity 18 (3) (2002) 739–767] proved an upper estimate for the covering numbers of the unit ball of Gaussian reproducing kernel Hilbert spaces (RKHSs), considered as a subset of the space of continuous functions.

In this note we determine the *exact asymptotic order* of these covering numbers, exploiting an explicit description of Gaussian RKHSs via orthonormal bases. We show that Zhou's estimate is almost sharp (up to a double logarithmic factor), but his conjecture on the correct asymptotic rate is far too optimistic. Moreover we give an application of our entropy results to small deviations of certain smooth Gaussian processes.

1. Introduction

The pioneering paper by Cucker and Smale [6] gave a new impetus to the statistical theory of learning, which studies how to "learn" unknown objects from random samples. In particular they demonstrated the important role of functional analytic methods in this context. For example, in order to estimate the probabilistic error and the number of samples required for a given confidence level and a given error bound, metric entropy quantities such as covering and entropy numbers are very useful, see also the monographs by Cucker and Zhou [7] and by Steinwart and Christmann [17] and the paper by Williamson et al. [21].

The concept of metric entropy is very basic and general, it has numerous applications in many other branches of mathematics, e.g. in approximation theory, probability theory (small deviation

E-mail address: kuehn@math.uni-leipzig.de.

0885-064X/$ – see front matter
doi:10.1016/j.jco.2011.01.005

problems for stochastic processes), operator theory (eigenvalue distributions of compact operators), PDEs (spectral theory of pseudodifferential operators). For more information on these subjects we refer to the articles by Kuelbs and Li [12] and Li and Linde [14] on small ball problems for Gaussian measures, the monographs by König [11] and Pietsch [16] on eigenvalues of compact operators in Banach spaces, and the book by Edmunds and Triebel [8] on function spaces and spectral theory of PDEs.

If A is a subset of a metric space M and $\varepsilon > 0$, the *covering number* $\mathcal{N}(\varepsilon, A; M)$ is defined as the minimal number of balls in M of radius ε which cover the set A. The centers of these balls form an *ε-net of A in M*. A possibly wider known but closely related notion is Kolmogorov's ε-entropy $\mathcal{H}(\varepsilon, A; M) := \log \mathcal{N}(\varepsilon, A; M)$, see e.g. [10]. Obviously,

$$A \text{ is precompact} \iff \mathcal{N}(\varepsilon, A; M) < \infty \quad \text{for all } \varepsilon > 0.$$

Thus the growth rate of $\mathcal{N}(\varepsilon, A; M)$ or $\mathcal{H}(\varepsilon, A; M)$ as $\varepsilon \to 0$ can be viewed as a measure of the "degree of compactness" or "massiveness" of the set A.

Many modern machine learning methods such as support vector machines use Gaussian radial basis functions, which generate a reproducing kernel Hilbert space. Motivated by these facts, Zhou [23] studied covering numbers of the unit ball of RKHSs, considered as a subset of the space of continuous functions. The results were expressed in terms of smoothness properties of the Fourier transform of the kernels. To illustrate his general results he gave an upper estimate for these covering numbers in the case of Gaussian RKHSs (see Example 4 on p. 761 ff. in [23]) and stated a conjecture about the exact asymptotic behaviour (p. 763 f. in [23]). For further results in this direction see also [24,20].

Using completely different methods we determine the exact asymptotic behaviour of these covering numbers. It turns out that Zhou's upper estimate is almost sharp, up to a double logarithmic factor, but his conjecture is too optimistic. Essential tools in our proof are recent results by Steinwart et al. [18] on the structure of Gaussian RKHSs, in particular we exploit the specific orthonormal bases (ONB) of these spaces given in [18]. A quite interesting detail of the proof of the lower bound is the fact that finite sections of the famous Hilbert matrix come into play.

As an application we obtain the sharp asymptotic rate of small deviation probabilities of certain smooth Gaussian processes. Here we use the close connection between metric entropy and small deviations that was discovered by Kuelbs and Li [12].

The organization of the paper is as follows. In Section 2 we describe the necessary background for our main result. We introduce covering numbers of (bounded linear) operators between Banach spaces, a variant of the covering numbers defined above, and state some simple properties that will be needed in the sequel. Moreover, we briefly recall some general facts from the theory of RKHSs and the result from [18] on ONBs in Gaussian RKHSs. In Section 3 we state and prove our main result on covering numbers, and Section 4 contains the application to small deviation probabilities.

2. Preliminaries

In this section we fix notation and recall some well-known basic facts concerning the two concepts mentioned in the title—covering numbers and reproducing kernel Hilbert spaces. Throughout the paper we consider only *real* Banach spaces, and "operator" always means "bounded linear operator between Banach spaces". The Euclidean norm in any $\mathbb{R}^d$ will be denoted by $\|\cdot\|_2$. For functions $f, g : (0, \infty) \to \mathbb{R}$ we write

$$f(\varepsilon) \sim g(\varepsilon) (\text{strong equivalence}), \quad \text{if } \lim_{\varepsilon \to 0} \frac{f(\varepsilon)}{g(\varepsilon)} = 1 \quad \text{and}$$

$$f(\varepsilon) \asymp g(\varepsilon) (\text{weak equivalence}), \quad \text{if } 0 < \liminf_{\varepsilon \to 0} \frac{f(\varepsilon)}{g(\varepsilon)} \le \limsup_{\varepsilon \to 0} \frac{f(\varepsilon)}{g(\varepsilon)} < \infty.$$

The same notation will be used for sequences and $n \to \infty$.

2.1. Covering numbers of operators

For the formulation and proofs of our results it is convenient to introduce a variant of covering numbers for operators. Let $\varepsilon > 0$ and X, Y be Banach spaces with unit balls B_X and B_Y, respectively. Then the *covering numbers* of an operator $T : X \to Y$ are defined as

$$\mathcal{N}(\varepsilon, T) = \min\left\{ n \in \mathbb{N} : \exists y_1, \ldots, y_n \in Y \text{ s.t. } T(B_X) \subseteq \bigcup_{k=1}^{n} (y_k + \varepsilon B_Y) \right\},$$

i.e. $\mathcal{N}(\varepsilon, T) = \mathcal{N}(\varepsilon, T(B_X); Y)$. Clearly, for $\varepsilon \geq \|T\|$ one has $\mathcal{N}(\varepsilon, T) = 1$.

In the following lemma we collect the properties of covering numbers that will be needed in the sequel. We skip the simple proofs which are along the same lines as the proofs for the corresponding properties of entropy numbers, see e.g. [11, Sections 1.d and 2.d]. In particular the last two inequalitities are due to volume arguments.

Lemma 1. *Let $S, T : X \to Y$ and $R : Z \to X$ be operators in (real) Banach spaces and $\varepsilon, \delta > 0$. Then*

$$\mathcal{N}(\varepsilon + \delta, T + S) \leq \mathcal{N}(\varepsilon, T) \cdot \mathcal{N}(\delta, S) \tag{1}$$

$$\mathcal{N}(\varepsilon\delta, TR) \leq \mathcal{N}(\varepsilon, T) \cdot \mathcal{N}(\delta, R) \tag{2}$$

$$\mathcal{N}(\varepsilon, T) \leq \left(1 + \frac{2\|T\|}{\varepsilon}\right)^{\operatorname{rank} T} \quad \text{if } \operatorname{rank} T < \infty \tag{3}$$

$$\mathcal{N}(\varepsilon, T) \geq |\det T| \cdot \left(\frac{1}{\varepsilon}\right)^{n} \quad \text{if } T : X \to X \text{ and } \dim X = n. \tag{4}$$

As an easy consequence we have for operators $T : H \to G$ acting between two (real) n-dimensional Hilbert spaces the estimate

$$\mathcal{N}(\varepsilon, T) \geq \left(\frac{1}{\varepsilon}\right)^{n} \sqrt{\det(T^*T)}. \tag{5}$$

Indeed, select any orthogonal map $A : G \to H$, then (2) implies $\mathcal{N}(\varepsilon, T) = \mathcal{N}(\varepsilon, AT)$. By elementary properties of determinants we have

$$(\det(AT))^2 = \det((AT)^*) \det(AT) = \det(T^*A^*AT) = \det(T^*T),$$

and applying (4) to the operator $AT : H \to H$ we obtain (5).

2.2. Gaussian reproducing kernel Hilbert spaces

The concept of reproducing kernel Hilbert spaces (RKHSs) is widely used in mathematics. The first systematic treatment was given by Aronszajn [1] in 1950, and his article is still a standard reference on this subject.

We consider here only *real-valued kernels* $K : X \times X \to \mathbb{R}$ defined on a non-empty set X. A kernel K is called *positive definite*, if for all $n \in \mathbb{N}$ and every choice of elements $x_1, \ldots, x_n \in X$ and $a_1, \ldots, a_n \in \mathbb{R}$ the inequality

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$$

holds. Consider the functions $k_x : X \to \mathbb{R}$, defined by $k_x(y) := K(x, y)$. The RKHS $H_K(X)$ generated by the kernel K is defined as the completion of $\operatorname{span}\{k_x : x \in X\}$ with respect to the inner product $\langle k_x, k_y \rangle := K(x, y)$. Since the completion is only defined up to isometric isomorphisms, one should say

more precisely that among all possible completions there is a unique space *whose elements are functions* from X to $\mathbb{R}$, and this is $H_K(X)$. The details of this construction can be found e.g. in [17, Chapter 4]. All RKHSs $H_K(X)$ possess the reproducing property

$$f(x) = \langle f, k_x \rangle \quad \text{for all } x \in X \text{ and } f \in H_K(X),$$

which is extremely useful in applications.

Prominent examples of positive definite kernels are Gaussian radial basis function (RBF) kernels, or short *Gaussian kernels*, i.e. functions of the form

$$K(x, y) = \exp(-\sigma^2 \|x - y\|_2^2) \tag{6}$$

defined on a subset $X \subset \mathbb{R}^d$ with non-empty interior. Here $\sigma > 0$ is an arbitrary parameter whose inverse $\frac{1}{\sigma}$ is called *width* in the machine learning literature. Although these kernels are frequently used in modern machine learning methods, very little was known about the structure of the corresponding RKHSs until this question was addressed in [18]. In particular, for Gaussian kernels of the form (6) the authors found explicit orthonormal bases of the corresponding RKHSs $H_\sigma(X)$. In the proofs they used methods from complex analysis for related complex-valued Gaussian kernels on subsets of $\mathbb{C}^d$.

For simplicity we will restrict ourselves to the special case when X is the d-dimensional unit cube $[0, 1]^d$. First we need some notation. For $\sigma > 0$ and $n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ consider the functions $e_n : [0, 1] \to \mathbb{R}$ given by

$$e_n(x) := \sqrt{\frac{(2\sigma^2)^n}{n!}} x^n \exp(-\sigma^2 x^2).$$

For any multi-index $\nu = (n_1, \ldots, n_d) \in \mathbb{N}_0^d$ let $|\nu| := n_1 + \cdots + n_d$ and define the function $e_\nu : [0, 1]^d \to \mathbb{R}$ by

$$e_\nu(x) = \prod_{j=1}^{d} e_{n_j}(x_j) \quad \text{for } x = (x_1, \ldots, x_d).$$

We also use the tensor product notation $e_\nu = e_{n_1} \otimes \cdots \otimes e_{n_d}$ for these functions.

Lemma 2 (*[18, Theorem 8]*). *Let $\sigma > 0$ and $d \in \mathbb{N}$. Then the system $\{e_\nu : \nu \in \mathbb{N}_0^d\}$ is an orthonormal basis in the reproducing kernel Hilbert space $H_\sigma([0, 1]^d)$ generated by the kernel (6).*

3. Main result and proof

In connection with mathematical learning theory Zhou [23] proved – among many other things – that for every $\sigma > 0$ and $d \in \mathbb{N}$ the covering numbers $\mathcal{N}(\varepsilon)$ of the unit ball of the RKHS $H_\sigma([0, 1]^d)$, considered as a subset of $C([0, 1]^d)$, behave asymptotically like

$$\log \mathcal{N}(\varepsilon) = \mathcal{O}\left(\left(\log \frac{1}{\varepsilon}\right)^{d+1}\right) \quad \text{as } \varepsilon \to 0$$

and conjectured that the correct bound is $\left(\log \frac{1}{\varepsilon}\right)^{\frac{d}{2}+1}$. Moreover he proved a lower bound of order $\left(\log \frac{1}{\varepsilon}\right)^{\frac{d}{2}}$, see [24, Example 1 on p. 1747]. Our main result in this note shows that Zhou's upper bound is almost correct, up to a double logarithmic factor, but his conjecture was much too optimistic.

Theorem 3. *Let $\sigma > 0$ and $d \in \mathbb{N}$. Then the covering numbers of the embedding $I_{\sigma,d} : H_\sigma([0, 1]^d) \to C([0, 1]^d)$ behave asymptotically like*

$$\log \mathcal{N}(\varepsilon, I_{\sigma,d}) \asymp \frac{\left(\log \frac{1}{\varepsilon}\right)^{d+1}}{\left(\log \log \frac{1}{\varepsilon}\right)^{d}} \quad \text{as } \varepsilon \to 0. \tag{7}$$

The same is true for $I_{\sigma,d} : H_\sigma([0, 1]^d) \to L_p([0, 1]^d)$, $2 \le p < \infty$.

Proof. Due to the obvious norm estimates $\|f\|_2 \le \|f\|_p \le \|f\|_\infty$ it is enough to prove the upper bound for the covering numbers with respect to the sup-norm and the lower bound with respect to the L_2-norm.

Step 1. *Upper estimate for the* sup*-norm.*

First we determine the norm of the operator $I_{\sigma,d} : H_\sigma([0,1]^d) \to C([0,1]^d)$. If $H_K(X)$ is a RKHS generated by a bounded kernel K on X, then

$$\|id : H_K(X) \to \ell_\infty(X)\| = \sup_{x\in X} \sqrt{K(x,x)}, \tag{8}$$

see [17, Lemma 4.23]. As a direct consequence we get

$$\|I_{\sigma,d} : H_\sigma([0,1]^d) \to C([0,1]^d)\| = 1.$$

Nevertheless we give an alternative proof that is based on the special ONB in $H_\sigma([0,1]^d)$, because the same arguments will be needed below in further norm estimates which do not follow from the general result (8). We have

$$\begin{aligned}
\|I_{\sigma,d}\|^2 &= \sup_{\|f\|_{H_\sigma}=1} \sup_{x\in[0,1]^d} |f(x)|^2 \\
&= \sup_{x\in[0,1]^d} \sup_{\|f\|_{H_\sigma}=1} \left| \sum_{\nu\in\mathbb{N}_0^d} \langle f, e_\nu\rangle e_\nu(x) \right|^2 \\
&= \sup_{x\in[0,1]^d} \sum_{\nu\in\mathbb{N}_0^d} e_\nu(x)^2 \\
&= \sup_{x\in[0,1]^d} \sum_{\nu\in\mathbb{N}_0^d} \prod_{j=1}^d \frac{(2\sigma^2 x_j^2)^{n_j}}{n_j!} \cdot \exp(-2\sigma^2\|x\|_2^2).
\end{aligned}$$

By the multinomial theorem the sum in the last expression equals

$$\begin{aligned}
\sum_{\nu\in\mathbb{N}_0^d} \prod_{j=1}^d \frac{(2\sigma^2 x_j^2)^{n_j}}{n_j!} &= \sum_{n=0}^\infty \frac{1}{n!} \sum_{|\nu|=n} \frac{n!}{n_1!\cdots n_d!} \left(2\sigma^2 x_1^2\right)^{n_1} \cdots \left(2\sigma^2 x_d^2\right)^{n_d} \\
&= \sum_{n=0}^\infty \frac{1}{n!} \left(2\sigma^2(x_1^2+\cdots+x_d^2)\right)^n = \exp(2\sigma^2\|x\|_2^2),
\end{aligned}$$

which implies $\|I_{\sigma,d} : H_\sigma([0,1]^d) \to C([0,1]^d)\| = 1$.

Next we introduce for all $N \in \mathbb{N}$ the orthogonal projections

$$\begin{aligned}
&P_N \quad \text{onto } \operatorname{span}\{e_\nu : |\nu| < N\}, \\
&Q_N \quad \text{onto } \operatorname{span}\{e_\nu : |\nu| \ge N\}
\end{aligned}$$

in $H_\sigma([0,1]^d)$. By the same arguments as above for $\|I_{\sigma,d}\|$ we obtain

$$\|I_{\sigma,d}Q_N\|^2 \le \sup_{x\in[0,1]^d} \sum_{n=N}^\infty \frac{1}{n!} \left(2\sigma^2\|x\|_2^2\right)^n \cdot \exp(-2\sigma^2\|x\|_2^2).$$

The sum in this formula is the tail of the Taylor series of the exponential function, and setting $t = 2\sigma^2\|x\|_2^2$ Taylor's theorem implies

$$\sum_{n=N}^\infty \frac{t^n}{n!} \le \frac{t^N}{N!} \exp(\xi) \quad \text{for some } \xi \in (0,t).$$

Combining this with the previous formula and using $N! \geq \left(\frac{N}{e}\right)^N$ we get

$$\|I_{\sigma,d}Q_N\| \leq \sup_{x\in[0,1]^d} \left(\frac{\left(2\sigma^2\|x\|_2^2\right)^N}{N!}\right)^{1/2} \leq \left(\frac{2\sigma^2 ed}{N}\right)^{N/2} =: \delta_N.$$

Given any sufficiently small real number $\varepsilon > 0$, there exists a natural number $N = N_\varepsilon \in \mathbb{N}$ with $\delta_N \leq \varepsilon < \delta_{N-1}$. By this choice we have

$$\log\frac{1}{\varepsilon} \sim \frac{N}{2}\log\left(\frac{N}{2e\sigma^2 d}\right) \sim \frac{N}{2}\log N \quad \text{and} \quad N \sim \frac{2\log\frac{1}{\varepsilon}}{\log\log\frac{1}{\varepsilon}}.$$

From $\|I_{\sigma,d}Q_N\| \leq \delta_N$ we get $\mathcal{N}(\delta_N, I_{\sigma,d}Q_N) = 1$, and since $\|P_N I_{\sigma,d}\| \leq \|I_{\sigma,d}\| \leq 1$ we can use (1)–(3) from Lemma 1 to bound the covering numbers of the operator $I_{\sigma,d}$ as follows.

$$\begin{aligned}\mathcal{N}(2\varepsilon, I_{\sigma,d}) &\leq \mathcal{N}(\varepsilon+\delta_N, I_{\sigma,d}P_N + I_{\sigma,d}Q_N)\\ &\leq \mathcal{N}(\varepsilon, I_{\sigma,d}P_N)\cdot\mathcal{N}(\delta_N, I_{\sigma,d}Q_N)\\ &= \mathcal{N}(\varepsilon, I_{\sigma,d}P_N) \leq \left(1+\frac{2}{\varepsilon}\right)^{\operatorname{rank}P_N}.\end{aligned}$$

By the trivial estimate $\operatorname{rank}P_N \leq N^d$ we finally get the desired upper bound

$$\log\mathcal{N}(2\varepsilon, I_{\sigma,d}) \leq N^d\log\left(1+\frac{2}{\varepsilon}\right) \asymp \frac{\left(\log\frac{1}{\varepsilon}\right)^{d+1}}{\left(\log\log\frac{1}{\varepsilon}\right)^d}. \tag{9}$$

Step 2. *Lower estimate for the L_2-norm.*

The idea is to apply the lower estimate (5) for covering numbers, which involves determinants. To this end we will relate the operator $I_{\sigma,d}$ to finite rank operators whose determinants can be computed explicitly. For the construction of such operators and estimates of their determinants we use tensor product techniques. We begin with the

Case $d = 1$. Recall that the functions

$$e_k(x) = \mu_k x^k \exp(-\sigma^2x^2) \quad (k \in \mathbb{N}_0)$$

form an ONB in $H_\sigma([0,1])$, where we have set

$$\mu_k := \sqrt{\frac{(2\sigma^2)^k}{k!}}.$$

For arbitrary $n \in \mathbb{N}$ we consider the operator

$$T_n : E_n \xrightarrow{J_n} H_\sigma([0,1]) \xrightarrow{I_{\sigma,1}} L_2([0,1]) \xrightarrow{M_\sigma} L_2([0,1]) \xrightarrow{P_n} F_n,$$

where

J_n is the embedding from $E_n := \operatorname{span}\{e_0, e_1, \ldots, e_{n-1}\}$ into $H_\sigma([0,1])$,
M_σ is the pointwise multiplication operator $M_\sigma f(x) = \exp(\sigma^2x^2)f(x)$,
P_n is the orthogonal projection onto $F_n :=$ range of $M_\sigma I_{\sigma,1}J_n$.

The operator $T_n : E_n \to F_n$ is uniquely determined by

$$T_n e_k(x) = \mu_k x^k \quad \text{for } k = 0, 1, \ldots, n-1.$$

The idea of working with the multiplication operator M_σ consists in the elimination of the "nasty" factor $\exp(-\sigma^2x^2)$. This simplifies the representing matrix $A_n = (a_{ij})_{i,j=0}^{n-1}$ of the operator $T_n^*T_n : E_n \to E_n$ with respect to the ONB $\{e_0, e_1, \ldots, e_{n-1}\}$ of E_n, so that we can *explicitly* compute its determinant. The entries of A_n are

$$a_{ij} = \langle T_n e_i, T_n e_j\rangle_{L_2} = \mu_i\mu_j\int_0^1 x^{i+j}\,dx = \frac{\mu_i\mu_j}{i+j+1}.$$

That means we have the factorization

$$A_n = D_n H_n D_n,$$

where

D_n is the $n \times n$-diagonal matrix with diagonal entries $\mu_0, \ldots, \mu_{n-1}$, and
H_n is the n-th section of the *Hilbert matrix*,

$$H_n = \left(\frac{1}{i+j+1}\right)_{i,j=0}^{n-1} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdot & \cdot & \cdot & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & & & & \cdot & \cdot \\ \frac{1}{3} & & & & \cdot & & \cdot \\ \cdot & & & \cdot & & & \cdot \\ \cdot & & \cdot & & & & \cdot \\ \cdot & \cdot & & & & & \cdot \\ \frac{1}{n} & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{1}{2n-1} \end{pmatrix}.$$

The well-known determinant formula (see e.g. [5, p. 306])

$$\det H_n = \frac{[1!\,2!\cdots(n-1)!]^4}{1!\,2!\cdots(2n-1)!}$$

implies

$$\det A_n = (\det D_n)^2 \det H_n = \frac{(2\sigma^2)^{\frac{n(n-1)}{2}}}{1!\,2!\cdots(n-1)!} \cdot \frac{[1!\,2!\cdots(n-1)!]^4}{1!\,2!\cdots(2n-1)!}.$$

Using elementary calculus one can easily check that

$$\log(1!\,2!\cdots(n-1)!) \sim \frac{n^2}{2}\log n \quad \text{as } n \to \infty.$$

(Remember that $\sim$ means *strong equivalence*.) This gives

$$\log(\det A_n) \sim \frac{n^2}{2}\log\left(2\sigma^2\right) + \frac{3}{2}n^2\log n - \frac{1}{2}(2n)^2\log(2n) \sim -\frac{n^2}{2}\log n.$$

Case $d > 1$. For the d-fold tensor product of A_n we have

$$\det(A_n^{\otimes d}) = (\det A_n)^{dn^{d-1}}. \tag{10}$$

This follows by induction from the formula

$$\det(A \otimes B) = (\det A)^m (\det B)^n$$

for the tensor product of an $n \times n$-matrix A and an $m \times m$-matrix B.

Alternatively we can argue as follows. If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the matrix A_n, then $\det A_n = \prod_{i=1}^n \lambda_i$. The eigenvalues of $A_n^{\otimes d}$ are all possible products $\lambda_{i_1} \cdots \lambda_{i_d}$ with $i_j \in \{1, \ldots, n\}$, whence

$$\det(A_n^{\otimes d}) = \prod_{i_1,\ldots,i_d=1}^{n} \lambda_{i_1} \cdots \lambda_{i_d}.$$

This product consists of dn^d factors of the form λ_i, where each of the n numbers λ_i appears equally often, i.e. dn^{d-1} times. This proves (10).

Now consider the d-fold tensor product $T_n^{\otimes d}$ of the operator T_n. Clearly the representing matrix of the operator $(T_n^{\otimes d})^* T_n^{\otimes d} : E_n^{\otimes d} \to E_n^{\otimes d}$ with respect to the ONB $\{e_{n_1} \otimes \cdots \otimes e_{n_d} : 0 \le n_j < n\}$ of $E_n^{\otimes d}$ is the tensor product $A_n^{\otimes d}$, whence we have for all $n \in \mathbb{N}$

$$\log\left(\det(A_n^{\otimes d})\right) = dn^{d-1}\log(\det A_n) \sim -\frac{d}{2}n^{d+1}\log n.$$

From the factorization

$$T_n^{\otimes d} : E_n^{\otimes d} \xrightarrow{J_n^{\otimes d}} H_\sigma([0,1]^d) \xrightarrow{I_{\sigma,d}} L_2([0,1]^d) \xrightarrow{M_\sigma^{\otimes d}} L_2([0,1]^d) \xrightarrow{P_n^{\otimes d}} F_n^{\otimes d}$$

and the norm estimates

$$\|M_\sigma^{\otimes d}\| \le \mathrm{e}^{\sigma^2 d} \quad \text{and} \quad \|J_n^{\otimes d}\| = \|P_n^{\otimes d}\| = 1,$$

taking also into account that $\dim E_n^{\otimes d} = \dim F_n^{\otimes d} = n^d$, we obtain from (5) the following bound for the covering numbers of $I_{\sigma,d}$,

$$\mathcal{N}(\mathrm{e}^{-\sigma^2 d}\varepsilon, I_{\sigma,d}) \ge \mathcal{N}(\varepsilon, T_n^{\otimes d}) \ge \sqrt{\det(A_n^{\otimes d})}\cdot\left(\frac{1}{\varepsilon}\right)^{n^d} \quad \text{for all } \varepsilon > 0.$$

Finally we optimize over n. For sufficiently small $\varepsilon > 0$ we choose

$$n = n_\varepsilon = \left[\frac{2}{d}\cdot\frac{\log\frac{1}{\varepsilon}}{\log\log\frac{1}{\varepsilon}}\right],$$

where $[t]$ denotes the greatest integer part of $t \in \mathbb{R}$. In particular this implies $\log n_\varepsilon \sim \log\log\frac{1}{\varepsilon}$, and putting all our estimates together we arrive at

$$\begin{aligned}\log\mathcal{N}(\mathrm{e}^{-\sigma^2 d}\varepsilon, I_{\sigma,d}) &\ge \frac{1}{2}\log\left(\det(A_{n_\varepsilon}^{\otimes d})\right) + n_\varepsilon^d \log\frac{1}{\varepsilon}\\ &\sim -\frac{d}{4}n_\varepsilon^{d+1}\log n_\varepsilon + n_\varepsilon^d\log\frac{1}{\varepsilon}\\ &\sim \frac{1}{2}\left(\frac{2}{d}\right)^d\frac{\left(\log\frac{1}{\varepsilon}\right)^{d+1}}{\left(\log\log\frac{1}{\varepsilon}\right)^d},\end{aligned}$$

which is equivalent to the desired lower bound. The proof is finished. □

Remark 4. Let us add some information on the constants that are hidden in the weak asymptotic equivalence (7). For simplicity of notation we set

$$\psi_d(\varepsilon) := \frac{\left(\log\frac{1}{\varepsilon}\right)^{d+1}}{\left(\log\log\frac{1}{\varepsilon}\right)^d}$$

for $d \in N$ and sufficiently small $\varepsilon > 0$. The functions ψ_d are slowly varying in the sense that for all constants $C > 0$ the relation

$$\lim_{\varepsilon\to 0}\frac{\psi_d(C\varepsilon)}{\psi_d(\varepsilon)} = 1$$

holds. In the proof of the upper bound for the covering numbers one can replace the rough estimate $\operatorname{rank} P_N \le N^d$ by the exact value

$$\operatorname{rank} P_N = \operatorname{card}\{\nu \in \mathbb{N}_0^d : |\nu| < N\} = \binom{N-1+d}{d}.$$

This can be shown by standard combinatorial arguments and gives the sharper upper bound

$$\operatorname{rank} P_N \le \left(\frac{e(N+d)}{d}\right)^d \sim \left(\frac{eN}{d}\right)^d.$$

From (9) and our choice of $N = N_\varepsilon \sim \frac{2\log\frac{1}{\varepsilon}}{\log\log\frac{1}{\varepsilon}}$ we obtain for the operator $I_{\sigma,d} : H_\sigma([0,1]^d) \to C([0,1]^d)$ and sufficiently small $\varepsilon > 0$

$$\log \mathcal{N}(\varepsilon, I_{\sigma,d}) \le \left(\frac{eN}{d}\right)^d \log\left(1+\frac{4}{\varepsilon}\right) \sim \left(\frac{2e}{d}\right)^d \psi_d(\varepsilon),$$

whence

$$\limsup_{\varepsilon\to 0} \frac{\log \mathcal{N}(\varepsilon, I_{\sigma,d})}{\psi_d(\varepsilon)} \le \left(\frac{2e}{d}\right)^d.$$

From the lower bound for the covering numbers that was shown in the proof we obtain

$$\liminf_{\varepsilon\to 0} \frac{\log \mathcal{N}(\varepsilon, I_{\sigma,d})}{\psi_d(\varepsilon)} \ge \frac{1}{2}\left(\frac{2}{d}\right)^d.$$

The same remains true for $I_{\sigma,d} : H_\sigma([0,1]^d) \to L_p([0,1]^d)$, $2 \le p < \infty$. It is an open question whether the limit $\lim_{\varepsilon\to 0} \frac{\mathcal{N}(\varepsilon, I_{\sigma,d})}{\psi_d(\varepsilon)}$ exists.

The influence of the parameter σ and the dimension d on the covering numbers of Gaussian RKHSs has been studied in several papers, e.g. in [23,19,20,22]. The corresponding results are weaker in ε than ours, but they give concrete constants depending on σ and d, which is important in learning theory. In particular we mention Theorem 2.1. in [19] which describes a suitable trade-off between the influence of ε and σ.

Remark 5. We are working in this note with *real spaces* only, but the real Gaussian RKHSs have natural complex extensions whose elements are analytic functions, see e.g. [18]. Therefore it is not surprising that our main entropy estimate (7) in Theorem 3 looks similar to the results by Kolmogorov and Tihomirov [10, Theorem XX] on metric entropy in classes of analytic functions on $\mathbb{C}^d$. The asymptotic entropy behaviour is the same in both cases, the proofs, however, are completely different.

4. An application to small deviations of Gaussian processes

Let $X = (X(t))_{t\in T}$ be a stochastic process indexed by a set T, typical examples are $T = [0,1]$, $(0,\infty)$, $\mathbb{R}$ or $[0,1]^d$. The small deviation problem (also called small ball or small value problem) for the process X consists in finding the asymptotic behaviour of the function

$$\varphi(\varepsilon) := -\log \mathbb{P}(\|X\| \le \varepsilon) \quad \text{as } \varepsilon \to 0,$$

where $\|.\|$ is for example the norm in some $L_p(T)$ or in $C(T)$. Small deviation probabilities play a fundamental role in many problems in probability theory such as the law of the iterated logarithm of Chung type, strong limit laws in statistics, quantization and approximation of stochastic processes, see e.g. the survey [15], the lecture notes [13] and the literature compilation [4]. For Gaussian processes many results are available in the literature, mostly showing a polynomial behaviour of the small deviation function $\varphi(\varepsilon)$. However, for smooth processes much less is known, see [3,2].

The RKHS H of a centered Gaussian process $X = (X(t))_{t\in T}$ is given by the positive definite kernel

$$K(s,t) := \mathbb{E}X(s)X(t), \quad s,t \in T,$$

which describes the covariance structure of X. If X attains values in a Banach space E, then the RKHS H embeds compactly into E. Kuelbs and Li [12] discovered a close, but quite complicated connection between the small deviation function $\varphi(\varepsilon)$ of the process and the metric entropy $\mathcal{H}(\varepsilon)$ of the unit ball

of the RKHS, considered as a subset of E. However, under certain regularity conditions on $\varphi(\varepsilon)$ and/or $\mathcal{H}(\varepsilon)$ this connection is more direct. For example, Aurzada et al. [3, Corollary 2.6] showed that for any $\alpha > 0$ and $\beta \in \mathbb{R}$

$$\varphi(\varepsilon) \asymp \left(\log \frac{1}{\varepsilon}\right)^{\alpha} \left(\log\log \frac{1}{\varepsilon}\right)^{\beta} \Longleftrightarrow \mathcal{H}(\varepsilon) \asymp \left(\log \frac{1}{\varepsilon}\right)^{\alpha} \left(\log\log \frac{1}{\varepsilon}\right)^{\beta}.$$

This enables us to translate our entropy estimates in Theorem 3 directly into the following small deviation result.

Theorem 6. *Let $\sigma > 0, d \in \mathbb{N}$ and $X_{\sigma,d} = \left(X_{\sigma,d}(t)\right)_{t\in[0,1]^d}$ be the centered Gaussian process with covariance structure*

$$K(s,t) := \mathbb{E} X_{\sigma,d}(s) X_{\sigma,d}(t) = \exp\left(-\sigma^2 \|s-t\|_2^2\right).$$

Then the small deviation probabilities of $X_{\sigma,d}(t)$ with respect to the sup*-norm behave asymptotically, as $\varepsilon \to 0$, like*

$$-\log \mathbb{P}\left(\sup_{t\in[0,1]^d} |X_{\sigma,d}(t)| \le \varepsilon\right) \asymp \frac{\left(\log \frac{1}{\varepsilon}\right)^{d+1}}{\left(\log\log \frac{1}{\varepsilon}\right)^{d}}.$$

The same estimates hold for small deviations with respect to any L_p-norm with $2 \le p < \infty$.

Remark 7. The one-dimensional case ($d = 1$) of our sup-norm estimate

$$-\log \mathbb{P}\left(\sup_{t\in[0,1]} |X_{\sigma,1}(t)| \le \varepsilon\right) \asymp \frac{\left(\log \frac{1}{\varepsilon}\right)^{2}}{\left(\log\log \frac{1}{\varepsilon}\right)}$$

coincides with a result in [3], see Theorem 1.1, case $\nu = 2$.

For other small deviation results for smooth Gaussian fields we refer to [9, Proposition 4.2]. In this paper the authors consider only deviations with respect to the L_2-norm, but in this case they even obtain exact constants.

Remark 8. Concerning *strong equivalence*$\sim$ of the small deviation probabilities our methods give no information. This is a challenging open problem which requires new techniques for its solution.

Acknowledgments

I thank M. Lifshits for discussions during the Seminar "Algorithms and Complexity for Continuous Problems" (Schloss Dagstuhl, September 2009) on the metric entropy results by Kolmogorov and Tihomirov [10] for classes of analytic functions, and Ding-Xuan Zhou for pointing out to me the papers [24,20].

Moreover, I am very grateful to Wenbo Li for drawing my attention (while I was visiting the University of Delaware in March 2010) to possible applications to small deviations of Gaussian processes. This was the motivation to add the final Section 4.

Finally I thank two anonymous referees for their careful reading of the manuscript and several useful hints.

The author's work was supported in part by Ministerio de Ciencia e Innovación (Spain), MTM2010-15814.

References

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (3) (1950) 337–404.
[2] F. Aurzada, F. Gao, T. Kühn, W.V. Li, Q.-M. Shao, Small deviations for a family of smooth Gaussian processes, 2010. Available at: http://arxiv.org/abs/arXiv:1009.5580.
[3] F. Aurzada, I.A. Ibragimov, M.A. Lifshits, J.H. van Zanten, Small deviations of smooth stationary Gaussian processes, Theory Probab. Appl. 53 (2009) 697–707.

[4] Bibliography compilation on small deviation probabilities. Available at: http://www.proba.jussieu.fr/pageperso/smalldev/biblio.html.
[5] M.-D. Choi, Tricks or treats with the Hilbert matrix, Amer. Math. Monthly 90 (5) (1983) 301–312.
[6] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. (NS) 39 (1) (2002) 1–49.
[7] F. Cucker, D.-X. Zhou, Learning Theory: An Approximation Theory Viewpoint, in: Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2007.
[8] D.E. Edmunds, H. Triebel, Function Spaces, Entropy Numbers, Differential Operators, in: Cambridge Tracts in Mathematics, vol. 120, Cambridge University Press, Cambridge, 1996.
[9] A.I. Karol', A.I. Nazarov, Small ball probabilities for smooth Gaussian fields and tensor products of compact operators, Preprint, 2010. Available at: http://arxiv.org/abs/arXiv:1009.4412.
[10] A.N. Kolmogorov, V.M. Tihomirov, ε-entropy and ε-capacity of sets in function spaces, Uspekhi Mat. Nauk 14 (2 (86)) (1959) 3–86 (in Russian); English translation: Amer. Math. Soc. Transl. (2) 17 (1961) 277–364.
[11] H. König, Eigenvalue Distribution of Compact Operators, in: Operator Theory: Advances and Applications, vol. 16, Birkhäuser Verlag, Basel, 1986.
[12] J. Kuelbs, W.V. Li, Metric entropy and the small ball problem for Gaussian measures, J. Funct. Anal. 116 (1993) 133–157.
[13] W.V. Li, Small value probabilities: techniques and applications, 2010. Available at: http://www.math.udel.edu/wli/svp.html.
[14] W.V. Li, W. Linde, Approximation, metric entropy and small ball estimates for Gaussian measures, Ann. Probab. 27 (3) (1999) 1556–1578.
[15] W.V. Li, Q.-M. Shao, Gaussian processes: inequalities, small ball probabilities and applications, in: Stochastic Processes: Theory and Methods, in: Handbook of Statist., vol. 19, 2001, pp. 533–597.
[16] A. Pietsch, Eigenvalues and s-Numbers, in: Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig, and Cambridge Studies in Advanced Mathematics, vol. 13, Cambridge University Press, Cambridge, 1987.
[17] I. Steinwart, A. Christmann, Support Vector Machines, Springer, New York, 2008.
[18] I. Steinwart, D. Hush, C. Scovel, An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels, IEEE Trans. Inform. Theory 52 (10) (2006) 4635–4643.
[19] I. Steinwart, C. Scovel, Fast rates for support vector machines using Gaussian kernels, Ann. Statist. 35 (2) (2007) 575–607.
[20] H.-W. Sun, D.-X. Zhou, Reproducing kernel Hilbert spaces associated with analytic translation-invariant Mercer kernels, J. Fourier Anal. Appl. 14 (1) (2008) 89–101.
[21] R.C. Williamson, A.J. Smola, B. Schölkopf, Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators, IEEE Trans. Inform. Theory 47 (6) (2001) 2516–2532.
[22] D.-H. Xiang, D.-X. Zhou, Classification with Gaussians and convex loss, J. Mach. Learn. Res. 10 (2009) 1447–1468.
[23] D.-X. Zhou, The covering number in learning theory, J. Complexity 18 (3) (2002) 739–767.
[24] D.-X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (7) (2003) 1743–1752.