

# A Closer Look at Covering Number Bounds for Gaussian Kernels

Ingo Steinwart and Simon Fischer

July 23, 2020

Institute for Stochastics and Applications

Faculty 8: Mathematics and Physics

University of Stuttgart

70569 Stuttgart Germany

{ingo.steinwart, simon.fischer}@mathematik.uni-stuttgart.de

## Abstract

We establish some new bounds on the log-covering numbers of (anisotropic) Gaussian reproducing kernel Hilbert spaces. Unlike previous results in this direction we focus on small explicit constants and their dependency on crucial parameters such as the kernel bandwidth and the size and dimension of the underlying space.

**Keywords** Gaussian kernels, Covering numbers

## 1 Introduction

Gaussian kernels and their reproducing kernel Hilbert spaces (RKHSs) play a central role for kernel-based learning algorithms such as support vector machines (SVMs), see e.g. [9, 2, 11], and Gaussian processes for machine learning, see e.g. [8, 5]. For the analysis of such learning algorithms one usually needs to bound both the *approximation error*, which quantitatively describes how well the considered RKHS approximates certain classes of smooth functions, and the *estimation error*, which bounds the uncertainty caused by the statistical nature of the observations the algorithm learns from. Moreover, the estimation error is typically analyzed with the help of bounds on certain entropy- or covering numbers of the involved RKHSs, and in the case of Gaussian RKHSs these bounds crucially depend on quantities such as the considered kernel width. The major focus of this work is to analyze this dependence. To be more precise, recall that the (isotropic) Gaussian kernels are given by

$$k_\sigma(x, x') := \exp(-\sigma^2 \|x - x'\|_{\ell_d^2}^2), \quad x, x' \in X, \quad (1)$$

where  $X$  is a subset of  $\mathbb{R}^d$ ,  $\sigma > 0$  is the so-called kernel width, and  $\|\cdot\|_{\ell_2^d}$  denotes the Euclidean norm on  $\mathbb{R}^d$ . Moreover, we write  $H_\sigma(X)$  for the corresponding RKHS, see [11, Chapter 4] for details about Gaussian RKHSs as well as for a general introduction to RKHSs.

In order to underpin the importance of log-covering number bounds with explicit and well-understood constants we will briefly sketch the analysis of SVMs using the least squares loss and the Gaussian kernel in the following. To this end, let us recall that the covering numbers of a bounded subset  $A \subseteq E$  of some Banach space  $E$  are defined by

$$\mathcal{N}(A, \varepsilon) := \min \left\{ n \in \mathbb{N} : \exists x_1, \dots, x_n \in E : A \subseteq \bigcup_{i=1}^n x_i + \varepsilon B_E \right\}, \quad \varepsilon > 0,$$

where  $B_E$  denotes the closed unit ball of  $E$ . Moreover, for a bounded linear operator  $T : E \rightarrow F$  between two Banach spaces  $E$  and  $F$ , the log-covering numbers are  $\mathcal{H}(T, \varepsilon) := \log(\mathcal{N}(TB_E, \varepsilon))$ . Now, if  $X \subseteq \mathbb{R}^d$  is bounded it is well-known that for all  $\sigma \geq 1$  and  $p \in (0, 1)$  we have a constant  $K_{X, \sigma, p}$  such that

$$\mathcal{H}(\text{Id} : H_\sigma(X) \rightarrow \ell_\infty(X), \varepsilon) \leq K_{X, \sigma, p} \cdot \varepsilon^{-p}, \quad \varepsilon \in (0, 1], \quad (2)$$

where  $\text{Id} : H_\sigma(X) \rightarrow \ell_\infty(X)$  denotes the canonical embedding of  $H_\sigma(X)$  into the space  $\ell_\infty(X)$  of bounded functions  $f : X \rightarrow \mathbb{R}$  equipped with the usual supremum norm  $\|\cdot\|_\infty$ , see e.g. [11, Theorem 6.27 and Exercise 6.8]. However, the dependency of the constant  $K_{X, \sigma, p}$  on  $X$ ,  $\sigma$ , and  $p$  is far from being well-understood.

Let us now briefly describe how this dependency influences the learning performance guarantees of SVMs using the least squares loss and a Gaussian kernel  $k_\sigma$ . To this recall that for a dataset  $D = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times [-1, 1])^n$  of length  $n$  and a regularization parameter  $\lambda > 0$ , such an SVM produces a decision function  $f_{D, \lambda, \sigma} : X \rightarrow [-1, 1]$  that minimizes some regularized empirical error quantity over  $H_\sigma(X)$ , in order to recover the true but unknown regression function  $f^* : X \rightarrow [-1, 1]$ . In this scenario [11, Theorem 7.23] in combination with (2) gives, for every  $f_0 \in H_\sigma(X)$ , the following over all error bound

$$\|f_{D, \lambda, \sigma} - f^*\|_{L_2(\nu)}^2 \leq 9(\lambda \|f_0\|_{H_\sigma(X)}^2 + \|f_0 - f^*\|_{L_2(\nu)}^2) + C_p \cdot \frac{K_{X, \sigma, p}}{\lambda^{p/2} n} + \tilde{\varepsilon}(n, \lambda, \tau, \sigma, p, f_0), \quad (3)$$

which holds true with probability not less than  $1 - 3e^{-\tau}$ . Here,  $C_p$  is a constant whose dependency on  $p$  is explicitly given, and  $\tilde{\varepsilon}(n, \lambda, \tau, \sigma, p, f_0)$  is an additional error term, that for common choices of  $f_0$ ,  $\lambda$ ,  $p$ ,  $\tau$ , and  $\sigma$  is dominated by the term

$$\varepsilon(n, \lambda, \sigma, p) := C_p \cdot \frac{K_{X, \sigma, p}}{\lambda^{p/2} n},$$

which in the following we call *estimation error*. Together with  $\tilde{\varepsilon}(n, \lambda, \tau, \sigma, p, f_0)$ , the estimation error bounds the error caused by statistical fluctuations. In contrast, the first error term in

(3), which does not depend on the sample size  $n$ , refers to the *approximation error*.

It is well-known that the Gaussian RKHS  $H_\sigma(X)$  only contains  $C^\infty$ -functions and that  $H_\sigma(X)$  is dense in  $L_p(\nu)$  for all  $p \in [1, \infty)$  and all finite measures  $\nu$  on  $X$ . Moreover, if  $X$  is compact, then  $H_\sigma(X)$  is also dense in  $C(X)$ . Again, we refer to [11, Chapter 4] for details. Now recall that these denseness results guarantee, for example, that the *minimal approximation error*

$$A(\lambda, \sigma, f^*) := \inf \left\{ \lambda \|f_0\|_{H_\sigma(X)}^2 + \|f_0 - f^*\|_{L_2(\nu)}^2 : f_0 \in H_\sigma(X) \right\}$$

satisfies  $A(\lambda, \sigma, f^*) \rightarrow 0$  for  $\lambda \rightarrow 0$  and *fixed*  $\sigma > 0$  and  $f^* \in L_2(\nu)$ . For  $n \rightarrow \infty$  and  $\lambda \rightarrow 0$  with  $\lambda^{p/2}n \rightarrow \infty$  this shows that (3) vanishes, i.e. SVMs using the least squares loss and a *fixed* Gaussian kernel can learn in a purely asymptotic sense, see e.g. [11, Chapters 5, 6, and 9] for details. However, a more detailed analysis that includes convergence rates for the learning process, requires convergence rates for the approximation error and especially for  $A(\lambda, \sigma, f^*) \rightarrow 0$ . Unfortunately, it has been shown in [10] that for fixed  $\sigma > 0$  any polynomial rate even for the minimal approximation error  $A(\lambda, \sigma, f^*) \rightarrow 0$  is impossible if  $f \notin C^\infty$ , and the latter is an unacceptable restriction from a learning theoretical point of view.

To address this issue and to be better aligned with empirical knowledge that strongly suggest to vary the width  $\sigma$  with the data set, one usually investigates the learning behavior in cases in which we have  $\lambda \rightarrow 0$  and  $\sigma \rightarrow \infty$  simultaneously. For example, [3] shows that for specific combinations of rates for  $\lambda \rightarrow 0$  and  $\sigma \rightarrow \infty$  the approximation error converges to 0 with a polynomial speed whenever  $f^*$  is contained in some Besov space.

While this approach solves the issues regarding the approximation error, it simultaneously makes the analysis of the estimation error  $\epsilon(n, \lambda, \sigma, p)$  more complicated. To be more precise, for fixed  $\sigma$  and  $p$  the dependence of  $K_{X, \sigma, p}$  on these parameters have no influence on the learning rate, however, if we consider  $\sigma \rightarrow \infty$  the behavior of  $K_{X, \sigma, p}$  plays a crucial role for the estimation error. Since it has been recently observed in [4] that the learning rates can be further improved, if we additionally let  $p = p_n \rightarrow 0$  sufficiently slowly, also the dependence of  $K_{X, \sigma, p}$  on  $p$  is of interest from a learning theoretical point of view. Moreover, the guarantees on the learning performance obviously become better, if, in addition,  $K_{X, \sigma, p}$  only depends on *small* universal constants. Therefore, the goal of this work is to derive bounds on  $\mathcal{H}(\text{Id} : H_\sigma(X) \rightarrow \ell_\infty(X), \varepsilon)$  that do not only have a desirable behavior for  $\varepsilon \rightarrow 0$ , but for which we can also control the behavior of the corresponding constants in  $\sigma$ ,  $X$ ,  $d$ , and if applicable, in  $p$ .

To this end, we first refine the analysis of [7] by carefully controlling the arising constants. It turns out that the final constants have both small absolute values and a reasonable behavior in the dimension  $d$ . Unfortunately, however, their behavior for  $\sigma \rightarrow \infty$  is far from being optimal. For this reason, we present another result that relates the log-covering numbers of  $\text{Id} : H_\sigma(X) \rightarrow \ell_\infty(X)$  to the log-covering numbers of  $\text{Id} : H_1(B_2^d) \rightarrow \ell_\infty(B_2^d)$ , where  $B_2^d \subseteq \mathbb{R}^d$  denotes the closed Euclidean unit ball, and to the covering numbers of the underlying space

$X$ . As a consequence, we do not only obtain a much better behavior for  $\sigma \rightarrow \infty$ , but also log-covering number bounds for *anisotropic* Gaussian kernels, which are defined by

$$k_\sigma(x, x') := \exp(-\|D_\sigma x - D_\sigma x'\|_{\ell_2^d}^2), \quad x, x' \in X, \quad (4)$$

where  $D_\sigma(x_1, \dots, x_d) := (\sigma_1 x_1, \dots, \sigma_d x_d)$ . Note that these kernels are an example of so-called automatic relevance determination (ARD) kernels, which are particularly popular in the Gaussian processes for machine learning context, see e.g. [8, Chapter 5].

The rest of this work is organized as follows: In the next section we present our main results, discuss their consequences, and compare them to results previously obtained in the literature such as [13, 14, 7]. All proofs can be found in Section 3.

## 2 Main Results

This section contains all main results of this work: In the first subsection we derive bounds on the log-covering numbers of the embedding  $\text{Id} : H_\sigma(B_2^d) \rightarrow \ell_\infty(B_2^d)$  of the isotropic Gaussian RKHS defined in (1). In the second subsection we then show how to generalize these bounds to anisotropic Gaussian kernels (4) on general bounded sets  $X \subseteq \mathbb{R}^d$ .

### 2.1 Isotropic Gaussian Kernels

Before we present the results of this subsection, let us introduce some notation: if two functions  $f, g : (0, \infty) \rightarrow (0, \infty)$  satisfy  $\lim_{t \rightarrow \infty} f(t)/g(t) = 1$  we write  $f(t) \sim g(t)$  for  $t \rightarrow \infty$ . Moreover, recall that, for  $k \in \mathbb{N}$  and  $t > 0$ , the *generalized binomial coefficient* is defined by

$$\binom{t}{k} := \frac{1}{k!} \prod_{i=1}^k (t - k + i).$$

Note that for  $t \in \mathbb{N}$  this definition coincides with the classical definition of binomial coefficients. In the following, generalized binomial coefficients mainly appear in the form

$$\binom{t+d}{d} = \frac{1}{d!} \prod_{i=1}^d (t+i) \quad (5)$$

where  $d \geq 1$  is an integer and  $t > 0$ . Then the functions  $t \mapsto \binom{t+d}{d}$  and  $d \mapsto \binom{t+d}{d}$  are increasing and the functions  $t \mapsto \binom{t+d}{d} \cdot t^{-d}$  and  $d \mapsto \binom{t+d}{d} \cdot d^{-t}$  are decreasing with

$$\binom{t+d}{d} \sim \frac{t^d}{d!} \quad \text{for } t \rightarrow \infty \quad \text{and} \quad \binom{t+d}{d} \sim \frac{d^t}{\Gamma(t+1)} \quad \text{for } d \rightarrow \infty,$$

where  $\Gamma$  denotes the Gamma function. See Lemma 3.6 for the non-obvious assertions. With these preparations our first result reads as follows.

**2.1 Theorem** For all  $d \geq 1$ , all  $\sigma > 0$ , and all  $0 < \varepsilon \leq 1$  we have

$$\mathcal{H}(\text{Id} : H_\sigma(B_2^d) \rightarrow \ell_\infty(B_2^d), \varepsilon) \leq \binom{2e(1 + \sigma^2) + d}{d} \cdot e^{-d} \cdot \frac{\log^{d+1}(4/\varepsilon)}{\log \log^d(4/\varepsilon)}.$$

Note that Theorem 2.1 recovers the asymptotic behavior of  $\varepsilon \mapsto \mathcal{H}(\text{Id} : H_\sigma(B_2^d) \rightarrow \ell_\infty(B_2^d), \varepsilon)$  found by Kühn in [7], which in turn improved the earlier results in [13, 14]. By presenting a corresponding lower bound on the log-covering numbers, [7] further shows that this behavior in  $\varepsilon$  is optimal. Unlike the upper bound in [7], however, Theorem 2.1 also provides an upper bound for the behavior in  $\sigma$  and  $d$  that is expressed by the constant

$$K_{d,\sigma} := \binom{2e(1 + \sigma^2) + d}{d} \cdot e^{-d}.$$

To better understand the behavior of this constant in  $d$ , let us first consider the case  $\sigma = 1$ . Since  $d \mapsto \binom{4e+d}{d} \cdot d^{-4e}$  is decreasing as mentioned above we then find

$$C_d := \frac{K_{d,1}}{d^{4e}e^{-d}} \leq \binom{4e+1}{1} \cdot 1^{-4e} = 4e+1 \approx 11.8731$$

for all  $d \geq 1$ . Moreover, some numerical calculations show that for  $d = 1$  we have  $K_{1,1} = 4 + 1/e \approx 4.3679$ , while for  $d \geq 2$  we have  $C_d \leq C_2 \approx 0.0407 \leq 0.05$ , and hence we obtain

$$K_{d,1} \leq 0.05 \cdot d^{4e} e^{-d}, \quad d \geq 2. \quad (6)$$

In this respect note that [13, Proposition 1] found a constant behaving like  $d^{d+1}$  for a  $\log^{d+1}(1/\varepsilon)$ -type bound on the log-covering numbers. Unfortunately, this result is not directly comparable to (6) since [13] considered the set  $X = [0, 1]^d$ . Since  $[0, 1]^d \subseteq 1/2 + \sqrt{d}/2 \cdot B_2^d$  combining (6) with the later established (12) for  $r = \sqrt{d}/2$  and  $\sigma = 1$ , however, we obtain a constant not exceeding  $0.05 \cdot d^{4e} (2e/3)^{-d} d^{d/2}$  for  $X = [0, 1]^d$  and  $d \geq 2$ . In other words, our analysis does improve the above mentioned results of [13] in both  $\varepsilon$  and  $d$ .

Furthermore, this inequality correctly describes the asymptotic behavior of  $K_{d,1}$  for  $d \rightarrow \infty$ , since our considerations at the beginning of this section show

$$\lim_{d \rightarrow \infty} C_d = \lim_{d \rightarrow \infty} \binom{4e+d}{d} \cdot d^{-4e} = \frac{1}{\Gamma(4e+1)} \approx 3.4130 \cdot 10^{-8}.$$

Finally, some additional numerical calculations give  $K_{d,1} \leq 30$  for all  $d \geq 1$ , and the maximal value of  $K_{d,1}$  is attained at  $d = 6$ .

Let us now consider the behavior of  $K_{d,\sigma}$  in  $\sigma$  for a fixed  $d \geq 1$ . To this end, we first observe that  $K_{d,\sigma}$  is increasing in  $\sigma$ , and hence we have  $K_{d,\sigma} > K_{d,0} = \binom{2e+d}{d} e^{-d} > 0$  for all  $\sigma > 0$ .

Moreover, the representation in (5) directly gives

$$\frac{2^d}{d!}(1 + \sigma^2)^d \leq K_{d,\sigma} \leq \frac{4^d}{d!}(1 + \sigma^2)^d \quad (7)$$

for all  $\sigma > 0$  satisfying  $2e(1 + \sigma^2) \geq d$ . Consequently the constant  $K_{d,\sigma}$  grows like  $\sigma^{2d}$  for  $\sigma \rightarrow \infty$ , compared to the  $\sigma^{2d+2}$ -behavior of the already discussed result in [13]. Below in Section 2.2 we will see that we can find another constant for the estimate of Theorem 2.1 that only grows like  $\sigma^d$  for  $\sigma \rightarrow \infty$ .

Our next goal is to show that the size of the constant in Theorem 2.1 is significantly influenced by the choice of the considered range of  $\varepsilon$ . More precisely, Theorem 2.1 considers the maximal range  $0 < \varepsilon \leq 1$ , since we have  $\|\text{Id} : H_\sigma(B_2^d) \rightarrow \ell_\infty(B_2^d)\| = 1$ , and thus we find

$$\mathcal{H}(\text{Id} : H_\sigma(B_2^d) \rightarrow \ell_\infty(B_2^d), \varepsilon) = 0, \quad \varepsilon \geq 1.$$

Our next theorem shows that by considering a smaller range for  $\varepsilon$ , we can substantially decrease the constant appearing in the estimate. For its formulation, we recall that Lambert's  $W$ -function is the inverse of  $t \mapsto te^t$ . Note that on  $(-1/e, 0)$  the inverse is multi-valued and throughout this work we use the upper branch  $W : [-1/e, \infty) \rightarrow [-1, \infty)$ , which is often denoted by  $W_0$  in the literature. Finally, recall that  $W$  is increasing and  $W(t) \sim \log(t)$  for  $t \rightarrow \infty$ . Now our second result reads as follows.

**2.2 Theorem** *For all  $d \geq 1$ , all  $\sigma > 0$ , and  $0 < \varepsilon_0 \leq 4 \exp(-e^{1+\sigma^{-2}})$  we define  $y_0 := \log(4/\varepsilon_0)$ ,  $x_0 := 2y_0/W(\frac{y_0}{e\sigma^2})$ , and*

$$K_{d,\sigma,\varepsilon_0} := \binom{x_0 + d}{d} \cdot \left(\frac{\log(y_0)}{y_0}\right)^d.$$

*Then for all  $0 < \varepsilon \leq \varepsilon_0$  we have*

$$\mathcal{H}(\text{Id} : H_\sigma(B_2^d) \rightarrow \ell_\infty(B_2^d), \varepsilon) \leq K_{d,\sigma,\varepsilon_0} \cdot \frac{\log^{d+1}(4/\varepsilon)}{\log \log^d(4/\varepsilon)}.$$

To appreciate Theorem 2.2 we note that for  $\varepsilon_0 \rightarrow 0$  we have  $y_0 \rightarrow \infty$  and  $x_0 \rightarrow \infty$ . Since  $W(t) \sim \log(t)$  and  $\binom{t+d}{d} \sim t^d/d!$  for  $t \rightarrow \infty$  we then find

$$\lim_{\varepsilon_0 \rightarrow 0^+} K_{d,\sigma,\varepsilon_0} = \lim_{\varepsilon_0 \rightarrow 0^+} \binom{x_0 + d}{d} \cdot x_0^{-d} \cdot \left(\frac{2 \log(y_0)}{W(\frac{y_0}{e\sigma^2})}\right)^d = \frac{2^d}{d!}.$$

This sharpens the result of [7, Remark 4] by a factor of approximately  $\sqrt{2\pi d}$ . Finally note that for  $\sigma = 1$  and  $\varepsilon_0 := 4 \exp(-e^2) \approx 0.0025$  we have  $y_0 = e^2$  and  $x_0 = 2e^2$ . Hence we find

$$K_{d,1,\varepsilon_0} \leq 16 \cdot d^{2e^2} \cdot (2/e^2)^d.$$

In this respect we like to mention that [13] established the constant  $4^d(6d+2)$  for  $0 < \varepsilon \leq \exp(-90d^2 - 11d - 3)$ , again however, for a  $\log^{d+1}(1/\varepsilon)$ -type bound on  $X = [0, 1]^d$ . To compare this result of [13] with our Theorem 2.2 we use  $[0, 1]^d \subseteq 1/2 + \sqrt{d}/2 \cdot B_2^d$  in combination with the later established Theorem 2.4 and (10) for  $r = \sqrt{d}/2$  and  $\sigma = 1$  as well as Lemma 3.7 for  $C := 1/\sqrt{360e}$  to obtain a constant not exceeding

$$(2\pi)^{-1/2} \cdot 16.84^d \cdot d^{-d/2-1/2}$$

for the range  $0 < \varepsilon \leq \varepsilon_0 := 4 \exp(-3d\sqrt{10e} \log(3d\sqrt{10/e}))$  on  $X = [0, 1]^d$ . This improves the result from [13] in both, the  $\varepsilon$  range and the constant for all  $d \geq 1$ .

For some applications, see e.g. [12, 4], it is sufficient and more convenient to work with a weaker bound in  $\varepsilon$  such as the one in (2). For this reason, the following theorem establishes an upper bound of the form (2) with an explicit constant.

**2.3 Theorem** *For  $d \geq 1$  and  $\sigma > 0$  we define*

$$t_0 := \frac{2(d+1) \cdot 4^{\frac{p}{d+1}}}{ep \cdot W\left(\frac{d+1}{p\sigma^2}\right)} \exp\left(\frac{1}{W\left(\frac{d+1}{p\sigma^2}\right)}\right).$$

*Then for all  $d \geq 1$ ,  $\sigma > 0$ ,  $p > 0$ , and all  $0 < \varepsilon \leq 1$  we have*

$$\mathcal{H}(\text{Id} : H_\sigma(B_2^d) \rightarrow \ell_\infty(B_2^d), \varepsilon) \leq \binom{t_0 + d}{d} \cdot \frac{d+1}{ep} \cdot 4^{\frac{p}{d+1}} \cdot \varepsilon^{-p}.$$

To better understand the constant appearing in Theorem 2.3, we denote it by

$$K_{d,\sigma,p} := \binom{t_0 + d}{d} \cdot \frac{d+1}{ep} \cdot 4^{\frac{p}{d+1}}. \quad (8)$$

For fixed  $\sigma, p > 0$ , Lemma 3.8 then shows that  $d \mapsto K_{d,\sigma,p}$  grows more slowly than any exponential function, i.e. for all  $a > 0$  we have  $K_{d,\sigma,p} e^{-ad} \rightarrow 0$  for  $d \rightarrow \infty$ . To be more precise, Lemma 3.8 provides constants  $c_{\sigma,p} > 0$  and  $C_{\sigma,p} > 0$  independent of  $d$  such that

$$K_{d,\sigma,p} \leq C_{\sigma,p} \sqrt{d \log(d)} \cdot \exp\left(c_{\sigma,p} \cdot d \cdot \frac{\log \log(d)}{\log(d)}\right), \quad d \geq 1.$$

Moreover, if we restrict our considerations to  $\sigma := 1$  and also fix a  $d \geq 1$  and a  $0 < p_0 \leq 1/e$ , then Lemma 3.9 shows that

$$K_{d,1,p} \leq 1/2 \cdot C_0^d \cdot \sqrt{d} \cdot \frac{(1/p)^{d+1}}{\log^d(1/p)}, \quad 0 < p \leq p_0,$$

where the constant  $C_0$  is given by

$$C_0 := ep_0 \cdot \log(1/p_0) + (2 + 1/e)2^{1+p_0} \exp\left(\frac{1}{W(2/p_0)}\right). \quad (9)$$

In particular,  $C_0$  only depends on  $p_0$ , and for  $p_0 := 1/e$  we find  $C_0 \approx 13.6481$ . In addition,  $C_0$  converges to  $4 + 2/e \approx 4.7358$  for  $p_0 \rightarrow 0$ . Finally, we note that the constant appearing in Theorem 2.3 can again be substantially improved if we restrict our consideration to a smaller range  $0 < \varepsilon \leq \varepsilon_0$ .

## 2.2 Anisotropic Gaussian Kernels

The goal of this subsection is to analyze how the constants in the log-covering number bounds depend on the kernel width  $\sigma$  and the size of the input space  $X$ . To this end, our next theorem reduces the problem of bounding the log-covering numbers of  $\text{Id} : H_\sigma(X) \rightarrow \ell_\infty(X)$  of *anisotropic* Gaussian RKHS to the estimation of the log-covering numbers of the embedding  $\text{Id} : H_1(B_2^d) \rightarrow \ell_\infty(B_2^d)$  of the *isotropic* Gaussian RKHS with width  $\sigma = 1$ .

**2.4 Theorem** *For all bounded subsets  $X \subseteq \mathbb{R}^d$ , all  $\sigma = (\sigma_1, \dots, \sigma_d) \in (0, \infty)^d$ , and all  $0 < \varepsilon \leq 1$  we have*

$$\mathcal{H}(\text{Id} : H_\sigma(X) \rightarrow \ell_\infty(X), \varepsilon) \leq \mathcal{N}(D_\sigma X, 1) \cdot \mathcal{H}(\text{Id} : H_1(B_2^d) \rightarrow \ell_\infty(B_2^d), \varepsilon),$$

where the covering numbers  $\mathcal{N}(D_\sigma X, 1)$  of  $D_\sigma X \subseteq \mathbb{R}^d$  are with respect to the Euclidean norm.

Before we proceed we like to remark that Theorem 2.4 actually holds for general bounded and translation invariant kernels, see Section 3.3 for details.

Now, to illustrate the impact of Theorem 2.4 we note that  $X$  is assumed to be bounded, and hence there are an  $x \in \mathbb{R}^d$  and an  $r > 0$  with  $X \subseteq x + rB_2^d$ . In the case of  $\min_i \sigma_i \geq 1/r$ , Lemma 3.13 then gives us

$$\mathcal{N}(D_\sigma X, 1) \leq \mathcal{N}(D_\sigma B_2^d, 1/r) \leq \sigma_1 \cdot \dots \cdot \sigma_d \cdot (3r)^d. \quad (10)$$

For the sake of completeness, we further mention that in the case of  $\max_i \sigma_i \leq 1/r$  we have  $\mathcal{N}(D_\sigma X, 1) = 1$ . Now, we can combine Theorem 2.4 with one of the theorems presented in Section 2.1. For example, by combining Theorem 2.4 with Theorem 2.1 and (10) we obtain

$$\mathcal{H}(\text{Id} : H_\sigma(rB_2^d) \rightarrow \ell_\infty(rB_2^d), \varepsilon) \leq \binom{4e + d}{d} \cdot (3r/e)^d \cdot \sigma_1 \cdot \dots \cdot \sigma_d \cdot \frac{\log^{d+1}(4/\varepsilon)}{\log \log^d(4/\varepsilon)} \quad (11)$$

for all  $0 < \varepsilon \leq 1$ , all  $r > 0$ , and all  $\sigma = (\sigma_1, \dots, \sigma_d) \in [1/r, \infty)^d$ . Finally, we mention that in the case of  $r \geq 1$  and an isotropic Gaussian kernel with width  $\sigma \geq 1$  the constant in (11), that

is

$$\tilde{K}_{d,\sigma,r} := K_{d,1} \cdot (3r\sigma)^d = \binom{4e+d}{d} \cdot (3/e)^d \cdot r^d \cdot \sigma^d, \quad (12)$$

grows like  $\sigma^d$  for  $\sigma \rightarrow \infty$ . In contrast, recall from (7) that the constant  $K_{d,\sigma}$  obtained in Theorem 2.1 grows like  $\sigma^{2d}$ . Consequently, (11) improves Theorem 2.1 in the dependency on  $\sigma$  by a factor of 2 in the exponent. In this respect note that [12] obtained the same behavior in  $\sigma$  but for a bound that does *not* include the double logarithmic factor  $\log \log^d(4/\varepsilon)$  in (11). Moreover, [11, Theorem 6.27] achieves the same behavior in  $\sigma$  for a polynomial bound of the form (2). Of course, the latter two results can be recovered from (11), and in addition, the results in [11, 12] do not take care of the explicit form of the constants.

### 3 Proofs

Before we present the proofs of our results, we briefly recall some basic facts about covering numbers. To this end, let  $S, T : U \rightarrow V$  and  $R : V \rightarrow W$  be some bounded operators between Banach spaces. Then the covering numbers satisfy

$$\mathcal{N}(S + T, \varepsilon + \|T\|) \leq \mathcal{N}(S, \varepsilon) \quad \text{and} \quad \mathcal{N}(RS, \varepsilon\|R\|) \leq \mathcal{N}(S, \varepsilon) \quad (13)$$

for all  $\varepsilon > 0$ . Furthermore, if  $T$  has a finite rank, then the covering numbers satisfy the following standard bound

$$\mathcal{N}(T, \varepsilon) \leq \left(1 + \frac{2\|T\|}{\varepsilon}\right)^{\text{rank } T} \quad (14)$$

For the proofs of these properties and a comprehensive introduction to this topic we refer to [1, Section 1.3], where we note that in [1] the proofs are for entropy numbers but they easily transfer to covering numbers.

#### 3.1 Isotropic Gaussian Kernels

Throughout this subsection the domain  $X := B_2^d \subseteq \mathbb{R}^d$  is fixed, and hence we simply write  $I_\sigma$  for the embedding  $\text{Id} : H_\sigma(B_2^d) \rightarrow \ell_\infty(B_2^d)$ . Before we prove the results of Subsection 2.1 we present several auxiliary lemmas. Our first result in this direction, which essentially repeats the key argument of [7, Theorem 3] on the input space  $X = B_2^d$  instead of  $X = [0, 1]^d$ , provides a general estimate for the log-covering numbers of  $I_\sigma$ .

**3.1 Lemma** *For all  $\sigma > 0$ ,  $\varepsilon > 0$ , and all integers  $N \geq 1$  we have*

$$\mathcal{H}\left(I_{\sigma,\varepsilon} + \sqrt{\frac{(2\sigma^2)^N}{N!}}\right) \leq \binom{N-1+d}{d} \cdot \log(1 + 2/\varepsilon).$$

*Proof.* For fixed  $\sigma > 0$ ,  $\varepsilon > 0$ , and  $N \geq 1$  we define

$$\varepsilon_0 := \sqrt{\frac{(2\sigma^2)^N}{N!}}.$$

In order to repeat the argument of [7, Theorem 3], we begin by recalling some notation: For every multi-index  $k = (k_1, \dots, k_d) \in \mathbb{N}_0^d$  we define the function  $e_k : B_2^d \rightarrow \mathbb{R}$  by

$$e_k(x) := \sqrt{\frac{(2\sigma^2)^{|k|}}{k!}} x^k \exp(-\sigma^2 \|x\|_{\ell_2^d}^2)$$

where we use  $|k| := k_1 + \dots + k_d$ ,  $k! = k_1! \dots k_d!$ , and  $x^k := x_1^{k_1} \dots x_d^{k_d}$  for  $x = (x_1, \dots, x_d) \in B_2^d$ . Since  $B_2^d$  has a non-empty interior the family of functions  $(e_k)_{k \in \mathbb{N}_0^d}$  forms an orthonormal basis (ONB) of  $H_\sigma(B_2^d)$  according to [11, Theorem 4.42]. Using this ONB we now consider, for  $N \geq 1$ , the orthogonal projections  $P_N, Q_N : H_\sigma(B_2^d) \rightarrow H_\sigma(B_2^d)$  onto  $\overline{\text{span}}\{e_k : |k| < N\}$  and  $\overline{\text{span}}\{e_k : |k| \geq N\}$ , respectively. From the first equation on page 494 of [7] we know

$$\|I_\sigma \circ Q_N\| \leq \sup_{x \in B_2^d} \sqrt{\frac{(2\sigma^2 \|x\|_{\ell_2^d}^2)^N}{N!}} = \sqrt{\frac{(2\sigma^2)^N}{N!}} = \varepsilon_0.$$

As a consequence of (13), (14), and  $\|I_\sigma \circ P_N\| = 1$  we get

$$\mathcal{H}(I_\sigma, \varepsilon + \varepsilon_0) = \mathcal{H}(I_\sigma \circ P_N + I_\sigma \circ Q_N, \varepsilon + \varepsilon_0) \leq \mathcal{H}(I_\sigma \circ P_N, \varepsilon) \leq \text{rank}(P_N) \log(1 + 2/\varepsilon).$$

Together with the formula

$$\text{rank}(P_N) = \binom{N-1+d}{d},$$

which was derived in [7, Remark 4], we thus obtain the assertion.  $\square$

Our next goal is to find suitable values of  $N \geq 1$  for the bound in Lemma 3.1. To this end, recall that Lambert's  $W$ -function is increasing and satisfies the relations  $W(x) > 0$  for  $x > 0$ ,  $W(x)e^{W(x)} = x$  for  $x \geq -1/e$ , and  $W(ye^y) = y$  for  $y \geq -1$ . In the following, we will often use these relations without referencing them.

**3.2 Lemma** For all  $\sigma > 0$ ,  $x > 0$ ,  $y \geq -\sigma^2$ ,

$$p_\sigma(x) := 2 \left( \frac{2e\sigma^2}{x} \right)^{x/2}, \quad \text{and} \quad h_\sigma(y) := 2e\sigma^2 \exp\left(W\left(\frac{y}{e\sigma^2}\right)\right)$$

the following statements are true:

- i). The function  $p_\sigma : (0, \infty) \rightarrow (0, \infty)$  is decreasing on  $(2\sigma^2, \infty)$  and  $\lim_{x \rightarrow \infty} p_\sigma(x) = 0$ .

ii). The function  $h_\sigma : [-\sigma^2, \infty) \rightarrow [2\sigma^2, \infty)$  is increasing and we have

$$h_\sigma(y) = \frac{2y}{W\left(\frac{y}{e\sigma^2}\right)}. \quad (15)$$

iii). The function  $p_\sigma : [2\sigma^2, \infty) \rightarrow (0, 2\exp(\sigma^2)]$  is bijective with inverse  $p_\sigma^{-1}$  given by

$$p_\sigma^{-1}(\varepsilon) = h_\sigma \circ \log(2/\varepsilon).$$

*Proof.* i). Some tedious calculations show that the derivative of  $p_\sigma$  is given by

$$p'_\sigma(x) = \frac{p_\sigma(x)}{2} \log\left(\frac{2\sigma^2}{x}\right), \quad x > 0.$$

From this identity the first assertion immediately follows. The second assertion is obvious.

ii). The monotonicity of  $h_\sigma$  is a consequence of the monotonicity of  $W$  and the definition of the function  $h_\sigma$ . Moreover, (15) follows from the identity  $W(x) \exp(W(x)) = x$ .

iii). By part i) we already know that  $p_\sigma : [2\sigma^2, \infty) \rightarrow (0, 2\exp(\sigma^2)]$  is bijective. To verify the formula for  $p_\sigma^{-1}$ , we fix some  $0 < \varepsilon \leq 2\exp(\sigma^2)$  and write  $y := \log(2/\varepsilon)$ . This immediately gives  $y \geq -\sigma^2$  and by the definition of  $h_\sigma$  we find

$$p_\sigma \circ h_\sigma(y) = 2 \left( \frac{2e\sigma^2}{h_\sigma(y)} \right)^{h_\sigma(y)/2} = 2 \exp\left(-W\left(\frac{y}{e\sigma^2}\right) \cdot \frac{h_\sigma(y)}{2}\right) = 2e^{-y} = \varepsilon,$$

i.e. we have shown the assertion.  $\square$

In the next lemma we choose a suitable parameter  $N \geq 1$  for the bound in Lemma 3.1 with the help of the functions introduced in Lemma 3.2.

**3.3 Lemma** For all  $\sigma > 0$  and all  $0 < \varepsilon \leq 1$ , we have

$$\mathcal{H}(I_\sigma, \varepsilon) \leq \binom{(h_\sigma \circ \log)(4/\varepsilon) + d}{d} \cdot \log(4/\varepsilon). \quad (16)$$

*Proof.* For a fixed  $0 < \varepsilon \leq 1$  we write  $y := \log(4/\varepsilon)$  and  $x := h_\sigma(y)$ . Since  $y > 1$  we have  $x > 2\sigma^2$ , and hence there is a unique integer  $N \geq 1$  with  $N - 1 < x \leq N$ . Using Lemma 3.1 with  $2\varepsilon/3$  instead of  $\varepsilon$ , the monotonicity of  $t \mapsto \binom{t+d}{d}$ , and  $1 \leq 1/\varepsilon$  we find

$$\mathcal{H}\left(I_\sigma, \frac{2\varepsilon}{3} + \sqrt{\frac{(2\sigma^2)^N}{N!}}\right) \leq \binom{N - 1 + d}{d} \cdot \log(1 + 3/\varepsilon) \leq \binom{x + d}{d} \cdot \log(4/\varepsilon).$$

Consequently, it remains to show that  $\sqrt{(2\sigma^2)^N/N!} \leq \varepsilon/3$  holds true. To this end, we use

Stirling's formula  $N! \geq \sqrt{2\pi N}(N/e)^N$  to get

$$\sqrt{\frac{(2\sigma^2)^N}{N!}} \leq \frac{1}{(2\pi N)^{1/4}} \cdot \left(\frac{2e\sigma^2}{N}\right)^{N/2} \leq \frac{p_\sigma(N)}{2(2\pi)^{1/4}}.$$

Moreover, the already observed  $x > 2\sigma^2$  together with parts *i)* and *iii)* of Lemma 3.2 yields

$$p_\sigma(N) \leq p_\sigma(x) = p_\sigma(h_\sigma(y)) = p_\sigma(h_\sigma \circ \log(4/\varepsilon)) = \varepsilon/2.$$

Combining both estimates and  $(2\pi)^{-1/4} \leq 4/3$  we get the assertion.  $\square$

Note that by an easy adaption of the above proof we can replace the 4 in  $y = \log(4/\varepsilon)$  by  $\gamma = 7/2$  if we choose  $4\varepsilon/5$  instead of  $2\varepsilon/3$  and use the bound  $(2\pi)^{-1/4} \approx 0.6316 \leq 7/10$ . Moreover, some tedious calculations show that the argument still works for

$$\gamma := \frac{3(2\pi)^{1/4} + 1 + \sqrt{9(2\pi)^{1/2} + 2(2\pi)^{1/4} + 1}}{2(2\pi)^{1/4}} \approx 3.4485.$$

Since these improvements have little impact we stick to  $\gamma = 4$  for convenience. The following lemma demonstrates the general technique we use to bound the right hand side of (16).

**3.4 Lemma** *Let  $d \geq 1$ ,  $\sigma > 0$ ,  $0 < \varepsilon_0 \leq 1$ , and  $t_0 > 0$ . If  $f : (0, \varepsilon_0] \rightarrow (0, \infty)$  is a function with  $f(\varepsilon) \geq t_0$  and*

$$(h_\sigma \circ \log)(4/\varepsilon) \leq f(\varepsilon)$$

for all  $0 < \varepsilon \leq \varepsilon_0$ , then we have

$$\binom{(h_\sigma \circ \log)(4/\varepsilon) + d}{d} \leq \binom{t_0 + d}{d} \cdot \left(\frac{f(\varepsilon)}{t_0}\right)^d.$$

*Proof.* In order to prove this statement we use the auxiliary function  $G_d(t) : (0, \infty) \rightarrow (0, \infty)$  defined by  $G_d(t) := \binom{t+d}{d} \cdot t^{-d} = \frac{1}{d!} \prod_{i=1}^d (1 + \frac{i}{t})$ . Since  $t \mapsto \binom{t+d}{d}$  is increasing and  $G_d$  is decreasing we get

$$\binom{(h_\sigma \circ \log)(4/\varepsilon) + d}{d} \leq \binom{f(\varepsilon) + d}{d} = G_d(f(\varepsilon)) \cdot f^d(\varepsilon) \leq G_d(t_0) \cdot f^d(\varepsilon)$$

for all  $0 < \varepsilon \leq \varepsilon_0$ , which gives the assertion.  $\square$

As final preparation we need the following simple lemma.

**3.5 Lemma** *For  $\sigma > 0$  consider  $t^* := \sigma^{-2} \exp(\sigma^{-2})$  and  $q_\sigma : (0, \infty) \rightarrow \mathbb{R}$  defined by*

$$q_\sigma(t) := \frac{1 + \log(\sigma^2) + \log(t)}{W(t)}.$$

Then  $q_\sigma$  is increasing on  $(0, t^*]$  and decreasing on  $[t^*, \infty)$ . Moreover,  $q_\sigma$  has a unique global maximum at  $t^*$  with  $q_\sigma(t^*) = 1 + \sigma^2$  and we have  $\lim_{t \rightarrow \infty} q_\sigma(t) = 1$ .

*Proof.* A simple but tedious calculation shows

$$q'_\sigma(t) = \frac{W(t) - \log(t\sigma^2)}{t \cdot W(t) \cdot (1 + W(t))}.$$

Since the denominator is positive for all  $t > 0$  we can focus on the numerator in order to investigate the monotonicity properties of  $q_\sigma$ . Consequently,  $q_\sigma$  is decreasing, if and only if  $W(t) < \log(t\sigma^2)$  and this is equivalent to

$$t = W(t)e^{W(t)} < \log(t\sigma^2) \cdot \exp \circ \log(t\sigma^2) = t\sigma^2 \log(t\sigma^2).$$

Rearranging this inequality for  $t$  shows that  $q_\sigma$  is decreasing on  $[t^*, \infty)$ . Analogously, we get that  $q_\sigma$  is increasing on  $(0, t^*]$  and that  $q_\sigma$  has a unique global maximum at  $t^*$ . Since  $W(t^*) = \sigma^{-2}$  and  $\log(t^*) = -\log(\sigma^2) + \sigma^{-2}$  we find  $q_\sigma(t^*) = 1 + \sigma^2$ . Finally, for  $t \geq 1$  we write  $s := te^t$  and from

$$q_\sigma(s) = \frac{1 + \log(\sigma^2) + \log(te^t)}{W(te^t)} = \frac{1 + \log(\sigma^2) + \log(t) + t}{t},$$

the assertion  $\lim_{t \rightarrow \infty} q_\sigma(t) = 1$  easily follows.  $\square$

*Proof of Theorem 2.1.* Let us define  $\varepsilon_0 := 1$  and  $y_0 := \log(4/\varepsilon_0)$ . For  $0 < \varepsilon \leq \varepsilon_0$  we further write  $y := \log(4/\varepsilon) \geq y_0 > 1$ . An application of Lemma 3.5 then yields

$$(h_\sigma \circ \log)(4/\varepsilon) = \frac{2y}{W(\frac{y}{e\sigma^2})} = \frac{2y}{\log(y)} \cdot \frac{1 + \log(\sigma^2) + \log(\frac{y}{e\sigma^2})}{W(\frac{y}{e\sigma^2})} \leq 2(1 + \sigma^2) \cdot \frac{y}{\log(y)} =: f(\varepsilon)$$

for all  $0 < \varepsilon \leq \varepsilon_0$ . Now, the derivative of  $\beta : (1, \infty) \rightarrow (0, \infty)$  defined by  $\beta(t) := \frac{t}{\log(t)}$  is

$$\beta'(t) = \frac{\log(t) - 1}{\log^2(t)}, \tag{17}$$

and consequently it is easy to check that, for  $t^* := e$ , the function  $\beta$  is decreasing on  $(1, t^*]$ , increasing on  $[t^*, \infty)$ , and has a unique global minimum at  $t^*$  with  $\beta(t^*) = e$ . As a result, we get  $f(\varepsilon) \geq 2e(1 + \sigma^2) =: t_0$ . Finally, combining Lemma 3.3 and Lemma 3.4 gives the assertion.  $\square$

*Proof of Theorem 2.2.* For a fixed  $0 < \varepsilon_0 \leq 4 \exp(-e^{1+\sigma^{-2}})$  we recall the definitions of  $y_0 := \log(4/\varepsilon_0)$  and  $x_0 := h_\sigma(y_0)$ . Moreover, for  $0 < \varepsilon \leq \varepsilon_0$  we write  $y := \log(4/\varepsilon) \geq y_0$ . Note that the restriction on  $\varepsilon_0$  ensures  $y_0 \geq \exp(1 + \sigma^{-2})$  and hence  $\frac{y_0}{e\sigma^2} \geq \sigma^{-2} \exp(\sigma^{-2})$ . As a consequence, the function  $y \mapsto \log(y)/W(\frac{y}{e\sigma^2})$  is decreasing on  $[y_0, \infty)$  according to Lemma 3.5

and we get

$$(h_\sigma \circ \log)(4/\varepsilon) = \frac{2 \log(y)}{W(\frac{y}{e\sigma^2})} \cdot \frac{y}{\log(y)} \leq \frac{2 \log(y_0)}{W(\frac{y_0}{e\sigma^2})} \cdot \frac{y}{\log(y)} =: f(\varepsilon).$$

Now, from (17) we know that the function  $\beta(t) = \frac{t}{\log(t)}$  is increasing on  $[e, \infty)$ , and hence

$$f(\varepsilon) \geq \frac{2 \log(y_0)}{W(\frac{y_0}{e\sigma^2})} \cdot \frac{y_0}{\log(y_0)} = \frac{2y_0}{W(\frac{y_0}{e\sigma^2})} = x_0 =: t_0$$

for all  $0 < \varepsilon \leq \varepsilon_0$ . Finally, combining Lemma 3.3 and Lemma 3.4 gives the assertion.  $\square$

*Proof of Theorem 2.3.* For  $0 < \varepsilon \leq 1$  we again write  $y := \log(4/\varepsilon) \geq \log(4)$ . In order to give a polynomial upper bound for  $\mathcal{H}(I_\sigma, \varepsilon)$  we use Lemma 3.3 and estimate the two factors,  $(h_\sigma(y)^d)$  and  $\log(4/\varepsilon)$ , appearing in (16) separately by a polynomial bound. To bound the first factor we fix a  $q_1 > 0$  and define the function

$$g_1(t) := 2 \frac{te^{-q_1 t}}{W(\frac{t}{e\sigma^2})}, \quad t > 0.$$

Using  $e^{-q_1 y} = (4/\varepsilon)^{-q_1}$  we then get

$$(h_\sigma \circ \log)(4/\varepsilon) = \frac{2y}{W(\frac{y}{e\sigma^2})} \left(\frac{4}{\varepsilon}\right)^{-q_1} \cdot \left(\frac{4}{\varepsilon}\right)^{q_1} \leq \left(\frac{4}{\varepsilon}\right)^{q_1} \sup_{t>0} g_1(t) =: f(\varepsilon)$$

and  $f(\varepsilon) \geq 4^{q_1} \cdot \sup_{t>0} g_1(t) =: t_0$ . A simple but tedious calculation shows

$$g_1'(t) = \frac{g_1(t)}{1 + W(\frac{t}{e\sigma^2})} \left( \sigma^{-2} \exp\left(-\left(1 + W\left(\frac{t}{e\sigma^2}\right)\right)\right) - q_1 \left(1 + W\left(\frac{t}{e\sigma^2}\right)\right) \right).$$

If we define

$$t^* := \frac{1}{q_1} \left(1 - \frac{1}{W(\frac{1}{q_1\sigma^2})}\right)$$

then another tedious calculation shows that  $g_1$  is increasing on  $(0, t^*]$ , decreasing on  $[t^*, \infty)$ , and has a unique global maximum at  $t^*$ . In order to evaluate the maximum  $g_1(t^*)$  we first give another representation of  $t^*$  using  $t/W(t) = \exp(W(t))$  for  $t = \frac{1}{q_1\sigma^2}$

$$t^* = \left(W\left(\frac{1}{q_1\sigma^2}\right) - 1\right) \cdot \frac{1/q_1}{W(\frac{1}{q_1\sigma^2})} = \sigma^2 \cdot \left(W\left(\frac{1}{q_1\sigma^2}\right) - 1\right) \cdot \exp \circ W\left(\frac{1}{q_1\sigma^2}\right).$$

Using this representation together with  $W(xe^x) = x$  for  $x = W(\frac{1}{q_1\sigma^2}) - 1$  gives us

$$\frac{t^*}{W(\frac{t^*}{e\sigma^2})} = \frac{t^*}{W(\frac{1}{q_1\sigma^2}) - 1} = \frac{1}{q_1 \cdot W(\frac{1}{q_1\sigma^2})}.$$

Using this identity we directly get

$$t_0 = 4^{q_1} g_1(t^*) = 2 \cdot 4^{q_1} \cdot \frac{e^{-q_1 t^*}}{q_1 \cdot W\left(\frac{1}{q_1 \sigma^2}\right)} = \frac{2 \cdot 4^{q_1}}{e q_1 \cdot W\left(\frac{1}{q_1 \sigma^2}\right)} \exp\left(1/W\left(\frac{1}{q_1 \sigma^2}\right)\right)$$

and Lemma 3.4 gives us

$$\binom{(h_\sigma \circ \log)(4/\varepsilon) + d}{d} \leq \binom{t_0 + d}{d} \cdot \left(\frac{f(\varepsilon)}{t_0}\right)^d = \binom{t_0 + d}{d} \cdot 4^{-q_1 d} \cdot (4/\varepsilon)^{q_1 d}. \quad (18)$$

Now, we estimate the second factor  $y = \log(4/\varepsilon)$  by a polynomial bound of order  $q_2 > 0$ . To this end, we define the function  $g_2(t) := t e^{-q_2 t}$ , for  $t > 0$ , and estimate

$$y = (4/\varepsilon)^{q_2} \cdot y \cdot (4/\varepsilon)^{-q_2} \leq (4/\varepsilon)^{q_2} \cdot \sup_{t>0} g_2(t).$$

An easy calculation shows that the derivative of  $g_2$  is given by  $g_2'(t) = g_2(t) \cdot (1/t - q_2)$  and consequently  $g_2$  has a global maximum at  $t^* := 1/q_2$  with  $g_2(t^*) = \frac{1}{e q_2}$ . Therefore, we get

$$y \leq \frac{(4/\varepsilon)^{q_2}}{e q_2}. \quad (19)$$

Finally, combining Lemma 3.3 with (18) and (19) yields

$$\mathcal{H}(I_\sigma, \varepsilon) \leq \binom{t_0 + d}{d} \cdot \frac{1}{e q_2 \cdot 4^{q_1 d}} \cdot (4/\varepsilon)^{q_1 d + q_2},$$

and for  $q_1 = q_2 = \frac{p}{d+1}$  we get the assertion.  $\square$

### 3.2 Auxiliary Results

In this section we collect additional results that are helpful to understand the quantities appearing in the bounds presented in Section 2.1.

**3.6 Lemma** *For an integer  $d \geq 1$  and a real number  $t > 0$  the (generalized) binomial coefficient from (5) satisfies*

$$\binom{t+d}{d} = \frac{\Gamma(t+d+1)}{\Gamma(t+1)\Gamma(d+1)}.$$

Moreover, for a fixed real number  $t > 0$  the sequence

$$a_d := \binom{t+d}{d} \cdot d^{-t}, \quad d \geq 1,$$

is decreasing and converges to  $1/\Gamma(t+1)$ .

*Proof.* First note that  $\Gamma(d+1) = d!$  and an  $d$ -times application of  $\Gamma(t+1) = t \cdot \Gamma(t)$  gives us

$$\binom{t+d}{d} = \frac{1}{d!} \prod_{i=1}^d (t+i) = \frac{\Gamma(t+1) \prod_{i=1}^d (t+i)}{\Gamma(d+1)\Gamma(t+1)} = \frac{\Gamma(d+t+1)}{\Gamma(d+1)\Gamma(t+1)}.$$

Using  $\frac{\Gamma(d+t+1)}{\Gamma(d+1)d^t} \rightarrow 1$  for  $d \rightarrow \infty$ , which is a well-known property of the Gamma function, we get

$$a_d = \frac{1}{\Gamma(t+1)} \cdot \frac{\Gamma(d+t+1)}{\Gamma(d+1)d^t} \rightarrow \frac{1}{\Gamma(t+1)}$$

for  $d \rightarrow \infty$  and it remains to show the monotonicity. Using  $\Gamma(t+1) = t \cdot \Gamma(t)$  twice we get

$$a_{d+1} = \frac{\Gamma(d+t+2)}{\Gamma(d+2)\Gamma(t+1)} (d+1)^{-t} = a_d \cdot \left(\frac{d}{d+1}\right)^t \cdot \frac{d+t+1}{d+1}.$$

Consequently,  $(a_d)_{d \geq 1}$  is decreasing if and only if  $\frac{d+t+1}{d+1} < \left(\frac{d+1}{d}\right)^t$  is satisfied for all  $d \geq 1$  and  $t > 0$ . In order to prove this we fix some  $d \geq 1$  and show that

$$f_d(t) := \left(1 + \frac{1}{d}\right)^t - \left(1 + \frac{t}{d+1}\right) > 0$$

is satisfied for all  $t > 0$ . To this end, we calculate the first and second derivative

$$\begin{aligned} f'_d(t) &= \left(1 + \frac{1}{d}\right)^t \log\left(1 + \frac{1}{d}\right) - \frac{1}{d+1} \\ f''_d(t) &= \left(1 + \frac{1}{d}\right)^t \log^2\left(1 + \frac{1}{d}\right). \end{aligned}$$

Using  $\log(1+x) \geq \frac{x}{1+x}$ , which holds for all  $x > -1$ , for  $x = 1/d$ , we get

$$f'_d(0) = \log\left(1 + \frac{1}{d}\right) - \frac{1}{d+1} \geq \frac{1/d}{1+1/d} - \frac{1}{d+1} = 0.$$

Together with  $f''_d(t) > 0$  we get  $f'_d(t) > 0$  for all  $t > 0$ . Finally,  $f_d(0) = 0$  and  $f'_d(t) > 0$  gives  $f_d(t) > 0$  for all  $t > 0$  and hence the assertion is proven.  $\square$

**3.7 Lemma** For all  $C > 0$ ,  $d \geq 2Ce^2$ , and  $\varepsilon_0 := 4 \exp\left(-\frac{d}{2C} \log\left(\frac{d}{2eC}\right)\right)$  the condition  $\varepsilon_0 \leq 4 \exp(-e^{1+\sigma^{-2}})$  for  $\sigma = 1$  in Theorem 2.2 is satisfied and the quantity  $K_{d,1,\varepsilon_0}$  defined in Theorem 2.2 satisfies

$$K_{d,1,\varepsilon_0} \leq (2\pi)^{-1/2} \cdot (4e)^d (1+C)^d \cdot d^{-(d+1/2)}.$$

Moreover, for  $\frac{1}{2e^2} \geq C \geq \frac{1}{\sqrt{360e}}$  we have

$$\exp(-90d^2 - 11d - 3) \leq \varepsilon_0.$$

*Proof.* Let us recall the definition of  $y_0 := \log(4/\varepsilon_0)$  and  $x_0 := \frac{2y_0}{W(y_0/e)} = h_1(y_0)$  in Theorem 2.2

where  $h_1$  is defined in Lemma 3.2. Using the function  $p_1$  defined in Lemma 3.2 we can write

$$\varepsilon_0 = 4 \left( \frac{2eC}{d} \right)^{\frac{d}{2C}} = 2 \cdot p_1(d/C).$$

Since  $d/C \geq 2e^2 \geq 2$  part *iii*) of Lemma 3.2, which states  $p_1^{-1} = h_1 \circ \log(2/\cdot)$ , is applicable and hence

$$x_0 = h_1 \circ \log(4/\varepsilon_0) = h_1 \circ \log\left(\frac{2}{p_1(d/C)}\right) = d/C.$$

Next, we prove the inequality  $\varepsilon_0 \leq u := 4 \exp(-e^2)$ . To this end, note that we have  $h_1 \circ \log(4/u) = h_1(e^2) = 2e \exp(W(e)) = 2e^2$  since  $W(e) = 1$ . Our assumption  $d \geq 2Ce^2$  implies

$$h_1 \circ \log(4/\varepsilon_0) = d/C \geq 2e^2 = h_1 \circ \log(4/u)$$

and since  $h_1$  is increasing according to part *ii*) of Lemma 3.2 we get  $\varepsilon_0 \leq u = 4 \exp(-e^2)$ . Now, we prove the bound on  $K_{d,1,\varepsilon_0}$ . To this end, we rewrite  $K_{d,1,\varepsilon_0}$  using the representation of the binomial coefficient from (5)

$$\begin{aligned} K_{d,1,\varepsilon_0} &= \binom{x_0 + d}{d} x_0^{-d} \cdot \left( \frac{x_0 \log(y_0)}{y_0} \right)^d \\ &= \frac{1}{d!} \prod_{i=1}^d (1 + i/x_0) \cdot \left( \frac{2 \log(y_0)}{W(y_0/e)} \right)^d. \end{aligned}$$

If we bound the first factor by using  $i/x_0 \leq d/x_0 = C$  and if we bound the second factor by using  $\log(y_0) \leq 2W(y_0/e)$  from Lemma 3.5 then we get

$$K_{d,1,\varepsilon_0} \leq \frac{4^d (1+C)^d}{d!}.$$

Together with Stirling's formula  $d! \geq \sqrt{2\pi d} \cdot (d/e)^d$  this gives the desired bound. Finally, note that  $\log(t) \leq t$  and  $C \geq 1/\sqrt{360e}$  yields

$$\varepsilon_0 \geq \exp\left(-\frac{d^2}{4eC^2}\right) \geq \exp(-90d^2),$$

which proves the lower bound on  $\varepsilon_0$ . □

**3.8 Lemma** *For  $\sigma, p > 0$  there are constants  $c_{\sigma,p}, C_{\sigma,p} > 0$  such that  $K_{d,\sigma,p}$  defined in (8) satisfies, for all  $d \geq 1$*

$$K_{d,\sigma,p} \leq C_{\sigma,p} \cdot \sqrt{d \log(d)} \cdot \exp\left(c_{\sigma,p} \cdot d \cdot \frac{\log \log(d)}{\log(d)}\right).$$

*Proof.* For this proof we use the usual notation  $a_n \asymp b_n$  for two sequences  $(a_n)_{n \geq 1}, (b_n)_{n \geq 1}$  iff there is a constant  $c > 0$  with  $a_n \leq cb_n$  for all  $n \geq 1$ . Moreover, we write  $a_n \asymp b_n$  iff both

$a_n \leq b_n$  and  $a_n \geq b_n$  hold. Using  $W(t) \sim \log(t)$  we get, for  $d \rightarrow \infty$ ,

$$t_0 = \frac{2(d+1) \cdot 4^{\frac{p}{d+1}}}{ep \cdot W(\frac{d+1}{p\sigma^2})} \exp\left(\frac{1}{W(\frac{d+1}{p\sigma^2})}\right) \asymp \frac{d}{W(\frac{d+1}{p\sigma^2})} \asymp \frac{d}{\log(\frac{d+1}{p\sigma^2})} \asymp \frac{d}{\log(d)}.$$

Since  $t_0 \rightarrow \infty$  for  $d \rightarrow \infty$ , Lemma 3.6 and Stirling's formula  $\Gamma(t+1) \sim \sqrt{2\pi t} (t/e)^t$  yield

$$\binom{t_0+d}{d} = \frac{\Gamma(t_0+d+1)}{\Gamma(t_0+1)\Gamma(d+1)} \asymp \left(\frac{1}{t_0} + \frac{1}{d}\right)^{1/2} \left(1 + \frac{t_0}{d}\right)^d \left(1 + \frac{d}{t_0}\right)^{t_0}$$

for  $d \rightarrow \infty$ . Using the inequality  $1+t \leq e^t$ , which holds for all  $t \in \mathbb{R}$ , for  $t = t_0/d$  we get

$$\binom{t_0+d}{d} \leq \sqrt{\frac{\log(d)}{d}} \cdot e^{t_0} \cdot (1+d/t_0)^{t_0}$$

for  $d \rightarrow \infty$ . Consequently, we find a constant  $C_{\sigma,p} > 0$  with

$$K_{d,\sigma,p} = \binom{t_0+d}{d} \cdot \frac{d+1}{ep} \cdot 4^{\frac{p}{d+1}} \leq C_{\sigma,p} \cdot \sqrt{d \log(d)} \cdot \exp\left(t_0 \cdot \log\left(e + e \frac{d}{t_0}\right)\right).$$

Since, for  $d \rightarrow \infty$ , the exponent behaves like

$$t_0 \cdot \log\left(e + e \frac{d}{t_0}\right) \asymp \frac{d}{\log(d)} \cdot \log\left(e + e \log(d)\right) \asymp d \cdot \frac{\log \log(d)}{\log(d)}$$

there is a constant  $c_{\sigma,p} > 0$  independent of  $d$  with the desired property.  $\square$

**3.9 Lemma** For  $0 < p_0 \leq 1/e$  the quantities  $K_{d,1,p}$  and  $C_0$  defined in (8) and (9), respectively, satisfy

$$K_{d,1,p} \leq 1/2 \cdot C_0^d \cdot \sqrt{d} \cdot \frac{(1/p)^{d+1}}{\log^d(1/p)}, \quad 0 < p \leq p_0, \quad d \geq 1.$$

*Proof.* As a first step we bound  $t_0$ . To this end, we write  $g(p_0) := 2^{1+p_0} \cdot \exp(1/W(2/p_0)) \cdot (2+1/e)$  and bound  $W$  by  $\log$  with the help of Lemma 3.5, that is

$$\frac{\log(1/p)}{W(\frac{d+1}{p})} = \frac{1 + \log(\frac{1}{(d+1)e}) + \log(\frac{d+1}{p})}{W(\frac{d+1}{p})} \leq 1 + \frac{1}{(d+1)e} \leq d \cdot \frac{2+1/e}{d+1}.$$

Together with  $p \leq p_0$  and  $d \geq 1$  we can bound  $t_0$  by

$$t_0 = \frac{2(d+1) \cdot 4^{\frac{p}{d+1}}}{ep \cdot W(\frac{d+1}{p})} \exp\left(\frac{1}{W(\frac{d+1}{p})}\right) \leq \frac{g(p_0)}{e} \cdot d \cdot \frac{1/p}{\log(1/p)} =: f(p).$$

Since  $\beta(t) = \frac{t}{\log(t)}$  is increasing on  $[e, \infty)$  according to (17) and  $1/p \geq 1/p_0 \geq e$  we get  $f(p) \geq f(p_0) = g(p_0)/e \cdot d \cdot z_0$ , where  $z_0 := \frac{1/p_0}{\log(1/p_0)}$ . Repeating the proof of Lemma 3.4

together with  $d \geq 1$  and  $1/p \geq 1/p_0$  yields

$$\begin{aligned} K_{d,1,p} &= \binom{t_0 + d}{d} \cdot \frac{d+1}{ep} \cdot 4^{\frac{p}{d+1}} \leq \binom{f(p_0) + d}{d} \cdot \left(\frac{f(p)}{f(p_0)}\right)^d \cdot d \cdot \frac{2^{1+p_0}}{e} \cdot 1/p \\ &= \binom{f(p_0) + d}{d} \cdot z_0^{-d} \cdot d \cdot \frac{2^{1+p_0}}{e} \cdot \frac{(1/p)^{d+1}}{\log^d(1/p)}. \end{aligned}$$

Now, we bound the quantities depending on  $d$ . To this end, we use Lemma 3.6 together with Stirling's formula,  $f(p_0) \geq 4$ ,  $d \geq 1$ , and  $1+t \leq e^t$  for  $t = d/f(p_0)$

$$\begin{aligned} \binom{f(p_0) + d}{d} \cdot z_0^{-d} \cdot d &\leq \frac{e^{1/60}}{\sqrt{2\pi}} \cdot \left(1 + \frac{1}{f(p_0)}\right)^{1/2} \left(1 + \frac{f(p_0)}{d}\right)^d \left(1 + \frac{d}{f(p_0)}\right)^{f(p_0)} \cdot z_0^{-d} \cdot d \\ &\leq \frac{e^{1/60}}{\sqrt{2\pi}} \cdot \left(1 + \frac{e}{g(p_0) \cdot z_0}\right)^{1/2} \cdot \left(\frac{1 + g(p_0)/e \cdot z_0}{z_0}\right)^d \cdot e^d \cdot d^{1/2}. \end{aligned}$$

Since  $C_0 = e/z_0 + g(p_0)$  and the arising quantities that are independent of  $p$  and  $d$  satisfy

$$\frac{e^{1/60}}{\sqrt{2\pi}} \cdot \left(1 + \frac{e}{g(p_0) \cdot z_0}\right)^{1/2} \cdot \frac{2^{1+p_0}}{e} \leq \frac{e^{1/60}}{\sqrt{2\pi}} \cdot \left(1 + \frac{1}{2(2+1/e)}\right)^{1/2} \cdot \frac{2^{1+1/e}}{e} \approx 0.4239 \leq 1/2$$

the assertion is proven.  $\square$

### 3.3 Anisotropic Gaussian Kernels

In this section we first provide some general theory about covering numbers of RKHSs and finally prove Theorem 2.4. To this end, we introduce some notation. For a fixed bounded kernel  $k$  defined on a set  $X$  we often consider its restriction to different subsets  $Y \subseteq X$ . Consequently, we highlight the considered domain by writing  $H(Y)$  for the corresponding RKHS and by using the abbreviation  $I[Y]$  for the corresponding embedding  $\text{Id} : H(Y) \rightarrow \ell_\infty(Y)$ . Recall that  $I[Y]$  is well-defined according to [11, Lemma 4.23].

**3.10 Lemma** *Let  $T : Y \rightarrow X$  be a mapping between two non-empty sets and  $k$  be a bounded kernel on  $X$  with RKHS  $H(X)$ . Then*

$$k_T(y, y') := k(T(y), T(y')), \quad y, y' \in Y, \quad (20)$$

*defines a bounded kernel on  $Y$  with RKHS  $H_T(Y) = \{f \circ T : f \in H(X)\}$  and the corresponding RKHS-norm satisfies*

$$\|f \circ T\|_{H_T(Y)} \leq \|f\|_{H(X)}$$

*for  $f \in H(X)$ . Moreover, the covering numbers satisfy*

$$\mathcal{N}(\text{Id} : H_T(Y) \rightarrow \ell_\infty(Y), \varepsilon) \leq \mathcal{N}(\text{Id} : H(X) \rightarrow \ell_\infty(X), \varepsilon), \quad \varepsilon > 0. \quad (21)$$

If, in addition,  $T$  is bijective, then equality holds in (21).

*Proof.* Let  $\Phi : X \rightarrow H(X)$  be the canonical feature map of  $k$ , that is  $\Phi(x) := k(x, \cdot)$  for  $x \in X$ . Then it is easy to see that  $\Phi_T := \Phi \circ T$  is a feature map for  $k_T$ . Consequently,  $k_T$  is a kernel on  $Y$ , and according to [11, Theorem 4.21] the RKHS of  $k_T$  has the claimed form, the claimed norm inequality is satisfied, and  $S_H : H(X) \rightarrow H_T(Y)$  defined by  $f \mapsto f \circ T$  is a metric surjection, i.e.  $S_H \mathring{B}_{H(X)} = \mathring{B}_{H_T(Y)}$ . Now, it remains to prove the covering number bounds. To this end, we define the mapping  $S_\infty : \ell_\infty(X) \rightarrow \ell_\infty(Y)$  by  $f \mapsto f \circ T$ . Using the metric surjectivity of  $S_H$  and  $I_T[Y] \circ S_H = S_\infty \circ I[X]$  we get, for  $\varepsilon > 0$ ,

$$\mathcal{N}(I_T[Y], \varepsilon) = \mathcal{N}(I_T[Y] \circ S_H, \varepsilon) = \mathcal{N}(S_\infty \circ I[X], \varepsilon).$$

Since  $\|S_\infty f\|_{\ell_\infty(Y)} = \sup_{y \in Y} |f(T(y))| \leq \|f\|_{\ell_\infty(X)}$  is satisfied for all  $f \in \ell_\infty(X)$  we have  $\|S_\infty\| \leq 1$  and together with (13) this yields the assertion. If  $T$  is bijective we can exchange the role of  $X$  and  $Y$  and hence we get the claimed equality.  $\square$

**3.11 Lemma** *Let  $X = X_1 \cup X_2$  be the disjoint union of non-empty sets  $X_1, X_2$  and  $k$  be a bounded kernel on  $X$  with RKHS  $H(X)$ . Then for all  $\varepsilon > 0$  we have*

$$\mathcal{N}(\text{Id} : H(X) \rightarrow \ell_\infty(X), \varepsilon) \leq \mathcal{N}(\text{Id} : H(X_1) \rightarrow \ell_\infty(X_1), \varepsilon) \cdot \mathcal{N}(\text{Id} : H(X_2) \rightarrow \ell_\infty(X_2), \varepsilon).$$

*Proof.* Let  $m := \mathcal{N}(I[X_1], \varepsilon)$  and  $n := \mathcal{N}(I[X_2], \varepsilon)$ . Moreover, choose corresponding  $\varepsilon$ -nets  $f_1, \dots, f_m \in \ell_\infty(X_1)$  and  $g_1, \dots, g_n \in \ell_\infty(X_2)$ . Then for each  $i \in \{1, \dots, m\}$  and each  $j \in \{1, \dots, n\}$  we define

$$h_{i,j}(x) := \begin{cases} f_i(x), & x \in X_1 \\ g_j(x), & x \in X_2, \end{cases} \quad \text{for } x \in X.$$

This defines at most  $m \cdot n$  different elements of  $\ell_\infty(X)$  and it remains to show that  $h_{i,j}$  for  $i = 1, \dots, m$  and  $j = 1, \dots, n$  defines an  $\varepsilon$ -net of  $B_{H(X)}$ . For  $h \in H(X)$  with  $\|h\|_{H(X)} \leq 1$  we have  $h|_{X_\ell} \in H(X_\ell)$  with  $\|h|_{X_\ell}\|_{H(X_\ell)} \leq 1$ , for  $\ell = 1, 2$ , see Lemma 3.10. Consequently, there is an  $i \in \{1, \dots, m\}$  and a  $j \in \{1, \dots, n\}$  with  $\|h|_{X_1} - f_i\|_{\ell_\infty(X_1)} \leq \varepsilon$  and  $\|h|_{X_2} - g_j\|_{\ell_\infty(X_2)} \leq \varepsilon$ , respectively. For this choice of  $i$  and  $j$  we have

$$\|h - h_{i,j}\|_{\ell_\infty(X)} = \max\{\|h|_{X_1} - f_i\|_{\ell_\infty(X_1)}, \|h|_{X_2} - g_j\|_{\ell_\infty(X_2)}\} \leq \varepsilon$$

and hence the assertion is proven.  $\square$

So far, we considered bounded kernels on general sets. In the following, we investigate bounded kernels  $k : V \times V \rightarrow \mathbb{R}$  on a vector space  $V$ . The kernel  $k$  is called *translation invariant* along  $a \in V$  if

$$k(v + a, v' + a) = k(v, v')$$

is satisfied for all  $v, v' \in V$ . In this case the transformation  $T(x) := x + a$  does not change the kernel, i.e.  $k = k_T$ . Since  $T$  is bijective as a mapping  $X \rightarrow a + X$ , Lemma 3.10 yields

$$\mathcal{N}(\text{Id} : H(X) \rightarrow \ell_\infty(X), \varepsilon) = \mathcal{N}(\text{Id} : H(X + a) \rightarrow \ell_\infty(X + a), \varepsilon), \quad \varepsilon > 0. \quad (22)$$

If  $k$  is translation invariant along all  $a \in U \subseteq V$  for some subspace  $U \subseteq V$ , then we call  $k$  translation invariant along  $U$ .

**3.12 Lemma** *Let  $(V, \|\cdot\|)$  be a Banach space with complemented subspaces  $V_1, V_2 \subseteq V$ , i.e.  $V = V_1 + V_2$  and  $V_1 \cap V_2 = \{0\}$ . Moreover, let  $X_i \subseteq V_i$  be non-empty subsets, for  $i = 1, 2$ , and  $k$  be a bounded kernel on  $V$ . If  $k$  is translation invariant along  $V_1$  and  $X_1$  is relatively compact, then the log-covering numbers satisfy, for  $\delta > 0$  and  $\varepsilon > 0$ ,*

$$\mathcal{H}(\text{Id} : H(X_1 + X_2) \rightarrow \ell_\infty(X_1 + X_2), \varepsilon) \leq \mathcal{N}(X_1, \delta) \cdot \mathcal{H}(\text{Id} : H(\delta B_{V_1} + X_2) \rightarrow \ell_\infty(\delta B_{V_1} + X_2), \varepsilon).$$

*Proof.* Let us fix some  $\varepsilon, \delta > 0$  and set  $n := \mathcal{N}(X_1, \delta)$ . For a minimal  $\delta$ -net  $x_{1,1}, \dots, x_{1,n} \in V_1$  of  $X_1$  we choose a partition  $X_{1,1}, \dots, X_{1,n}$  of  $X_1$  with  $X_{1,i} \subseteq x_{1,i} + \delta B_{V_1}$  for all  $i = 1, \dots, n$ . Since we have chosen a minimal  $\delta$ -net  $X_{1,i} \neq \emptyset$  is satisfied for  $i = 1, \dots, n$ . Because  $X_i \subseteq V_i$ , for  $i = 1, 2$ , and  $V_1, V_2$  are complemented subspaces the sets  $X_{1,i} + X_2$ , for  $i = 1, \dots, n$ , form a partition of  $X_1 + X_2$  with  $X_{1,i} + X_2 \subseteq x_{1,i} + \delta B_{V_1} + X_2$ . A multiple application of Lemma 3.11 and an application of Lemma 3.10 for  $T = \text{Id}$  yield

$$\mathcal{H}(I[X], \varepsilon) \leq \sum_{i=1}^n \mathcal{H}(I[X_{1,i} + X_2], \varepsilon) \leq \sum_{i=1}^n \mathcal{H}(I[x_{1,i} + (\delta B_{V_1}) + X_2], \varepsilon).$$

Since  $k$  is translation invariant along  $V_1$ , Equation (22) yields the assertion.  $\square$

*Proof of Theorem 2.4.* Let  $X \subseteq \mathbb{R}^d$  be a bounded subset and  $\sigma = (\sigma_1, \dots, \sigma_d) \in (0, \infty)^d$ . With the notation introduced in (20) the Gaussian kernel then writes as  $k_\sigma = k_{D_\sigma}$ . Since the diagonal operator  $D_\sigma : X \rightarrow D_\sigma X$  is bijective, Lemma 3.10 yields  $\mathcal{H}(I_\sigma[X], \varepsilon) = \mathcal{H}(I_1[D_\sigma X], \varepsilon)$ . Together with Lemma 3.12 for  $\delta = 1$ ,  $V_1 = \mathbb{R}^d$  (equipped with the Euclidean norm),  $V_2 = \{0\}$ , and  $X_1 = D_\sigma X$ ,  $X_2 = \{0\}$  we get the assertion.  $\square$

Finally, we present a lemma bounding the covering numbers of convex sets  $X \subseteq \mathbb{R}^d$ . This result is well-known but we did not find exactly this one in the literature and hence we included a proof for convenience.

**3.13 Lemma** *Let  $X \subseteq \mathbb{R}^d$  be a convex set and  $r_0 > 0$  such that there is an  $a \in \mathbb{R}^d$  with  $a + r_0 B_2^d \subseteq X$ . Then we have*

$$\mathcal{N}(X, 2\varepsilon) \leq \frac{\lambda^d(X)}{\lambda^d(B_2^d)} \left( \frac{1}{r_0} + \frac{1}{\varepsilon} \right)^d, \quad \varepsilon > 0, \quad (23)$$

where the covering numbers are with respect to the Euclidean norm and  $\lambda^d$  denotes the  $d$ -dimensional Lebesgue measure.

*Proof.* For this proof we use *packing numbers*, which for  $\varepsilon > 0$  are defined by

$$\mathcal{P}(X, \varepsilon) := \max\left\{n \geq 1 : \exists x_1, \dots, x_n \in X \text{ with } \|x_i - x_j\|_{\ell_2^d} > 2\varepsilon \forall i \neq j\right\}.$$

Recall that  $\mathcal{P}(X, 2\varepsilon) \leq \mathcal{N}(X, 2\varepsilon) \leq \mathcal{P}(X, \varepsilon)$  holds for all  $\varepsilon > 0$ , see e.g. [6, Theorem IV]. Consequently, it is enough to bound  $\mathcal{P}(X, \varepsilon)$  by the right hand side of (23). For  $\varepsilon > 0$  we set  $n := \mathcal{P}(X, \varepsilon)$  and choose  $x_1, \dots, x_n \in X$  with  $\|x_i - x_j\|_{\ell_2^d} > 2\varepsilon$  for all  $i \neq j$ . Then the sets  $x_i + \varepsilon B_2^d$  are disjoint subsets of  $X + \varepsilon B_2^d$  and hence

$$n\varepsilon^d \lambda^d(B_2^d) = \lambda^d\left(\bigcup_{i=1}^n (x_i + \varepsilon B_2^d)\right) \leq \lambda^d(X + \varepsilon B_2^d).$$

Since  $X$  is convex we have  $s_1 X + s_2 X = (s_1 + s_2)X$  for  $s_1, s_2 > 0$ . Together with  $r_0 B_2^d \subseteq X - a$  we get

$$X + \varepsilon B_2^d = X + \frac{\varepsilon}{r_0} \cdot r_0 B_2^d \subseteq X + \frac{\varepsilon}{r_0} \cdot (X - a) = \left(1 + \frac{\varepsilon}{r_0}\right)X - \frac{\varepsilon}{r_0}a.$$

Both bounds together yield  $n\varepsilon^d \lambda^d(B_2^d) \leq \lambda^d(X)(1 + \varepsilon/r_0)^d$ , which gives the assertion.  $\square$

## Acknowledgment

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Simon Fischer.

## References

- [1] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*, Cambridge University Press, Cambridge, 1990.
- [2] F. Cucker and D.-X. Zhou. *Learning Theory*, Cambridge University Press, Cambridge, 2007.
- [3] M. Eberts and I. Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.*, 7:1–42, 2013.
- [4] M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *Mach. Learn.*, 108:203–227, 2018.
- [5] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv e-prints*, 1807.02582v1, 2018.

- [6] A. N. Kolmogorov and V. M. Tikhomirov.  $\varepsilon$ -entropy and  $\varepsilon$ -capacity of sets in functional spaces. *Uspekhi Mat. Nauk*, 17, 1961.
- [7] T. Kühn. Covering numbers of Gaussian reproducing kernel Hilbert spaces. *J. Complexity*, 27:489–499, 2011.
- [8] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, 2006.
- [9] B. Schölkopf and A. J. Smola. *Learning with Kernels*, MIT Press, Cambridge, 2001.
- [10] S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl. (Singap.)*, 01:17–41, 2002.
- [11] I. Steinwart and A. Christmann. *Support Vector Machines*, Springer, New York, 2008.
- [12] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Ann. Statist.*, 37:2655–2675, 2009.
- [13] D.-X. Zhou. The covering number in learning theory. *J. Complexity*, 18:739–767, 2002.
- [14] D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Trans. Inf. Theory*, 49:1743–1752, 2003.