An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels*†

Ingo Steinwart, Don Hush, Clint Scovel
Modeling, Algorithms and Informatics Group, CCS-3
Los Alamos National Laboratory
{ingo,dhush,jcs}@lanl.gov

February 1, 2005

Abstract

Although Gaussian RBF kernels are one of the most often used kernels in modern machine learning methods such as support vector machines (SVMs), almost nothing is known about the structure of their reproducing kernel Hilbert spaces (RKHSs). In this work we give two distinct explicit descriptions of the RKHSs corresponding to Gaussian RBF kernels and discuss some consequences. Furthermore, we present an orthonormal system for these spaces. Finally we discuss how our results can be used for analyzing the learning performance of SVMs.

Index Terms: Learning Theory, Support Vector Machines, Gaussian RBF Kernels

1 Introduction

In recent years support vector machines and related kernel-based algorithms (see e.g. [1] and [2] for introductions) have become the state-of-the-art methods for many machine learning problems. The common feature of these methods is that they are based on an optimization problem over a reproducing kernel Hilbert space (RKHS). If the underlying input space X of the machine learning problem has a specific structure, e.g. text strings or DNA sequences, one often uses a RKHS which is suitable to this structure (see e.g. [3] for a recent and thorough overview). If however X is a subset of $\mathbb{R}^d$ then the commonly recommended choice are the RKHSs of the Gaussian RBF kernels (see e.g. [4]). Although there has been substantial progress in understanding these RKHSs and their role in the learning process (see e.g. [5] and [6]) some simple questions are still open. For example, it is still unknown which functions are contained in these RKHSs, how the corresponding norms can be computed, and how the RKHSs for different widths correlate to each other. The aim of this paper is to answer these questions. In addition we discuss how our results can be used to bound the approximation error function of SVMs which plays a crucial role in the analysis of the learning performance of these learning algorithms.

The rest of the paper is organized as follows. In Section 2 we recall the definition and basic facts on kernels and RKHSs. In Section 3 we present our main results and discuss their consequences. Finally, Section 4 contains the proofs of the main theorems.

*Los Alamos Unclassified Report LA-UR 04-8274
†Submitted to *IEEE Transactions on Information Theory* on 12/06/04.

2 Preliminaries

So far, in the machine learning literature only $\mathbb{R}$-valued kernels have been considered. However, to describe the reproducing kernel Hilbert space (RKHS) of Gaussian kernels we will use $\mathbb{C}$-valued kernels and therefore we recall the basic facts on RKHSs for both cases (see e.g. [7], [8] [9], and [10]). To this end let us first recall that for a complex number $z = x + iy \in \mathbb{C}$, $x, y \in \mathbb{R}$, its conjugate is defined by $\bar{z} := x - iy$ and its absolute value is $|z| := \sqrt{z\bar{z}} = \sqrt{x^2 + y^2}$. In particular we have $\bar{x} = x$ and $|x| = \sqrt{x^2}$ for all $x \in \mathbb{R}$. Furthermore, we use the symbol $\mathbb{K}$ whenever we want to treat the real and the complex case simultaneously. For example, a $\mathbb{K}$-Hilbert space is a real Hilbert space when $\mathbb{K} = \mathbb{R}$ and a complex one when $\mathbb{K} = \mathbb{C}$. Recall, that in the latter case the inner product $\langle ., . \rangle$ is sesqui-linear and Hermitian. This fact forces us to be a bit pedantic with the ordering in inner products such as in the following definition.

Definition 2.1 Let X be a non-empty set. Then a function $k : X \times X \to \mathbb{K}$ is called a *kernel* on X if there exists a $\mathbb{K}$-Hilbert space H and a map $\Phi : X \to H$ such that for all $x, x' \in X$ we have

$$k(x, x') \;=\; \langle \Phi(x'), \Phi(x) \rangle \,. \tag{1}$$

We call Φ a *feature map* and H a *feature space* of k.

Note that in the real case condition (1) can be replaced by the well-known equation $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$. In the complex case however, $\langle ., . \rangle$ is Hermitian and hence (1) is equivalent to $k(x, x') = \overline{\langle \Phi(x), \Phi(x') \rangle}$.

Given a kernel neither the feature map nor the feature space are uniquely determined. However, one can always construct a canonical feature space, namely the RKHS. Let us now recall the basic theory of these spaces.

Definition 2.2 Let $X \neq \emptyset$ and H be a Hilbert function space over X, i.e. a Hilbert space which consists of functions mapping from X into $\mathbb{K}$.

i) The space H is called a *reproducing kernel Hilbert space (RKHS)* over X if for all $x \in X$ the *Dirac functional* $\delta_x : H \to \mathbb{K}$ defined by $\delta_x(f) := f(x)$, $f \in H$, is continuous.

ii) A function $k : X \times X \to \mathbb{K}$ is called a *reproducing kernel* of H if we have $k(., x) \in H$ for all $x \in X$ and the *reproducing property*

$$f(x) = \langle f, k(., x) \rangle$$

holds for all $f \in H$ and all $x \in X$.

Recall that reproducing kernel Hilbert spaces have the remarkable and important property that norm convergence implies pointwise convergence. More precisely, let H be a RKHS, $f \in H$, and $(f_n) \subset H$ be a sequence with $\|f_n - f\|_H \to 0$ for $n \to \infty$. Then for all $x \in X$ we have

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \delta_x(f_n) = \delta_x(f) = f(x)\,. \tag{2}$$

Furthermore, reproducing kernels are actually kernels in the sense of Definition 2.1 since $\Phi : X \to H$ defined by $\Phi(x) := k(., x)$ is a feature map of k. Moreover, the reproducing property says that each Dirac functional can be represented by the reproducing kernel. Consequently, a Hilbert function space H that has a reproducing kernel k is always a RKHS. The following theorem shows that conversely, every RKHS has a (unique) reproducing kernel and that this kernel can be determined by the Dirac functionals.

Theorem 2.3 *Let H be a RKHS over X. Then $k : X \times X \to \mathbb{K}$ defined by $k(x, x') := \langle \delta_x, \delta_{x'} \rangle$, $x, x' \in X$, is the only reproducing kernel of H. Furthermore, if $(e_i)_{i \in I}$ is an orthonormal basis (ONB) of H then for all $x, x' \in X$ we have*

$$k(x, x') = \sum_{i \in I} e_i(x) \overline{e_i(x')}\,, \tag{3}$$

where the convergence is absolute.

Theorem 2.3 shows that a RKHS uniquely determines its reproducing kernel. The following theorem states, that conversely every kernel has a unique RKHS.

Theorem 2.4 *Let $X \neq \emptyset$ and k be a kernel over X with feature space H_0 and feature map $\Phi_0 : X \to H_0$. Then*

$$H := \big\{ \langle w, \Phi_0(.) \rangle_{H_0} : w \in H_0 \big\} \tag{4}$$

equipped with the norm

$$\|f\|_H := \inf \big\{ \|w\|_{H_0} : w \in H_0 \text{ with } f = \langle w, \Phi_0(.) \rangle_{H_0} \big\} \tag{5}$$

is the only RKHS of k. In particular both definitions are independent of the choice of H_0 and Φ_0 and the operator $V : H_0 \to H$ defined by

$$Vw := \langle w, \Phi_0(.) \rangle_{H_0}\,, \qquad\qquad w \in H_0$$

is a metric surjection, i.e. $V \overset{\circ}{B}_{H_0} = \overset{\circ}{B}_H$, where $\overset{\circ}{B}_{H_0}$ and $\overset{\circ}{B}_H$ are the open unit balls of H_0 and H, respectively.

3 Results

Before we state our main results we need to recall the definition of the Gaussian RBF kernels. To this end we always denote the j-th component of a complex vector $z \in \mathbb{C}^d$ by z_j. Now let us write

$$k_{\sigma, \mathbb{C}^d}(z, z') := \exp\Big(-\sigma^2 \sum_{j=1}^{d} (z_j - \bar{z}'_j)^2 \Big)$$

for $d \in \mathbb{N}$, $\sigma > 0$, and $z, z' \in \mathbb{C}^d$. Then it can be shown that $k_{\sigma, \mathbb{C}^d}$ is a $\mathbb{C}$-valued kernel on $\mathbb{C}^d$ which we call the *complex Gaussian RBF kernel with width σ*. Furthermore, its restriction $k_\sigma := (k_{\sigma, \mathbb{C}^d})_{|\mathbb{R}^d \times \mathbb{R}^d}$ is an $\mathbb{R}$-valued kernel, which we call the *(real) Gaussian RBF kernel with width σ*. Obviously, this kernel satisfies $k_\sigma(x, x') = \exp(-\sigma^2 \|x - x'\|_2^2)$ for all $x, x' \in \mathbb{R}^d$, where $\|.\|_2$ denotes the Euclidian norm on $\mathbb{R}^d$.

Besides the the Gaussian RBF kernels we also have to introduce a family of spaces. To this end let $\sigma > 0$ and $d \in \mathbb{N}$. For a given holomorphic function $f : \mathbb{C}^d \to \mathbb{C}$ we define

$$\|f\|_{\sigma, \mathbb{C}^d} := \left(\frac{2^d \sigma^{2d}}{\pi^d} \int_{\mathbb{C}^d} |f(z)|^2 e^{\sigma^2 \sum_{j=1}^d (z_j - \bar{z}_j)^2} dz \right)^{1/2},$$

where dz stands for the complex Lebesgue measure on $\mathbb{C}^d$. Furthermore, we write

$$H_\sigma(\mathbb{C}^d) := \big\{ f : \mathbb{C}^d \to \mathbb{C} \,|\, f \text{ holomorphic and } \|f\|_{\sigma, \mathbb{C}^d} < \infty \big\}\,.$$

Obviously, $H_\sigma(\mathbb{C}^d)$ is a complex function space with pre-Hilbert norm $\|.\|_{\sigma, \mathbb{C}^d}$. Let us now state a lemma which will help us to show that $H_\sigma(\mathbb{C}^d)$ is a RKHS. Its proof can be found in Section 4.

Lemma 3.1 *For all $\sigma > 0$ and all compact subsets $K \subset \mathbb{C}^d$ there exists a constant $c_{K,\sigma} > 0$ such that for all $z \in K$ and all $f \in H_\sigma(\mathbb{C}^d)$ we have*

$$|f(z)| \leq c_{K,\sigma} \, \|f\|_{\sigma,\mathbb{C}^d} \, .$$

The above lemma shows that convergence in $\|.\|_{\sigma,\mathbb{C}^d}$ implies *compact convergence*, i.e. uniform convergence on every compact subset. Using the well-known fact from complex analysis that a compactly convergent sequence of holomorphic functions has a holomorphic limit (see e.g. [11, Thm. I.1.9]) we then immediately obtain the announced

Corollary 3.2 *The space $H_\sigma(\mathbb{C}^d)$ with norm $\|.\|_{\sigma,\mathbb{C}^d}$ is a RKHS for every $\sigma > 0$.*

We have seen in Theorem 2.3 that the reproducing kernel of a RKHS is determined by an arbitrary ONB of this RKHS. Therefore, to determine the reproducing kernel of $H_\sigma(\mathbb{C}^d)$ our next aim is to describe an orthonormal basis (ONB). To this end let us recall that the tensor product $f \otimes g : X \times X \to \mathbb{K}$ of two functions $f, g : X \to \mathbb{K}$ is defined by $f \otimes g(x, x') := f(x)g(x')$, $x, x' \in X$. Furthermore, the d-fold tensor product is defined analogously. Now we can formulate the following theorem whose proof can be found in Section 4.

Theorem 3.3 *For $\sigma > 0$ and $n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ we define the function $e_n : \mathbb{C} \to \mathbb{C}$ by*

$$e_n(z) := \sqrt{\frac{(2\sigma^2)^n}{n!}} \, z^n e^{-\sigma^2 z^2}$$

for all $z \in \mathbb{C}$. Then the system $(e_{n_1} \otimes \cdots \otimes e_{n_d})_{n_1,\ldots,n_d \geq 0}$ is an ONB of $H_\sigma(\mathbb{C}^d)$.

We have seen in Theorem 2.3 that an ONB of a RKHS can be used to determine the reproducing kernel. In our case this yields the following theorem whose proof can again be found in Section 4.

Theorem 3.4 *Let $\sigma > 0$ and $d \in \mathbb{N}$. Then the complex Gaussian RBF kernel $k_{\sigma,\mathbb{C}^d}$ is the reproducing kernel of $H_\sigma(\mathbb{C}^d)$.*

With the help of Theorem 3.4 we can now better understand the structure of the RKHSs of real Gaussian RBF kernels. Indeed, using the Definition of k_σ, Theorem 3.4, Theorem 2.4 and the fact that $H_\sigma(\mathbb{C}^d)$ is a feature space of k_σ we immediately obtain the following description of the RKHS of k_σ:

Corollary 3.5 *Let $X \subset \mathbb{R}^d$ and $\sigma > 0$ then the RKHS of the Gaussian RBF kernel k_σ on X is*

$$H_\sigma(X) := \big\{ f : X \to \mathbb{R} \,|\, \exists \hat{f} \in H_\sigma(\mathbb{C}^d) \text{ with } \hat{f}_{|X} = f \big\}$$

and for $f \in H_\sigma$ the norm is given by

$$\|f\|_{\sigma,X} := \inf\big\{ \|\hat{f}\|_{\sigma,\mathbb{C}^d} : \hat{f} \in H_\sigma(\mathbb{C}^d) \text{ with } \hat{f}_{|X} = f \big\} \, .$$

The above corollary shows that one can understand the structure of $H_\sigma(X)$ by the complex extension $k_{\sigma,\mathbb{C}^d}$ of k_σ. In particular, if X is "sufficiently large", e.g. it contains a non-empty open set, then every $f \in H_\sigma(X)$ has a *unique* (see e.g. [11, Rem. I.1.20]) entire extension $\hat{f} \in H_\sigma(\mathbb{C}^d)$ and thus the extension operator $f \mapsto \hat{f}$ is an isometric isomorphism. Furthermore, if $f \in H_\sigma(X)$ is constant on a non-empty open subset of X then its entire extension $\hat{f}$ is constant. However, an easy calculation shows that for non-vanishing constant functions we have $\|\hat{f}\|_{\sigma,\mathbb{C}^d} = \infty$, and therefore we have obtained

Corollary 3.6 *Assume that $X \subset \mathbb{R}^d$ contains an open set $A \neq \emptyset$ and let $f \in H_\sigma(X)$ be a function which is constant on A. Then we have $f(x) = 0$ for all $x \in X$.*

The above corollary shows that for sufficiently large X the space $H_\sigma(X)$ does not contain non-trivial constant functions. In particular we have $\mathbf{1}_X \notin H_\sigma(X)$ for such X.

Let us now determine an ONB of $H_\sigma(X)$ for "sufficiently large" $X \subset \mathbb{R}^d$. To this end we observe that for the functions $e_n : \mathbb{C} \to \mathbb{C}$ defined in Theorem 3.3 we obviously have

$$e_n(x) = \sqrt{\frac{(2\sigma^2)^n}{n!}}\, x^n e^{-\sigma^2 x^2} \in \mathbb{R}$$

for all $x \in \mathbb{R}$ and thus the restriction of e_n onto $\mathbb{R}$ is a real-valued function. Furthermore recall that the extension operator $\hat{}: H_\sigma(X) \to H_\sigma(\mathbb{C}^d)$ is an isometric isomorphism and hence so is its inverse—the restriction operator $\cdot_{|X} : H_\sigma(\mathbb{C}^d) \to H_\sigma(X)$. Therefore we obtain

Corollary 3.7 *Let $X \subset \mathbb{R}^d$ contain a non-empty open subset. For $\sigma > 0$ and $n \in \mathbb{N}_0$ we define the function $e_n : X \to \mathbb{R}$ by*

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

for all $x \in X$. Then the system $(e_{n_1} \otimes \cdots \otimes e_{n_d})_{n_1,\ldots,n_d \geq 0}$ is an ONB of $H_\sigma(X)$.

Finally let us compare the norms $\|.\|_\sigma$ for different values of σ. To this end we first observe that the weight function in the definition of $\|.\|_\sigma$ satisfies

$$e^{\sigma^2 \sum_{j=1}^d (z_j - \bar{z}_j)^2} = e^{-4\sigma^2 \sum_{j=1}^d y_j^2},$$

if $y_j = \operatorname{Im} z_j$, $j = 1, \ldots, d$. This immediately leads to

Corollary 3.8 *Let $X \subset \mathbb{R}^d$ and $0 < \sigma \leq \tau < \infty$. Then $H_\sigma(X) \subset H_\tau(X)$ and for all $f \in H_\sigma(X)$ we have*

$$\|f\|_{\tau,X} \leq \left(\frac{\tau}{\sigma}\right)^d \|f\|_{\sigma,X}\,,$$

i.e. the inclusion $\mathrm{id} : H_\sigma(X) \to H_\tau(X)$ *satisfies* $\|\mathrm{id}\| \leq \left(\frac{\tau}{\sigma}\right)^d$.

Remark 3.9 In this paper it was our aim to describe the RKHSs of the Gaussian RBF kernels with *elementary* methods. However, as observed by Saitoh [10] one can obtain the same results using the so-called Bargmann spaces introduced in [12]. Indeed, [12] shows that these spaces are the RKHSs of the exponential kernels $(z, z') \mapsto \exp(\langle z, \bar{z}' \rangle)$ on $\mathbb{C}^d$, $d \geq 1$, and therefore one can determine the RKHSs of $k_{\sigma,\mathbb{C}^d}$ by using the relation between the exponential and the Gaussian RBF kernels.

It is well known that a RKHS corresponding to a kernel has many different feature spaces and feature maps. Let us now present another feature space and feature map for k_σ which add insight into the nature of the inclusion $\mathrm{id} : H_\sigma(X) \to H_\tau(X)$. As one consequence, we will obtain e.g. a stronger bound than stated in Corollary 3.8 on the norm of the inclusion. Furthermore, it will turn out that for suffiently large X this bound is exact.

In order to formulate all these results let $L_2(\mathbb{R}^d)$ be the space of square-integrable functions on $\mathbb{R}^d$ equipped with the usual norm $\|\cdot\|_2$. Our first result shows that $L_2(\mathbb{R}^d)$ is a feature space of k_σ.

Lemma 3.10 *Let $0 < \sigma < \infty$, $X \subset \mathbb{R}^d$. We define $\Phi_\sigma : X \to L_2(\mathbb{R}^d)$ by*

$$\Phi_\sigma(x) \;:=\; \frac{(2\sigma)^{\frac{d}{2}}}{\pi^{\frac{d}{4}}} e^{-2\sigma^2 \|x - \cdot\,\|_2^2}, \qquad\qquad x \in X.$$

Then $L_2(\mathbb{R}^d)$ is a feature space and $\Phi_\sigma : X \to L_2(\mathbb{R}^d)$ is a feature map of k_σ.

With the help of the above feature space and map we will now present a representation of the inclusion $\mathrm{id} : H_\sigma(X) \to H_\tau(X)$. To this end recall (see e.g. [13]) that for $t > 0$ the Gauss-Weierstraß integral operator $W_t : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$ is defined by

$$W_t g(x) \;:=\; (4\pi t)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{-\frac{\|x-y\|_2^2}{4t}} g(y) dy$$

for all $g \in L_2(\mathbb{R}^d)$, $x \in \mathbb{R}^d$. Now we can formulate the announced result.

Proposition 3.11 *For $0 < \sigma < \tau < \infty$ we define $\delta := \frac{1}{8}\big(\frac{1}{\sigma^2} - \frac{1}{\tau^2}\big)$. Furthermore, let $X \subset \mathbb{R}^d$ and W_δ be as above. Then we obtain a commutative diagram*

$$\begin{array}{ccc} H_\sigma(X) & \xrightarrow{\quad \mathrm{id} \quad} & H_\tau(X) \\ \Big\uparrow{\scriptstyle V_\sigma} & & \Big\uparrow{\scriptstyle V_\tau} \\ L_2(\mathbb{R}^d) & \xrightarrow[\;(\frac{\tau}{\sigma})^{\frac{d}{2}} W_\delta\;]{} & L_2(\mathbb{R}^d) \end{array}$$

where the vertical maps V_σ and V_τ are the metric surjections of Theorem 2.4.

Since V_σ of the above proposition is a metric surjection we obtain $\|\mathrm{id} \circ V_\sigma\| = \|\mathrm{id}\|$, and hence the commutativity of the diagram implies

$$\|\mathrm{id} : H_\sigma(X) \to H_\tau(X)\| = \|\mathrm{id} \circ V_\sigma\| = \Big(\frac{\tau}{\sigma}\Big)^{\frac{d}{2}} \|V_\tau \circ W_\delta\| \le \Big(\frac{\tau}{\sigma}\Big)^{\frac{d}{2}} \|W_\delta\|.$$

However it is well known (see e.g. [13]) that $\|W_\delta\| \le 1$. Therefore we have established the following corollary.

Corollary 3.12 *Let $X \subset \mathbb{R}^d$ and $0 < \sigma \le \tau < \infty$. Then we have*

$$\|\mathrm{id} : H_\sigma(X) \to H_\tau(X)\| \le \Big(\frac{\tau}{\sigma}\Big)^{\frac{d}{2}}.$$

Our last result which is proved in Section 4 shows that for sufficiently large X the metric surjections $V_\sigma : L_2(\mathbb{R}^d) \to H_\sigma(X)$ are isometric isomorphisms, and consequently $\mathrm{id} : H_\sigma(X) \to H_\tau(X)$ shares many important properties with W_δ.

Corollary 3.13 *Let $X \subset \mathbb{R}^d$ contain a non-empty open subset. Then $V_\sigma : L_2(\mathbb{R}^d) \to H_\sigma(X)$ is an isometric isomorphism for all $\sigma > 0$. In addition, for all $0 < \sigma < \tau < \infty$ and $\delta := \frac{1}{8}\big(\frac{1}{\sigma^2} - \frac{1}{\tau^2}\big)$ we* *have the following cummutative diagram*

$$
\begin{array}{ccc}
H_\sigma(X) & \xrightarrow{\ \mathrm{id}\ } & H_\tau(X) \\
{\scriptstyle V_\sigma^{-1}}\big\downarrow & & \big\uparrow{\scriptstyle V_\tau} \\
L_2(\mathbb{R}^d) & \xrightarrow[(\frac{\tau}{\sigma})^{\frac{d}{2}} W_\delta]{} & L_2(\mathbb{R}^d)
\end{array}
$$

and consequently the following statements are true:

i) $\mathrm{id} : H_\sigma(X) \to H_\tau(X)$ *is not compact.*

ii) $\mathrm{id} : H_\sigma(X) \to H_\tau(X)$ *is not surjective, i.e.* $H_\sigma(X) \subsetneq H_\tau(X)$*.*

iii) The estimate of Corollary 3.12 is exact, i.e. we have

$$\|\mathrm{id} : H_\sigma(X) \to H_\tau(X)\| = \Big(\frac{\tau}{\sigma}\Big)^{\frac{d}{2}} .$$

Finally, let us briefly discuss how the above result can be used in the analysis of support vector machines (see [2] for these learning algorithms). For the sake of simplicity we only consider the support vector machines (SVMs) with Gaussian RBF kernels and with hinge loss $L(y,t) := \max\{0, 1-yt\}$, $y \in Y := \{-1,1\}$, $t \in \mathbb{R}$, which are used for binary classification problems (see [14] for an introduction to classification). Let $X \subset \mathbb{R}^d$ be as in the above corollary and P be a probability measure on $X \times Y$. Then for a measurable $f : X \to \mathbb{R}$ we define the *L-risk* by

$$\mathcal{R}_{L,P}(f) := \int_{X\times Y} L(y, f(x))dP(x,y) .$$

Furthermore, the *minimal L-risk* is denoted by $\mathcal{R}^*_{L,P} := \inf_f \mathcal{R}_{L,P}(f)$, where the infimum runs over all measurable functions. Now, it has recently been discovered that for analyzing the learning performance of SVMs the behaviour of the *approximation error function*

$$a_\sigma(\lambda) := \inf_{f\in H_\sigma(X)} \lambda\|f\|^2_{\sigma,X} + \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \qquad (6)$$

for $\lambda \to 0$ plays an important role. Indeed, $a_\sigma(\lambda) \to 0$ for $\lambda \to 0$ was used in [15] to show that SVMs can learn in the sense of universal consistency (see [14] for an introduction to this notion of learning). Furthermore, [16], [17] and [6] established small bounds on $a_\sigma(\lambda)$ for certain P, σ and λ which were used for stronger guarantees on the learning performance of SVMs. Unfortunately, the used techniques are rather involved and in particular it is completely open whether the obtained bounds are sharp. Now, Corollary 3.13 shows

$$a_\sigma(\lambda) = \inf_{g\in L_2(\mathbb{R}^d)} \lambda\|g\|^2_{L_2(\mathbb{R}^d)} + \mathcal{R}_{L,P}(V_\sigma g) - \mathcal{R}^*_{L,P} , \qquad (7)$$

which may significantly help in understanding the behaviour of $a_\sigma(\lambda)$. Indeed, in order to establish a small bound of $a_\sigma(\lambda)$ via (6) one has to simultaneously control both the shape and the $\|\,.\,\|_{\sigma,X}$-norm of certain $f \in H_\sigma(X)$ which is rather challenging because of the analyticity of these f. In contrast to this, we see that when considering (7) the task is to simultaneously control $\|g\|_{L_2(\mathbb{R}^d)}$

and the shape of $V_\sigma g$ for suitable $g \in L_2(\mathbb{R}^d)$. Obviously, the first term is easy to determine for many g and the second term can be investigated by e.g. the well-established theory of the Gauss-Weierstraß integral operator, or more generally, convolution operators. Remarkably, this approach was already used implicitly in the proof of [6], however arising technical difficulties in [6] make it hard to see the simple structure there. We hope that by outlining (7) and its usability the existing bounds on $a_\sigma(\lambda)$ can be further improved.

4 Proofs

For the proof of Lemma 3.1 we need the following technical lemma.

Lemma 4.1 *For all $d \in \mathbb{N}$, all holomorphic functions $f : \mathbb{C}^d \to \mathbb{C}$, all $r_1, \dots, r_d > 0$, and all $z \in \mathbb{C}^d$ we have*

$$|f(z)|^2 \le \frac{1}{(2\pi)^d} \int_0^{2\pi} \cdots \int_0^{2\pi} \left| f(z_1 + r_1 e^{i\theta_1}, \dots, z_d + r_d e^{i\theta_d}) \right|^2 d\theta_1 \cdots d\theta_d \,. \tag{8}$$

Proof: We proceed by induction over d. For $d = 1$ the assertion follows from Hardy's convexity theorem (see e.g. [18, p. 9]) which states that the function

$$r \mapsto \frac{1}{2\pi} \int_0^{2\pi} |f(z + re^{i\theta})|^2 d\theta$$

is non-decreasing on $[0, \infty)$.

Now let us suppose that we have already shown the assertion for $d \in \mathbb{N}$. Let $f : \mathbb{C}^{d+1} \to \mathbb{C}$ be a holomorphic function, and choose $r_1, \dots, r_{d+1} > 0$. Since for fixed $(z_1, \dots, z_d) \in \mathbb{C}^d$ the function $z_{d+1} \mapsto f(z_1, \dots, z_d, z_{d+1})$ is holomorphic by the induction hypothesis, we obtain

$$|f(z_1, \dots, z_{d+1})|^2 \le \frac{1}{2\pi} \int_0^{2\pi} |f(z_1, \dots, z_d, z_{d+1} + r_{d+1} e^{i\theta_{d+1}})|^2 d\theta_{d+1} \,.$$

Now applying the induction hypothesis to the holomorphic function $(z_1, \dots, z_d) \mapsto f(z_1, \dots, z_d, z_{d+1} + r_{d+1} e^{i\theta_{d+1}})$ on $\mathbb{C}^d$ gives the assertion for $d + 1$. ■

Proof of Lemma 3.1: Let us define $c := \max\{e^{-\sigma^2 \sum_{j=1}^d (z_j - \bar{z}_j)^2} : (z_1, \dots, z_d) \in K + (B_\mathbb{C})^d\}$, where $B_\mathbb{C}$ denotes the closed unit ball of $\mathbb{C}$. Now, by Lemma 4.1 we have

$$2^d r_1 \cdots r_d |f(z)|^2 \le \frac{r_1 \cdots r_d}{\pi^d} \int_0^{2\pi} \cdots \int_0^{2\pi} \left| f(z_1 + r_1 e^{i\theta_1}, \dots, z_d + r_d e^{i\theta_d}) \right|^2 d\theta_1 \cdots d\theta_d$$

and integrating this inequality with respect to $r = (r_1, \dots, r_d)$ over $[0, 1]^d$ then yields

$$\begin{aligned} |f(z)|^2 &\le \frac{1}{\pi^d} \int\limits_{z+(B_\mathbb{C})^d} |f(z')|^2 dz' \\ &\le \frac{c}{\pi^d} \int\limits_{z+(B_\mathbb{C})^d} |f(z')|^2 e^{\sigma^2 \sum_{j=1}^d (z_j - \bar{z}_j)^2} dz' \\ &\le \frac{c}{(2\sigma^2)^d} \, \|f\|^2_{\sigma, \mathbb{C}^d} \,. \end{aligned}$$

■

For the proof of Theorem 3.3 we need the following technical lemma.

Lemma 4.2 *For all $n, m \in \mathbb{N}_0$ and all $\sigma > 0$ we have*

$$\int_{\mathbb{C}} z^n (\bar{z})^m e^{-2\sigma^2 z\bar{z}} dz = \begin{cases} \frac{\pi\, n!}{(2\sigma^2)^{n+1}} & \text{if } n = m \\ 0 & \text{otherwise.} \end{cases}$$

Proof: Let us first consider the case $n = m$. Then we have

$$\begin{aligned} \int_{\mathbb{C}} z^n (\bar{z})^m e^{-2\sigma^2 z\bar{z}} dz &= \int_0^\infty \int_0^{2\pi} r^{2n} e^{-2\sigma^2 r^2} d\theta\, r\, dr \\ &= 2\pi \int_0^\infty r^{2n+1} e^{-2\sigma^2 r^2} dr \\ &= \frac{\pi}{(2\sigma^2)^{n+1}} \int_0^\infty t^n e^{-t} dt \\ &= \frac{\pi\, n!}{(2\sigma^2)^{n+1}}\,. \end{aligned}$$

Now let us assume $n \neq m$. Then we obtain

$$\int_{\mathbb{C}} z^n (\bar{z})^m e^{-2\sigma^2 z\bar{z}} dz = \int_0^\infty r \int_0^{2\pi} r^{n+m} e^{i(n-m)\theta} e^{-2\sigma^2 r^2} d\theta dr = 0\,.$$

■

Proof of Theorem 3.3: In order to avoid cumbersome technical notations that hide the structure of the proof we first consider the case $d = 1$.

Let us show that $(e_n)_{n\geq 0}$ is an orthonormal system. To this end for $n, m \in \mathbb{N}_0$, and $z \in \mathbb{C}$ we observe

$$e_n(z)\overline{e_m(z)} e^{\sigma^2(z-\bar{z})^2} = \sqrt{\frac{(2\sigma^2)^{n+m}}{n!\,m!}} z^n (\bar{z})^m e^{-\sigma^2 z^2 - \sigma^2 \bar{z}^2} e^{\sigma^2(z-\bar{z})^2} = \sqrt{\frac{(2\sigma^2)^{n+m}}{n!\,m!}} z^n (\bar{z})^m e^{-2\sigma^2 z\bar{z}}\,.$$

Therefore for $n, m \geq 0$ we obtain

$$\langle e_n, e_m \rangle = \frac{2\sigma^2}{\pi} \int\limits_{\mathbb{C}} e_n(z)\overline{e_m(z)} e^{\sigma^2(z-\bar{z})^2} dz = \frac{2\sigma^2}{\pi} \cdot \sqrt{\frac{(2\sigma^2)^{n+m}}{n!\,m!}} \int\limits_{\mathbb{C}} z^n (\bar{z})^m e^{-2\sigma^2 z\bar{z}} dz = \begin{cases} 1 & \text{if } n = m \\ 0 & \text{otherwise} \end{cases}$$

by Lemma 4.2. This shows that $(e_n)_{n\geq 0}$ is indeed an orthonormal system.

Now, let us show that this system is also complete. To this end let $f \in H_\sigma(\mathbb{C})$. Then $z \mapsto e^{\sigma^2 z^2} f(z)$ is an entire function, and therefore there exists a sequence $(a_n) \subset \mathbb{C}$ such that

$$f(z) = \sum_{n=0}^\infty a_n z^n e^{-\sigma^2 z^2} = \sum_{n=0}^\infty a_n \sqrt{\frac{n!}{(2\sigma^2)^n}}\, e_n(z) \tag{9}$$

for all $z \in \mathbb{C}$. Obviously, it suffices to show that the above convergence also holds with respect to $\|.\|_{\sigma,\mathbb{C}}$. To prove this we first recall from complex analysis that the series in (9) converges absolutely and compactly. Therefore for $n \geq 0$ Lemma 4.2 yields

$$
\begin{aligned}
\langle f, e_n \rangle &= \frac{2\sigma^2}{\pi} \int\limits_{\mathbb{C}} f(z)\overline{e_n(z)} e^{\sigma^2(z-\bar{z})^2} dz \\
&= \frac{2\sigma^2}{\pi} \sum_{m=0}^{\infty} a_m \int\limits_{\mathbb{C}} z^m e^{-\sigma^2 z^2} \overline{e_n(z)} e^{\sigma^2(z-\bar{z})^2} dz \\
&= \frac{2\sigma^2}{\pi} \sqrt{\frac{(2\sigma^2)^n}{n!}} \sum_{m=0}^{\infty} a_m \int\limits_{\mathbb{C}} z^m (\bar{z})^n e^{-2\sigma^2 z\bar{z}} dz \\
&= a_n \sqrt{\frac{n!}{(2\sigma^2)^n}} \, . \qquad (10)
\end{aligned}
$$

Furthermore, since (e_n) is an orthonormal system we have $(\langle f, e_n \rangle) \in \ell_2$ by Bessel's inequality. Using again that (e_n) is an orthonormal system in $H_\sigma(\mathbb{C})$ we hence find a function $g \in H_\sigma(\mathbb{C})$ with $g = \sum_{n=0}^{\infty} \langle f, e_n \rangle e_n$, where the convergence take place in $H_\sigma(\mathbb{C})$. Now, using (9), (10), and the fact that norm convergence in RKHSs implies point-wise convergence we find $g = f$, i.e. the series in (9) with respect to $\|.\|_{\sigma,\mathbb{C}}$.

Now, let us briefly treat the general, d-dimensional case. In this case the orthonormality of $(e_{n_1} \otimes \cdots \otimes e_{n_d})_{n_1,\ldots,n_d \geq 0}$ follows from

$$
\langle e_{n_1} \otimes \cdots \otimes e_{n_d}, e_{m_1} \otimes \cdots \otimes e_{m_d} \rangle = \prod_{j=1}^{d} \langle e_{n_j}, e_{m_j} \rangle \, .
$$

Furthermore, the completeness can be shown analogously to the 1-dimensional case using a d-dimensional Taylor series (see e.g. [11, Thm. I.1.18]) instead of (9). ■

Proof of Theorem 3.4: Let k be the reproducing kernel of $H_{\sigma,\mathbb{C}^d}$. Then using the ONB of Theorem 3.3 we obtain

$$
\begin{aligned}
k(z, z') &= \sum_{n_1,\ldots,n_d=0}^{\infty} e_{n_1} \otimes \cdots \otimes e_{n_d}(z) \overline{e_{n_1} \otimes \cdots \otimes e_{n_d}(z')} \\
&= \sum_{n_1,\ldots,n_d=0}^{\infty} \prod_{j=1}^{d} \frac{(2\sigma^2)^{n_j}}{n_j!} (z\bar{z}')^{n_j} e^{-\sigma^2 z_j^2 - \sigma^2 (\bar{z}_j')^2} \\
&= \prod_{j=1}^{d} \sum_{n_j=0}^{\infty} \frac{(2\sigma^2)^{n_j}}{n_j!} (z\bar{z}')^{n_j} e^{-\sigma^2 z_j^2 - \sigma^2 (\bar{z}_j')^2} \\
&= \prod_{j=1}^{d} e^{-\sigma^2 z_j^2 - \sigma^2 (\bar{z}_j')^2 + 2\sigma^2 z_j \bar{z}_j'} \\
&= e^{-\sigma^2 \sum_{j=1}^{d} (z_j - \bar{z}_j')^2} \, ,
\end{aligned}
$$

which shows the assertion. ■

Proof of Lemma 3.10: We beging by collecting some well known facts about manipulating Gaussians that are useful in proving Lemma 3.10, Theorem 4.3, and Corollary 3.13. First it is well known that for all $t > 0$ and $x \in \mathbb{R}^d$ we have

$$\int_{\mathbb{R}^d} e^{-\frac{\|y-x\|_2^2}{t}} dy = (\pi t)^{\frac{d}{2}} . \tag{11}$$

Second, an elementary calculation shows that for any $\alpha \geq 0$ that

$$\|y-x\|_2^2 + \alpha \|y-x'\|_2^2 = \frac{\alpha}{1+\alpha}\|x-x'\|_2^2 + (1+\alpha)\left\|y - \frac{x+\alpha x'}{1+\alpha}\right\|_2^2 \tag{12}$$

for all $y, x, x' \in \mathbb{R}^d$. Now by using (11) and setting $\alpha := 1$ in (12) we obtain

$$\begin{aligned}
\langle \Phi_\sigma(x), \Phi_\sigma(x')\rangle_{L_2(\mathbb{R}^d)} &= \frac{(2\sigma)^d}{\pi^{\frac{d}{2}}} \int_{\mathbb{R}^d} e^{-2\sigma^2\|x-z\|_2^2} e^{-2\sigma^2\|x'-z\|_2^2} dz \\
&= \frac{(2\sigma)^d}{\pi^{\frac{d}{2}}} e^{-\sigma^2\|x-x'\|_2^2} \int_{\mathbb{R}^d} e^{-4\sigma^2\|z-\frac{x+x'}{2}\|_2^2} dz \\
&= \frac{(2\sigma)^d}{\pi^{\frac{d}{2}}} \cdot e^{-\sigma^2\|x-x'\|_2^2} \left(\frac{\pi}{4\sigma^2}\right)^{\frac{d}{2}} \\
&= k_\sigma(x,x').
\end{aligned}$$

Therefore Φ_σ is a feature map and $L_2(\mathbb{R}^d)$ is a feature space of k_σ. ∎

Proof of Proposition 3.11: Theorem 2.4 shows that we can compute the metric surjection $V_\sigma : L_2(\mathbb{R}^d) \to H_\sigma(X)$ by

$$V_\sigma g(x) = \langle g, \Phi_\sigma(x)\rangle_{L_2(\mathbb{R}^d)} = \frac{(2\sigma)^{\frac{d}{2}}}{\pi^{\frac{d}{4}}} \int_{\mathbb{R}^d} e^{-2\sigma^2\|x-y\|_2^2} g(y) dy \,, \qquad g \in L_2(\mathbb{R}^d),\ x \in X \,,$$

where Φ_σ is the feature map defined in Lemma 3.10. Note, that in this formula the *computation* of V_σ is independent of the chosen domain X. Therefore let us first consider the case $X = \mathbb{R}^d$. Then the relationships

$$V_\sigma = \left(\frac{\pi}{\sigma^2}\right)^{\frac{d}{4}} W_{\frac{1}{8\sigma^2}} \qquad \text{and} \qquad W_{\frac{1}{8\tau^2}} = \left(\frac{\tau^2}{\pi}\right)^{\frac{d}{4}} V_\tau$$

are easily derived. Furthermore it is well known (see e.g. Hille and Phillips [13]) that the Gauss-Weierstraß integral operator corresponds to a solution of the heat equation and so satisfies the semigroup identity

$$W_s = W_t W_{s-t}$$

for all $0 < t < s$. Combining this with the relations between the operators W_t and V_σ we obtain

$$V_\sigma = \left(\frac{\pi}{\sigma^2}\right)^{\frac{d}{4}} W_{\frac{1}{8\sigma^2}} = \left(\frac{\pi}{\sigma^2}\right)^{\frac{d}{4}} W_{\frac{1}{8\tau^2}} W_{\frac{1}{8}\left(\frac{1}{\sigma^2}-\frac{1}{\tau^2}\right)} = \left(\frac{\tau}{\sigma}\right)^{\frac{d}{2}} V_\tau W_{\frac{1}{8}\left(\frac{1}{\sigma^2}-\frac{1}{\tau^2}\right)} \tag{13}$$

for all $0 < \sigma < \tau$, and thus the diagram commutes in the case of $X = \mathbb{R}^d$. The general case $X \subset \mathbb{R}^d$ follows from (13) using the fact that the *computation* of V_σ is independent of X. ∎

For the proof of Corollary 3.13 we have to recall the following important theorem which for completeness is proved below.

Theorem 4.3 *The Gauss-Weierstraß integral operator* $W_t : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$ *is not compact for all* $t > 0$*.*

Proof: Let $\mathbb{Z}^d$ be the lattice of integral vectors in $\mathbb{R}^d$. For $n = (n_1, ..., n_d) \in \mathbb{Z}^d$ and $s > 0$ we define

$$g_n^{(s)}(x) := (2\pi s)^{-\frac{d}{4}} e^{-\frac{\|x-n\|^2}{4s}}, \qquad x \in \mathbb{R}^d.$$

Then (11) shows $\|g_n^{(s)}\|_2^2 = 1$, i.e. $g_n^{(s)}$ is contained in the closed unit ball $B_{L_2(\mathbb{R}^d)}$ of $L_2(\mathbb{R}^d)$. Furthermore from (11) and (12) we infer

$$W_t g_n^{(s)}(x) = (4\pi t)^{-\frac{d}{2}} (2\pi s)^{-\frac{d}{4}} \int_{\mathbb{R}^d} e^{-\frac{\|x-y\|_2^2}{4t}} e^{-\frac{\|y-n\|_2^2}{4s}} dy = (2\pi s)^{-\frac{d}{4}} \Big(\frac{s}{s+t}\Big)^{\frac{d}{2}} e^{-\frac{\|x-n\|_2^2}{4(s+t)}}.$$

Consequently, by utilizing (11) and (12) yet again, we obtain for $n, m \in \mathbb{Z}^d$ that

$$\langle W_t g_n^{(s)}, W_t g_m^{(s)} \rangle = (2\pi s)^{-\frac{d}{2}} \Big(\frac{s}{s+t}\Big)^{d} \int_{\mathbb{R}^d} e^{-\frac{\|y-n\|_2^2}{4(s+t)}} e^{-\frac{\|y-m\|_2^2}{4(s+t)}} dy = \Big(\frac{s}{s+t}\Big)^{\frac{d}{2}} e^{-\frac{\|m-n\|_2^2}{8(s+t)}}. \tag{14}$$

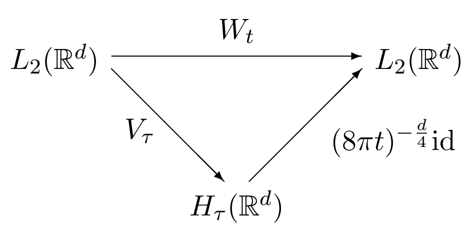
Therefore for $n \neq m \in \mathbb{Z}^d$ and $s := t$ we have

$$\begin{aligned} \|W_t g_n^{(t)} - W_t g_m^{(t)}\|_2^2 = \|W_t g_n^{(t)}\|_2^2 + \|W_t g_m^{(t)}\|_2^2 - 2\langle W_t g_n^{(t)}, W_t g_m^{(t)} \rangle \; &= \; 2^{1-\frac{d}{2}} \Big(1 - e^{-\frac{\|m-n\|_2^2}{16t}}\Big) \\ &\geq \; 2^{1-\frac{d}{2}} (1 - e^{-\frac{1}{16t}}), \end{aligned}$$

and hence $\{W_t g_n^{(t)} : n \in \mathbb{Z}^d\} \subset W_t B_{L_2(\mathbb{R}^d)}$ is not precompact. This implies the assertion. ■

Proof of Corollary 3.13: Let us first show that V_σ is an isometric isomorphism. In view of Theorem 2.4 it suffices to prove that V_σ is injective. To this end let $g \in L_2(\mathbb{R}^d)$ with $V_\sigma g = 0$. Since X contains an open subset the analytic extension $\hat{V}_\sigma g : \mathbb{R}^d \to \mathbb{R}$ of $V_\sigma g$ also satisfies $\hat{V}_\sigma g = 0$. Now, the unique continuation property of Itô and Yamabe [19] for the heat equation implies $g = 0$, and hence V_σ is injective by its linearity. Obviously, the asserted diagram is an immediate consequence of the injectivity of V_σ and the diagram in Proposition 3.11.

Now the remaining assertions can be shown by the established diagram. Indeed, *i)* follows from Theorem 4.3. Beckner's [20] work on sharp Young's inequalities implies $\|W_\delta\| = 1$ which establishes *iii)* but the result can be easily obtained from (14). Indeed, the latter implies $\|W_\delta g_n\|_2 = (\frac{s}{s+t})^{\frac{d}{4}}$ and we obtain *iii)* by letting $s \to \infty$. Finally, by considering the case $X = \mathbb{R}^d$ we note that for $t > 0$ and $\tau := \frac{1}{\sqrt{8t}}$ we have $W_t = (8\pi t)^{-\frac{d}{4}} V_\tau$, i.e. we have the following commutative diagram



Now, since $H_\tau(\mathbb{R}^d)$ consists of analytic functions we obviously have $H_\tau(\mathbb{R}^d) \subsetneq L_2(\mathbb{R}^d)$ and hence W_t is not surjective. ■

References

[1] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[2] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.

[3] N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[4] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, National Taiwan University, 2003.

[5] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93, 2001.

[6] C. Scovel and I. Steinwart. Fast rates for support vector machines. *Ann. Statist.*, submitted, 2003. `http://www.c3.lanl.gov/~ingo/publications/ann-03.ps`.

[7] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[8] H. Meschkowski. *Hilbertsche Räume mit Kernfunktion*. Springer, Berlin, 1962.

[9] E. Hille. Introduction to general theory of reproducing kernels. *Rocky Mt. J. Math.*, 2:321–368, 1972.

[10] S. Saitoh. *Integral transforms, reproducing kernels and their applications*. Longman Scientific & Technical, Harlow, 1997.

[11] R.M. Range. *Holomorphic Functions and Integral Representations in Several Complex Variables*. Springer, 1986.

[12] V. Bargmann. On a Hilbert space of analytic functions and an associated integral transform, part 1. *Comm. Pure Appl. Math.*, 14:187–214, 1961.

[13] E. Hille and R. S. Phillips. *Functional Analysis and Semi-groups*. American Mathematical Society Colloquium Publications Vol. XXXI, Providence, revised edition, 1957.

[14] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[15] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. Inform. Theory*, to appear, 2005.

[16] Q. Wu and D.-X. Zhou. Analysis of support vector machine classification. Tech. Report, City University of Hong Kong, 2003.

[17] S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl.*, 1:17–41, 2003.

[18] P.L. Duren. *Theory of H^p spaces*. Academic Press, 1970.

[19] S. Itô and H. Yamabe. A unique continuation theorem for solutions of a parabolic differential equation. *J. Math. Soc. Japan*, 10:314–321, 1958.

[20] W. Beckner. Inequalities in Fourier analysis. *Ann. of Math.*, 102:159–182, 1975.