LA-UR-09-01900

Approved for public release;
distribution is unlimited.

Title: Radial kernels and their Reproducing Kernel Hilbert Spaces

Author(s): Clint Scovel, Don Hush, Ingo Steinwart, and James Theiler

Intended for: Journal of Complexity

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Form 836 (7/06)

This page intentionally left blank.

Radial kernels and their Reproducing Kernel Hilbert Spaces

Clint Scovel, Don Hush, Ingo Steinwart, and James Theiler

Los Alamos National Laboratory

jcs@lanl.gov
dhush@lanl.gov
ingo@lanl.gov
jt@lanl.gov

March 23, 2010

Abstract

We describe how to use Schoenberg's theorem for a radial kernel combined with existing bounds on the approximation error functions for Gaussian kernels to obtain a bound on the approximation error function for the radial kernel. The result is applied to the exponential kernel and Student's kernel. To establish these results we develop a general theory regarding mixtures of kernels. We analyze the reproducing kernel Hilbert space (RKHS) of the mixture in terms of the RKHS's of the mixture components and prove a type of Jensen inequality between the approximation error function for the mixture and the approximation error functions of the mixture components.

1 Introduction

Gaussian kernels have been popular in Learning Theory for some time. However it is only recently that they have been shown to allow efficient learning. For example, Steinwart et. al. [32, 33, 30, 31] show that one can achieve fast learning rates with the Gaussian kernels. See [29] for a more complete history. Moreover, efficient learning algorithms have been developed for arbitrary kernels in e. g. [17, 16, 20]. However, Gaussian kernels can suffer numerically in practice when the underlying space is large or the kernel parameter t is large since the function $e^{-t^2\|x-x'\|^2}$ may be evaluated by the computer as having only values 0 and 1. Consequently, other radial kernels such as $e^{-\alpha\|x-x'\|}$ or $(1+m^{-1}\|x-x'\|_2^2)^{-\alpha}$ are often used. However, the above mentioned analysis of learning rates has yet to be developed for these kernels. One reason for this is that we have no good bounds on their approximation error properties. In this paper, we will in particular provide bounds on the so-called approximation error functions, defined in the papers mentioned above, for a large class of radial kernels which includes the above examples.

In practice it appears advantageous to have radial functions which are kernels independent of the dimension d of the underlying space $\mathbb{R}^d$. Due to theorems of Bernstein [8, 35], Bochner [9], Schoenberg [25], and Moore [22], this set of kernels, which we denote by $\mathcal{K}_{rad}$, corresponds to the

set of finite Borel measures on $\mathbb{R}^+$ through an integral representation in terms of Gaussian kernels. That is, denote $\mathbb{R}^+ := [0, \infty)$ and let $k_t(x, x') = e^{-t^2\|x-x'\|^2}, t \in \mathbb{R}^+$ denote the family of Gaussian kernels. Then $k \in \mathcal{K}_{rad}$ if and only if there is a finite Borel measure μ on $\mathbb{R}^+$ such that for all $d \geq 1$ we have

$$k(x, x') = \int_{\mathbb{R}^+} k_t(x, x') d\mu(t), \quad x, x' \in \mathbb{R}^d . \tag{1}$$

See Theorem 1.1 below for a precise statement. Henceforth we will use the term "radial kernel" to refer to elements of $\mathcal{K}_{rad}$. Micchelli et. al. [21] have used the integral representation (1) to show that all nonconstant radial kernels are universal for all compact subsets, in the sense that their RKHSs are dense in the Banach space $C(X)$ of continuous functions. However, to obtain learning rates Steinwart et. al. [32, 33, 30, 31] utilized the important concept of the approximation error function $A_k(\lambda)$ corresponding to the kernel k defined as follows: Let $\mathcal{R}$ be a continuous convex function on the reproducing kernel Hilbert space H_k associated with the kernel k and define the regularized functions by $\mathcal{R}_{\lambda,k}(f) := \lambda\|f\|^2_{H_k} + \mathcal{R}(f)$, $\lambda \geq 0$. Also let $\mathcal{R}^*_{\lambda,k} := \inf_{f\in H_k} \mathcal{R}_{\lambda,k}(f)$ denote their minimum values (i. e. greatest lower bounds). The approximation error function defined by $A_k(\lambda) := \mathcal{R}^*_{\lambda,k} - \mathcal{R}^*_{0,k}$ measures how minimizing the regularized function $\mathcal{R}_{\lambda,k}$ approximately minimizes the function $\mathcal{R} = \mathcal{R}_{0,k}(f)$. Now suppose that we consider a radial kernel k and ask how the representation (1) can be used to provide bounds for its approximation error function $A_k(\lambda)$ in terms of bounds on the approximation error functions $A_{k_t}(\lambda)$ for the Gaussian kernels and the measure μ. Indeed, our main result Corollary 3.5 is that for a radial kernel $k = \int k_t d\mu(t)$ we have

$$A_k(\lambda) \leq \int_{\mathbb{R}^+} A_{k_t}(\lambda) d\mu(t) , \quad \lambda \geq 0. \tag{2}$$

Using existing bounds on the approximation error functions for Gaussian kernels, this result is then used to obtain bounds on the approximation error functions for the two radial kernels mentioned above.

Most of the results we present are relatively easy to obtain for finite sums of kernels. However, obtaining them for radial kernels using the integral representation (1) requires that a large part of the paper is concerned with the technical issues of measure and integration theory. To prove the main result (2) we first consider how can we represent the reproducing kernel Hilbert space (RKHS) H_k of the kernel $k = \int k_t d\mu(t)$ in terms of the Gaussian RKHSs $H_{k_t}, t \geq 0$ and the representing measure μ. Recall that [3] shows that if $k = k_1 + k_2$ is the sum of two kernels on X that k is a kernel and its corresponding RKHS H_k has the representation $H_k = \{f_1 + f_2 | f_1 \in H_{k_1},\, f_2 \in H_{k_2}\}$ with norm defined by

$$\|f\|^2_{H_k} = \inf_{\substack{f=f_1+f_2 \\ f_1\in H_{k_1},\, f_2\in H_{k_2}}} \Big(\|f_1\|^2_{H_{k_1}} + \|f_2\|^2_{H_{k_2}}\Big).$$

In addition, it is easy to show that for $\alpha > 0$ we have that $H_{\alpha k} = H_k$ and that $\|\alpha f\|^2_{H_{\alpha k}} = \alpha\|f\|^2_{H_k}$ so that for all $\alpha \in [0, 1]$ we have

$$\|f\|^2_{H_{\alpha k_1+(1-\alpha)k_2}} = \inf_{\substack{f=\alpha f_1+(1-\alpha)f_2 \\ f_1\in H_{k_1},\, f_2\in H_{k_2}}} \Big(\alpha\|f_1\|^2_{H_{k_1}} + (1-\alpha)\|f_2\|^2_{H_{k_2}}\Big), \quad f \in H_{\alpha k_1+(1-\alpha)k_2} , \tag{3}$$

suggesting that integral versions of these representations may be available.

Now let k_1 and k_2 be two kernels and let $\mathcal{R}$ be a continuous convex function on $H_{k_1} + H_{k_2}$ such that $\inf_{H_{k_1}} \mathcal{R} = \inf_{H_{k_2}} \mathcal{R}$. Then we can show

$$A_{\alpha k_1+(1-\alpha)k_2}(\lambda) \leq \alpha A_{k_1}(\lambda) + (1-\alpha)A_{k_2}(\lambda). \tag{4}$$

That is, in a certain sense, the function $k \mapsto A_k(\lambda)$ is a convex function.

Inequalities (3) and (4) suggest the existence of integral versions of these inequalities, which may then be used to analyze radial kernels, As a consequence of a general theory developed in this paper, Corollary 3.5 shows that we do indeed posses the desired integral inequalities for radial kernels. Roughly stated, if k is a radial kernel and μ is its representing measure so that $k = k_\mu := \mathbb{E}_{t\sim\mu} k_t$ where $\mathbb{E}_{t\sim\mu}$ denotes integration we obtain $H_{k_\mu} = \{\mathbb{E}_{t\sim\mu} f_t,\ f_t \in H_{k_t}, \forall t \in T\}$,

$$\|f\|_{k_\mu}^2 = \inf_{\substack{f=\mathbb{E}_{t\sim\mu} f_t \\ f_t \in H_{k_t}, \forall t \in T}} \mathbb{E}_{t\sim\mu} \|f_t\|_{k_t}^2,$$

and the approximation error function inequality (2).

Before we proceed, we follow [29] to fix terminology, set notation, and formally state the integral representation theorem we use for radial kernels. Let X be a nonempty set. Then a bivariate function $k : X \times X \to \mathbb{R}$ will be called a *kernel* if there exists a Hilbert space H and a map $\Phi : X \to H$ such that, for all $x, x' \in X$, we have $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$. H is called a feature space and Φ is called a feature map for k. Moreover, a Hilbert space H of real-valued functions on X is called the reproducing kernel Hilbert space (RKHS) corresponding to a bivariate function $k : X \times X \to \mathbb{R}$ if $k(\cdot, x) \in H$ for all $x \in X$, and we have the reproducing property $f(x) = \langle f, k(\cdot, x) \rangle$ for all $f \in H$ and $x \in X$. It is well known (see e.g. [29, Ch. 4]) that there exists a bijection between kernels and RKHSs although a kernel has many feature spaces in general. We denote the RKHS associated to the kernel k by H_k. Let us denote by $\mathbb{E}_{t\sim\mu}$ the process of integration with respect to the measure μ over a measurable space T. Moreover, for kernels, $\mathbb{E}_{t\sim\mu} k_t$ means that the integration is defined pointwise by $k_\mu(x, x') := \mathbb{E}_{t\sim\mu} k_t(x, x')$, $x, x' \in X$. A function $g : \mathbb{R}^+ \to \mathbb{R}$ is called completely monotone if

$$(-1)^k \frac{d^k}{dt^k} g(t) \geq 0, \quad t > 0$$
$$\lim_{t \downarrow 0} g(t) = g(0)\,.$$

In the representation theorem below we consider the family $\mathcal{G}$ of Gaussian kernels

$$\mathcal{G} := \big(k_t\big)_{t \in \mathbb{R}^+}\,, \tag{5}$$

where for $t \geq 0$ the Gaussian kernel k_t is defined by $k_t(x, x') \ := \ e^{-t^2 \|x - x'\|^2}, x, x' \in \mathbb{R}^d$.

Theorem 1.1 *Consider a real function $g : \mathbb{R}^+ \to \mathbb{R}$ and its corresponding radial function*

$$k_g(x, x') \ := \ g(\|x - x'\|), \quad x, x' \in \mathbb{R}^d.$$

Then the following assertions are equivalent:

i) k_g is a kernel for all dimensions $d \geq 1$.

ii) There exists a finite Borel measure μ on $\mathbb{R}^+$ such that $k_g = \mathbb{E}_{t\sim\mu} k_t$.

iii) $g(\sqrt{\cdot}\,)$ is completely monotone.

2 RKHS of mixtures

Before we proceed to analyze the RKHS corresponding to a mixture of kernels in terms of its mixture components, let us recall some basic facts about RKHSs. Suppose a kernel k has a feature map $\Phi : X \to H$ to a Hilbert space H. Let $\mathcal{F}(X)$ denote the set of real-valued functions on X and consider the mapping $\Phi^* : H \to \mathcal{F}(X)$ defined by

$$\big(\Phi^* g\big)(x) \;:=\; \langle g, \Phi(x)\rangle_H, \quad x \in X, g \in H.$$

Where no confusion should arise, we write H_k for the RKHS $H_k(X)$ associated with the kernel k. Then by [29, Thm. 4.21] we have that the RKHS H_k corresponding to k can be described as

$$H_k = \{\Phi^* g : g \in H\}, \tag{6}$$

$$\|f\|^2_{H_k} = \inf_{g \in H\,:\, f = \Phi^* g} \|g\|^2_H, \tag{7}$$

and that $\Phi^* : H \to H_k$ is a metric surjection, that is $\Phi^* \mathring{B}(H) = \mathring{B}(H_k)$, where $\mathring{B}(\cdot)$ denotes the open unit ball of its argument. Consequently, if Φ^* is injective it follows that it is an isometric isomorphism. Furthermore, let $P_\Phi : H \to ker(\Phi^*)^\perp \subset H$ be the orthogonal projection onto the orthogonal complement of the null space $ker(\Phi^*)$ of Φ^*. Let us observe that the proof of [29, Thm. 4.21] proves that the infimum in (7) is actually attained, that is

$$\|f\|^2_{H_k} = \min_{\substack{g \in H \\ f = \Phi^* g}} \|g\|^2_H, \tag{8}$$

and that the minimum is attained at

$$\hat{g} \;:=\; P_\Phi g \tag{9}$$

for any g that satisfies $\Phi^* g = f$.

To analyze the RKHS corresponding to a mixture of kernels in terms of its mixture components, we essentially follow the proof of [3, Sec. 6] for sums. However, due to the infinite nature of the mixtures we need to utilize some measurability and integrability considerations in the context of Lebesgue-Bochner spaces. To that end, consider a measurable space (T, Σ) equipped with a measure μ. In this paper, we will only consider nontrivial measures. For a Banach space E, a function $\mathfrak{f} : T \to E$ is said to be E-measurable if it is the pointwise limit of a sequence of step functions. Let $\mathcal{H}$ denote a Hilbert space and consider equivalence classes of $\mathcal{H}$-measurable functions, where functions are equivalent if they differ only on sets of μ–measure zero. The Lebesgue-Bochner space $L_2(\mu, \mathcal{H})$ consists of those equivalence classes such that the square of the norm

$$\|\mathfrak{f}\|^2_{L_2(\mu,\mathcal{H})} \;:=\; \mathbb{E}_{t\sim\mu}\|\mathfrak{f}(t)\|^2_{\mathcal{H}}$$

is finite. $L_2(\mu, \mathcal{H})$ is known to be complete [10], the proof being essentially the same as for real Lebesgue space $L_2(\mu)$. Therefore it is a Hilbert space. We will also need the following notions. For a Banach space E a function $\mathfrak{f} : T \to E$ is said to be weakly E-measurable if $t \mapsto \langle b^*, \mathfrak{f}(t)\rangle$ is measurable for all $b^* \in E^*$. Clearly an E-measurable function is weakly E-measurable. On the other hand, by Petti's Theorem (see e.g. [13, Prop. 1.20, Pg. 9]), if $\mathfrak{f} : T \to E$ is weakly E-measurable and has separable range then it is E-measurable. In addition, let E_1 and E_2 be Banach spaces and denote by $L(E_1, E_2)$ the Banach space of bounded linear operators from E_1 to E_2. Then we say that a function $f : T \to L(E_1, E_2)$ is simply E_2-measurable if $t \mapsto f(t)b$ is E_2-measurable for all $b \in E_1$.

In this paper we will be concerned with a family $(k_t)_{t\in T}$ of kernels and their mixtures $k_\mu := \mathbb{E}_{t\sim\mu}k_t$ corresponding to finite measures μ. We use the notation $H_t := H_{k_t}$ and $H_\mu := H_{k_\mu}$. Since we never consider the trivial measure $\mu = 0$ no confusion should arise as to the meaning of $H_0 := H_{k_0}$. Roughly stated, our main result of this section, Theorem 2.2, states that if, corresponding to the family of kernels, we have a family $(\Phi_t)_{t\in T}$ of feature maps $\Phi_t : X \to \mathcal{H}$ to a *common* feature space $\mathcal{H}$ such that the induced map $\Phi : T \times X \to \mathcal{H}$ defined by $\Phi(t, x) := \Phi_t(x)$ has some regularity, then the square norm of H_μ is related to a μ mixture of the square of the H_t norms of the function's mixture components. Namely we have an integral version of (3) mentioned in the introduction. However, before we prove this theorem we establish a preparatory lemma of independent interest. Recall that a Suslin space is a continuous image of a Polish space.

Lemma 2.1 *Let T be a measurable space, X be a Suslin space equipped with its Borel σ-algebra, and $\mathcal{H}$ a separable Hilbert space. Consider a family $(\Phi_t)_{t\in T}$ of maps $\Phi_t : X \to \mathcal{H}$ and the corresponding family $(P_{\Phi_t})_{t\in T}$ of orthogonal projections $P_{\Phi_t} : \mathcal{H} \to \mathcal{H}$ onto the orthogonal complement of the null space $ker(\Phi_t^*)$. Suppose that the map $\Phi : T \times X \to \mathcal{H}$ defined by $\Phi(t, x) := \Phi_t(x)$ is weakly $\mathcal{H}$-measurable. Then the map $t \mapsto P_{\Phi_t}$ is simply $\mathcal{H}$-measurable.*

We can now state our main theorem that describes the RKHS of mixtures.

Theorem 2.2 *Let (T, Σ, μ) be a measure space and consider a family $(k_t)_{t\in T}$ of reproducing kernels on X equipped with a family $(\Phi_t)_{t\in T}$ of feature maps $\Phi_t : X \to \mathcal{H}$ to a common feature space $\mathcal{H}$. For each $x \in X$ consider the map $\Psi_x : T \to \mathcal{H}$ defined by $\Psi_x(t) := \Phi_t(x)$. Suppose that for each $x \in X$ we have $\Psi_x \in L_2(\mu, \mathcal{H})$. Then the function $t \mapsto k_t(x, x')$ is integrable for all $x, x' \in X$ and the map $\Psi : X \mapsto L_2(\mu, \mathcal{H})$ defined by $x \mapsto \Psi_x$ is a feature map for $k_\mu := \mathbb{E}_{t\sim\mu}k_t$. In addition, we have*

$$H_\mu = \{\Psi^*\mathfrak{f} : \mathfrak{f} \in L_2(\mu, \mathcal{H})\}. \tag{10}$$

and

$$\|f\|_{H_\mu}^2 = \min_{\substack{\mathfrak{f}\in L_2(\mu,\mathcal{H}) \\ f=\Psi^*\mathfrak{f}}} \|\mathfrak{f}\|_{L_2(\mu,\mathcal{H})}^2 , \tag{11}$$

where

$$\big(\Psi^*\mathfrak{f}\big)(x) = \mathbb{E}_{t\sim\mu}\Big(\big(\Phi_t^*\mathfrak{f}(t)\big)(x)\Big) = \mathbb{E}_{t\sim\mu}\big\langle \mathfrak{f}(t), \Phi_t(x)\big\rangle_{\mathcal{H}}. \tag{12}$$

Moreover, let X be a Suslin space equipped with its Borel σ-algebra, $\mathcal{H}$ be a separable Hilbert space, and suppose the map $\Phi : T \times X \to \mathcal{H}$ defined by $\Phi(t, x) := \Phi_t(x)$ is weakly $\mathcal{H}$-measurable. Then, in addition to (11), we have

$$\|f\|_{H_\mu}^2 = \min_{\substack{\mathfrak{f}\in L_2(\mu,\mathcal{H}) \\ f=\Psi^*\mathfrak{f}}} \mathbb{E}_{t\sim\mu}\|\langle \mathfrak{f}(t), \Phi_t(\cdot)\rangle_{\mathcal{H}}\|_{H_t}^2.$$

Note that, by definition, the last assertion of Theorem 2.2 can be stated as

$$\|f\|_{H_\mu}^2 = \min_{\substack{\mathfrak{f}\in L_2(\mu,\mathcal{H}) \\ f=\Psi^*\mathfrak{f}}} \mathbb{E}_{t\sim\mu}\|\Phi_t^*\mathfrak{f}(t)\|_{H_t}^2 .$$

3 The approximation error function inequality

In this section we will establish the integral approximation error function inequality (2). Although the following analysis is easier when the risk function is the expectation of a loss function, some

important risk functions are not of this type. For example, the two-class Neyman-Pearson classification problem (see e.g. [11, 26, 27]) is to minimize one type of error while constraining the other type of error. To handle this more general case, consider a risk function defined on a space which contains all $H_t, t \in T$. Our first order of business is then to consider when H_μ also lies in this space. For simplicity, we consider the case when $H_t \subset L_2(\nu), t \in T$, where ν is a measure on X. To be precise about the meaning of this, for $f : X \to \mathbb{R}$, let $[f]_\sim$ denote the equivalence class of functions which equal f ν-a.e. For $t \in T$ we say that $H_t \subset L_2(\nu)$ if for all $f \in H_t$ we have $[f]_\sim \in L_2(\nu)$.

Lemma 3.1 *In addition to the assumptions of Theorem 2.2, let X be a measurable space. Let ν be a measure on X and let $\hat{k} : T \times X \to \mathbb{R}$ defined by $\hat{k}(t,x) := k_t(x,x)$ satisfy $\hat{k}(t,\cdot) \in L_1(\nu), t \in T$, and $\hat{k} \in L_1(\mu \otimes \nu)$. Then we have inclusions $I_t : H_t \hookrightarrow L_2(\nu)$ satisfying*

$$\|I_t\| \leq \|\hat{k}(t,\cdot)\|_{L_1(\nu)}^{\frac{1}{2}}, \quad t \in T$$

and $I_\mu : H_\mu \hookrightarrow L_2(\nu)$ satisfying

$$\|I_\mu\| \leq \|\hat{k}\|_{L_1(\mu\otimes\nu)}^{\frac{1}{2}}.$$

Note that the inclusions above may not be injective. Indeed, [29, Thm. 4.26] shows that, for $t \in T$, the inclusion I_t is injective if and only if the image of the integral operator associated with the kernel is dense in the RKHS. For more about this topic see the discussion after [29, Thm. 4.26].

Therefore, if Lemma 3.1 applies and we have a risk function $\mathcal{R} : L_2(\nu) \to \mathbb{R}$ we can define risk functions on $H_t, t \in T$ and H_μ through the injections. Moreover, for all $\lambda \geq 0$ define the regularized risk functions $\mathcal{R}_{\lambda,t} : H_t \to \mathbb{R}\,, t \in T$ and $\mathcal{R}_{\lambda,\mu} : H_\mu \to \mathbb{R}$ by

$$\mathcal{R}_{\lambda,t}(f) \;:=\; \lambda\|f\|_{H_t}^2 + \mathcal{R}(I_t f), \quad f \in H_t,$$

$$\mathcal{R}_{\lambda,\mu}(f) \;:=\; \lambda\|f\|_{H_\mu}^2 + \mathcal{R}(I_\mu f), \quad f \in H_\mu.$$

Finally, consider their minimum values $\mathcal{R}^*_{\lambda,t} := \inf_{f\in H_t} \mathcal{R}_{\lambda,t}(f)$ and $\mathcal{R}^*_{\lambda,\mu} := \inf_{f\in H_\mu} \mathcal{R}_{\lambda,\mu}(f)$.

Before we state our main result concerning a relationship between the minimum regularized risk associated with H_μ and that of $H_t, t \in T$, we establish a result that will be useful in its proof. Let $f \in H_\mu$ and consider the case when Lemma 3.1 applies. Since $I_\mu : H_\mu \hookrightarrow L_2(\nu)$ we have $I_\mu f \in L_2(\nu)$. On the other hand, by Theorem 2.2 we have $f(x) = \mathbb{E}_{t\sim\mu}\Big(\big(\Phi_t^* \mathfrak{f}(t)\big)(x)\Big)$ for some $\mathfrak{f} \in L_2(\mu, \mathcal{H})$. Since $\Phi_t^*\mathfrak{f}(t) \in H_t$ and $I_t : H_t \hookrightarrow L_2(\nu)$ for all $t \in T$, we can consider whether the $L_2(\nu)$-valued Bochner integral $\mathbb{E}_{t\sim\mu} I_t \Phi_t^* \mathfrak{f}(t)$ exists and if it exists, whether $I_\mu f = \mathbb{E}_{t\sim\mu} I_t \Phi_t^* \mathfrak{f}(t)$. The following theorem gives sufficient conditions for this to be the case.

Theorem 3.2 *In addition to the assumptions of Theorem 2.2 and Lemma 3.1, let ν be a σ-finite measure on X and μ be σ-finite measure on T. Suppose that the map $\Phi : T \times X \to \mathcal{H}$ defined by $\Phi(t,x) := \Phi_t(x)$ is weakly $\mathcal{H}$-measurable and that the map $T \to L\big(\mathcal{H}, L_2(\nu)\big)$ defined by $t \mapsto I_t\Phi_t^*$ is simply $L_2(\nu)$-measurable. Then the function $t \mapsto I_t\Phi_t^*\mathfrak{f}(t)$ is Bochner integrable for all $\mathfrak{f} \in L_2(\mu,\mathcal{H})$ and so defines an integral operator $\mathcal{I} : L_2(\mu,\mathcal{H}) \to L_2(\nu)$ by*

$$\mathcal{I}\mathfrak{f} \;:=\; \mathbb{E}_{t\sim\mu} I_t \Phi_t^* \mathfrak{f}(t).$$

This integral operator satisfies

$$\mathcal{I} = I_\mu \Psi^*$$

with $\|\mathcal{I}\| \leq \|\hat{k}\|_{L_1(\mu\otimes\nu)}^{\frac{1}{2}}$.

Remark 3.3 Theorem 3.2 is also true when $L_2(\nu)$ is replaced by the space $C_b(X)$ of bounded continuous functions or $L_p(\nu), p \geq 1$ since the extension is trivial for $C_b(X)$ and the Dunford-Schwartz Theorem [14, Thm. 17, Pg. 198] concerning scalar representations for Bochner integrals is all that is needed for the extension to $L_p(\nu), p \geq 1$. Moreover, we suspect that the Dunford-Schwartz Theorem also applies to Köthe spaces [19, Pg. II.28].

We can now establish our main result concerning a relationship between the minimum regularized risk associated with H_μ and that of $H_t, t \in T$. To simplify we state the result only for probability measures.

Theorem 3.4 *In addition to the assumptions of Theorems 2.2 and 3.2, let μ be a probability measure and suppose that $\mathcal{R} : L_2(\nu) \to \mathbb{R}$ is a continuous convex function. Then we have*

$$\mathcal{R}^*_{\lambda,\mu} \leq \mathbb{E}_{t\sim\mu}\mathcal{R}^*_{\lambda,t}\,, \quad \lambda \geq 0. \tag{13}$$

Now let us apply Theorem 3.4 to the approximation error functions. We define the approximation error functions to be $A_t(\lambda) := \mathcal{R}^*_{\lambda,t} - \mathcal{R}^*_{0,t}, t \geq 0$ and $A_\mu(\lambda) := \mathcal{R}^*_{\lambda,\mu} - \mathcal{R}^*_{0,\mu}$ and the Bayes risk to be $\mathcal{R}^* := \inf_{f\in L_2(\nu)} \mathcal{R}(f)$. We have the following corollary.

Corollary 3.5 *In addition to the assumptions of Theorem 3.4, suppose that*

$$\mu\big(\{t \in T : \mathcal{R}^*_{0,t} \neq \mathcal{R}^*\}\big) = 0.$$

Then we have $\mathcal{R}^*_{0,\mu} = \mathcal{R}^*$ and

$$A_\mu(\lambda) \leq \mathbb{E}_{t\sim\mu}A_t(\lambda)\,, \quad \lambda \geq 0.$$

4 Radial kernels

We now show that all the previous results apply to the radial kernels $\mathcal{K}_{rad}$ and then apply Corollary 3.5 to bound the approximation error function corresponding to the hinge-loss risk for the two kernels mentioned in the introduction. To that end, we introduce some notations and representations. Suppose that $Y \subset \mathbb{R}$ is measurable and P a probability measure on $X \times Y$. Then, according to [29, Def. 2.16], a function $L : X \times Y \times \mathbb{R} \to \mathbb{R}^+$ is said to be a convex continuous P-integrable Nemitski loss of order $p \in [1, 2]$ if it is convex and continuous in its last variable for all $x \in X, y \in Y$, and there exits a P-integrable function b and a constant $c > 0$ such that for all $x \in X, y \in Y, t \in \mathbb{R}$ we have $L(x, y, t) \leq b(x, y) + c|t|^p$. Also, let $T = \mathbb{R}^+$ and define $\mathcal{H} := L_2(\mathbb{R}^d)$. Consider the family $(\Phi_t)_{t\in\mathbb{R}^+}$ of maps $\Phi_t : \mathbb{R}^d \to \mathcal{H}$ defined as follows. For $t = 0$ select $z \in \mathcal{H}$ such that $\|z\|_{\mathcal{H}} = 1$ and define

$$\Phi_0(x) \;:=\; z, \quad x \in \mathbb{R}^d\,.$$

For $t > 0$, define

$$\Phi_t(x) \;:=\; \frac{t^{\frac{d}{2}}2^{\frac{d}{2}}}{\pi^{\frac{d}{4}}}e^{-2t^2\|x-\cdot\|_2^2}, \quad x \in \mathbb{R}^d\,.$$

Then [29, Lem. 4.45] implies that the family $(\Phi_t)_{t\in\mathbb{R}^+}$ of maps $\Phi_t : X \to \mathcal{H}$ obtained by restricting to an arbitrary subset $X \subset \mathbb{R}^d$ are feature maps for the Gaussian kernels $k_t \in \mathcal{G}, t \in \mathbb{R}^+$, defined on X.

Theorem 4.1 *Consider the family $(\Phi_t)_{t\in\mathbb{R}^+}$ defined above.*

i) Let $X \subset \mathbb{R}^d$ be a Borel subset. Then Lemma 2.1 applies.

ii) Consider a radial kernel $k \in \mathcal{K}_{rad}$ and a finite Borel representing measure μ such that $k = k_\mu$. Moreover, consider the family $(k_t)_{t\in\mathbb{R}^+}$ of Gaussian kernels equipped with the above defined family $(\Phi_t)_{t\in\mathbb{R}^+}$ of feature maps. Then Theorem 2.2 applies.

iii) Suppose further that ν is a finite Borel measure on X . Then Lemma 3.1 and Theorem 3.2 apply and assert that

$$\begin{aligned} \|I_t\| &\leq \sqrt{\nu(X)}, \quad t \in T, \\ \|I_\mu\| &\leq \sqrt{\mu(\mathbb{R}^+)\nu(X)} < \infty, \\ \|\mathcal{I}\| &\leq \sqrt{\mu(\mathbb{R}^+)\nu(X)} < \infty. \end{aligned}$$

iv) Suppose further that $k(x,x) = 1, x \in X$, and $\mathcal{R} : L_2(\nu) \mapsto \mathbb{R}$ is a continuous convex function. Then Theorem 3.4 applies.

v) In addition to the assumptions of i),ii), and iii), suppose that $Y \subset \mathbb{R}$ is measurable and P a probability measure on $X \times Y$. Moreover, let $L : X \times Y \times \mathbb{R} \to \mathbb{R}^+$ be convex continuous P-integrable Nemitski loss of order $p \in [1,2]$, and consider the corresponding risk function $\mathcal{R}(f) := \mathbb{E}_{(x,y)\sim P} L(x,y,f(x))$. Finally, suppose that $\lim_{x\mapsto\infty} k(x,x') = 0$, $x' \in X$. Then, $\mathcal{R} : L_2(P_X) \to \mathbb{R}^+$ and Corollary 3.5 applies.

We now use Corollary 3.5 via Theorem 4.1 to bound the approximation error function for the hinge risk and the RKHSs corresponding to the two kernels mentioned in the introduction. Let P be a probability measure on $X \times \{-1,1\}$ and consider the hinge loss $L : \mathbb{R} \times \{-1,1\} \to \mathbb{R}^+$ defined by $L(s,y) := \max\{0, 1-ys\}$ and the hinge risk $\mathcal{R}(f) := \mathbb{E}_{(x,y)\sim P} L(f(x),y)$. Then, as defined in [29, Def. 8.15], we say that P has margin-noise exponent $\beta \in [0,\infty)$ if

$$\int_{\Delta(x)<t} |2\eta(x) - 1| dP_X(x) \leq ct^\beta, \quad t \geq 0$$

for some version $\eta : X \to [0,1]$ of the conditional probability $\eta(x) := P(y=1|x)$, where $\Delta(x)$ is the distance to the decision boundary defined by $\{x : \eta(x) = \frac{1}{2}\}$. The noise exponent quantifies the concentration of mass around the decision boundary and is used to bound the approximation error function for the hinge risk and Gaussian kernels in [29, Thm. 8.18]. The following result can easily be extended to arbitrary measurable subsets $X \subset \mathbb{R}^d$ with the assumption of a tail exponent for P_X, thus extending [29, Thm. 8.18] to the two kernels mentioned in the introduction.

Corollary 4.2 *Let $X \subset \mathbb{R}^d$ be the closed unit ball and let P be a probability measure on $X \times \{-1,1\}$ with margin-noise exponent $\beta \in (0,\infty)$. Let $\mathcal{R}$ denote the hinge risk function. Moreover, for $\alpha > 0$, consider the exponential kernel*

$$k(x,x') := e^{-\alpha\|x-x'\|}, \quad x,x' \in X$$

and its RKHS H_k. Then, for $d \geq 2$, we have

$$A_k(\lambda) \leq C_{d,\beta}\big(\lambda^{\frac{1}{d}}\alpha + \alpha^{-\beta}\big), \quad \lambda > 0$$

where $C_{d,\beta}$ is a constant depending only on d and β.

Corollary 4.3 *With the assumptions of Corollary 4.2, instead consider the kernel*

$$k(x,x') := \big(1 + m^{-1}\|x-x'\|_2^2\big)^{-\alpha}, \quad x,x' \in X$$

and its RKHS H_k, for $m > 0, \alpha > 0$ and $d \geq 1$. Then for $2\alpha - \beta \geq 1$ we have

$$A_k(\lambda) \leq C_{d,\beta}\Big(\lambda m^{-\frac{d}{2}} \frac{\Gamma(\frac{d}{2}+\alpha)}{\Gamma(\alpha)} + m^{\frac{\beta}{2}} \frac{\Gamma(\alpha - \frac{\beta}{2})}{\Gamma(\alpha)}\Big), \quad \lambda > 0$$

and for $2\alpha - \beta < 1$ we have

$$A_k(\lambda) \leq C_{d,\beta}\Big(\lambda m^{-\frac{d}{2}} \frac{\Gamma(\frac{d}{2}+\alpha)}{\Gamma(\alpha)} + m^{\frac{\alpha\beta}{\beta+2}} \frac{1+\alpha}{\alpha\Gamma(\alpha)}\Big), \quad \lambda > 0$$

where $\Gamma(z) := \int_{\mathbb{R}^+} t^{z-1}e^{-t}, z > 0$ is the Gamma function and $C_{d,\beta}$ is a constant depending only on d and β.

Remark 4.4 The inequalities of Corollary 4.3 can be simplified using the inequality $\frac{\Gamma(\frac{d}{2}+\alpha)}{\Gamma(\alpha)} \leq \alpha^{\frac{d}{2}}$, found in [23] and the references therein. Moreover, the inequalities of Corollaries 4.2 and 4.3 can easily be sharpened using simple modifications of the proofs. However, our preliminary analysis lead to more complex results than those presented. The development of sharper, yet simple, bounds on the approximation error function is out of the scope of this paper.

4.1 Additional results for radial kernels

We now utilize the fact (see e.g. [29, Prop. 4.46]) that the family $\mathcal{G}$ of Gaussian kernels is nested in the sense that $H_{t_1} \subset H_{t_2}$, $0 < t_1 \leq t_2$. We first prove that $H_\mu(\mathbb{R}^d)$ does not contain constants if $\mu(\{0\}) = 0$. Let $\mathbf{1}$ denote the constant function with value 1.

Theorem 4.5 *If $\mu(\{0\}) = 0$, then we have $\mathbf{1} \notin H_\mu(\mathbb{R}^d)$.*

Note that if we choose $\mu := \delta_t$, the Dirac measure situated at $t > 0$, we obtain $\mathbf{1} \notin H_t(\mathbb{R}^d)$ which is a special case of the "no constants" theorem for Gaussian RKHSs [29, Cor. 4.44].

Each fixed $\alpha \geq 0$ determines an operator $\alpha^* : \mathcal{M} \to \mathcal{M}$ on measures defined by $\big(\alpha^*\mu\big)(A) := \mu(\alpha A)$. Therefore any μ determines a one parameter family of radial kernels $(k_{\alpha^*\mu})_{\alpha\geq 0}$. From [29, Prop. 4.46] we know that, for all $0 < \alpha_1 \leq \alpha_2$, $t > 0$, we have $H_{\alpha_1 t} \subset H_{\alpha_2 t}$ and that

$$\|\mathrm{id} : H_{\alpha_1 t} \to H_{\alpha_2 t}\| \leq \Big(\frac{\alpha_2}{\alpha_1}\Big)^{\frac{d}{2}}. \tag{14}$$

The following result shows we have the same results for $(H_{\alpha^*\mu})_{\alpha>0}$.

Lemma 4.6 *Consider a finite Borel measure μ and the family of kernels $(k_{\alpha^*\mu})_{\alpha\geq 0}$ on X. Then for all $0 < \alpha_1 \leq \alpha_2$ we have $H_{\alpha_1^*\mu} \subset H_{\alpha_2^*\mu}$ and*

$$\|\mathrm{id} : H_{\alpha_1^*\mu} \to H_{\alpha_2^*\mu}\| \leq \Big(\frac{\alpha_2}{\alpha_1}\Big)^{\frac{d}{2}}.$$

The following two theorems demonstrate sufficient conditions to have $H_\mu \subset H_t$ or $H_t \subset H_\mu$ for some t.

Theorem 4.7 *Let $X \subset \mathbb{R}^d$ and consider a finite Borel measure μ such that $\mu(\{0\}) = 0$ and $\mathbb{E}_{t\sim\mu} t^{-d} < \infty$. Furthermore, assume that, for some $t^* > 0$, we have $\mu\big([t^*, \infty)\big) = 0$. Then we have $H_\mu \subset H_{t^*}$ and*

$$\|\mathrm{id} : H_\mu \to H_{t^*}\| \leq (t^*)^{\frac{d}{2}} \sqrt{\mathbb{E}_{t\sim\mu} t^{-d}}.$$

Theorem 4.8 *Let $X \subset \mathbb{R}^d$ and consider a finite Borel measure μ satisfying $\mu(0,\infty) > 0$. Then there exists a $t_1 > 0$ such that $\mu([t_1,\infty)) > 0$ and for any such t_1 we have $H_{t_1} \subset H_\mu$. Moreover, for any such t_1 there exists a t_2 such that $\mu([t_1,t_2]) > 0$ and for any such t_2 we have*

$$\|\mathrm{id} : H_{t_1} \to H_\mu\| \leq \frac{t_1^{-\frac{d}{2}}}{\mu([t_1,t_2])} \Big(\int_{[t_1,t_2]} t^d d\mu(t) \Big)^{\frac{1}{2}} \leq \Big(\frac{t_2}{t_1}\Big)^{\frac{d}{2}} \mu([t_1,t_2])^{-\frac{1}{2}}.$$

The following corollary in particular generalizes the universality result of [21] to noncompact X.

Corollary 4.9 *Let $X \subset \mathbb{R}^d$ and consider a non-constant radial kernel k. Then the following hold:*

i) $H_k(\mathbb{R}^d)$ is dense in $L_p(\nu)$ for all $p \in [1,\infty)$ and all finite measures ν on $\mathbb{R}^d$.

ii) If $X \subset \mathbb{R}^d$ is compact, then k is universal.

iii) If $\mu([t,\infty)) > 0$ for all $t > 0$, we have $\cup_{t>0} H_t \subset H_\mu$.

iv) k is strictly positive definite.

5 Proofs

Proof of Theorem 1.1: Recall that a symmetric bivariate function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called positive definite ($k \gg 0$) if for all n, and $x_i \in \mathbb{R}^d$, $a_i \in \mathbb{R}, i = 1,..,n$ we have

$$\sum_{i,j=1}^n a_i a_j k(x_i,x_j) \geq 0\,. \tag{15}$$

Observe that Moore's result [22] (see e.g. [3]) asserts that k_g is a reproducing kernel, if and only if it is positive definite. Moreover, Schoenberg's result [25, Thm. 2], which heavily uses the representation of translation invariant functions of Bochner [9], states that k_g is positive definite for all d if and only if there exists a finite Borel measure μ on $\mathbb{R}^+$ such that $g(s) = \mathbb{E}_{t\sim\mu} e^{-t^2 s^2}$. Substituting $s := \|x - x'\|$ yields the equivalence between i) and ii). The equivalence between ii) and iii) is the result of Berstein [8, 35] (see also Schoenberg [25, Thm. 3]). For a thorough discussion of this topic see [7]. ■

Proof of Lemma 2.1: We need to show that for $f \in \mathcal{H}$ the function $t \mapsto P_{\Phi_t} f$ is $\mathcal{H}$-measurable. To that end, fix an $f \in \mathcal{H}$, and consider the function $h : T \times X \times \mathcal{H} \to \mathbb{R}$ defined by $h(t,x,g) := \big(\langle g - f, \Phi_t(x)\rangle\big)^2$. Since the map $(t,x) \mapsto \Phi_t(x)$ is weakly $\mathcal{H}$-measurable it follows, for fixed g, that h is measurable in (t,x). Moreover, h is obviously continuous in g for (x,t) fixed. Since $\mathcal{H}$ is separable and complete it is Polish. Therefore, it follows from Carathéodory's Lemma [12, Lem. III.39] (see also [29, Lem. A.3.17]) that h is measurable. Since X is Suslin it follows from [12, Lem. III.39] that $\acute{h} : T \times \mathcal{H} \to \mathbb{R}$ defined by $\acute{h}(t,g) := \sup_{x\in X} h(t,x,g)$ is measurable. Now observe that since $ker(\Phi_t^*) = \{w \in \mathcal{H} : \Phi_t^* w = 0\} = \{w \in \mathcal{H} : \langle w, \Phi_t(x)\rangle = 0, x \in X\}$ the set-valued function $F : \mathbb{R}^+ \to 2^{\mathcal{H}}$ defined by

$$F(t) = f + ker(\Phi_t^*).$$

satisfies $F(t) = \{g : \acute{h}(t,g) = 0\}$. Since the function $\omega : T \times \mathcal{H} \to \mathbb{R}$ defined by

$$\omega(t,g) = \|g\|_{\mathcal{H}}^2$$

is measurable and equations (8) and (9) assert that the infimum $\inf_{g\in F(t)}\omega(t,g)=\inf_{g:\Phi_t^*g=\Phi_t^*f}\|g\|_{\mathcal{H}}^2$ is attained at $g(t) := P_{\Phi_t}f$, Aumann's selection principle [29, Lem. A.3.18] implies that : $t\mapsto P_{\Phi_t}f$ is measurable. Since $\mathcal{H}$ is separable the assertion follows from [13, Cor. 1.9, Pg. 6].

■

Proof of Theorem 2.2: First observe that the assumption $\Psi_x,\Psi_{x'}\in L_2(\mu,\mathcal{H})$ implies that the function

$$t\mapsto 4\langle\Psi_x(t),\Psi_{x'}(t)\rangle_{\mathcal{H}}=\|\Psi_x(t)+\Psi_{x'}(t)\|^2-\|\Psi_x(t)-\Psi_{x'}(t)\|^2$$

is integrable for all $x,x'\in X$. Consequently, we obtain

$$\langle\Psi_x,\Psi_{x'}\rangle_{L_2(\mu,\mathcal{H})}=\mathbb{E}_{t\sim\mu}\langle\Psi_x(t),\Psi_{x'}(t)\rangle_{\mathcal{H}}=\mathbb{E}_{t\sim\mu}\langle\Phi_t(x),\Phi_t(x')\rangle_{\mathcal{H}}=\mathbb{E}_{t\sim\mu}k_t(x,x')$$

and so conclude that the function $t\mapsto k_t(x,x')$ is μ integrable for all $x,x'\in X$, and Ψ is a feature map for $k_\mu:=\mathbb{E}_{t\sim\mu}k_t$. Therefore, we obtain (10) and (11) from (6) and (7) respectively. Since

$$\big(\Psi^*\mathfrak{f}\big)(x)=\langle\mathfrak{f},\Psi_x\rangle_{L_2(\mu,\mathcal{H})}=\mathbb{E}_{t\sim\mu}\langle\mathfrak{f}(t),\Psi_x(t)\rangle_{\mathcal{H}}=\mathbb{E}_{t\sim\mu}\langle\mathfrak{f}(t),\Phi_t(x)\rangle_{\mathcal{H}}=\mathbb{E}_{t\sim\mu}\Big(\big(\Phi_t^*\mathfrak{f}(t)\big)(x)\Big),$$

we then obtain (12).

For the last assertion, let us first show that for $\mathfrak{f}\in L_2(\mu,\mathcal{H})$ the function $\hat{\mathfrak{f}}:T\to\mathcal{H}$ defined by $\hat{\mathfrak{f}}(t):=P_{\Phi_t}\mathfrak{f}(t)$ satisfies $\|\hat{\mathfrak{f}}(t)\|_{\mathcal{H}}^2=\|\Phi_t^*\mathfrak{f}(t)\|_{H_t}^2$, $\hat{\mathfrak{f}}(t)\in L_2(\mu,\mathcal{H})$, and $\Psi^*\mathfrak{f}=\Psi^*\hat{\mathfrak{f}}$. That is, for $\mathfrak{f}\in L_2(\mu,\mathcal{H})$, defining $\hat{\mathfrak{f}}(t):=P_{\Phi_t}\mathfrak{f}(t), t\in T$, we have

$$\|\hat{\mathfrak{f}}(t)\|_{\mathcal{H}}^2=\|\Phi_t^*\mathfrak{f}(t)\|_{H_t}^2,\ t\in T \tag{16}$$

and, for $f\in H_\mu$, we have

$$\Big\{\hat{\mathfrak{f}}:T\to\mathbb{R}\Big|\exists\mathfrak{f}\in L_2(\mu,\mathcal{H}):f=\Psi^*\mathfrak{f}\text{ and }\hat{\mathfrak{f}}(t)=P_{\Phi_t}\mathfrak{f}(t),t\in T\Big\}\subset\Big\{\mathfrak{f}\in L_2(\mu,\mathcal{H}):f=\Psi^*\mathfrak{f}\Big\}. \tag{17}$$

To that end, first observe that (8) and (9) imply that $\|\hat{\mathfrak{f}}(t)\|_{\mathcal{H}}^2=\|\Phi_t^*\mathfrak{f}(t)\|_{H_t}^2$. Moreover, Lemma 2.1 and [13, Prop. 1.13, Pg. 7] imply that $\hat{\mathfrak{f}}$ is $\mathcal{H}$-measurable for $\mathfrak{f}\in L_2(\mu,\mathcal{H})$. Since $\|\hat{\mathfrak{f}}(t)\|_{\mathcal{H}}^2\le\|\mathfrak{f}(t)\|_{\mathcal{H}}^2$, we conclude that $\hat{\mathfrak{f}}(t)\in L_2(\mu,\mathcal{H})$. Now fix $t\in T$. Since P_{Φ_t} is an orthogonal projection it follows that $\mathfrak{f}(t)-\hat{\mathfrak{f}}(t)=\mathfrak{f}(t)-P_{\Phi_t}\mathfrak{f}(t)\in ker(\Phi_t^*)$. Consequently, we obtain $\Phi_t^*\hat{\mathfrak{f}}(t)=\Phi_t^*P_{\Phi_t}\mathfrak{f}(t)=\Phi_t^*\mathfrak{f}(t)$ and therefore

$$\big(\Psi^*\mathfrak{f}\big)(x)=\mathbb{E}_{t\sim\mu}\big(\Phi_t^*\mathfrak{f}(t)\big)(x)=\mathbb{E}_{t\sim\mu}\big(\Phi_t^*\hat{\mathfrak{f}}(t)\big)(x)=\big(\Psi^*\hat{\mathfrak{f}}\big)(x).$$

That is, $\Psi^*\mathfrak{f}=\Psi^*\hat{\mathfrak{f}}$, establishing the claim.

To prove the last assertion, consider $f\in H_\mu$. It follows from the first assertion, (16), and (17), that

$$\begin{aligned}\|f\|_{H_\mu}^2=\inf_{\substack{\mathfrak{f}\in L_2(\mu,\mathcal{H})\\ f=\Psi^*\mathfrak{f}}}\|\mathfrak{f}\|_{L_2(\mu,\mathcal{H})}^2 &= \inf_{\substack{\mathfrak{f}\in L_2(\mu,\mathcal{H})\\ f=\Psi^*\mathfrak{f}}}\mathbb{E}_{t\sim\mu}\|\mathfrak{f}(t)\|_{\mathcal{H}}^2\\ &\le \inf_{\substack{\hat{\mathfrak{f}}(t)=P_{\Phi_t}\mathfrak{f}(t),t\in T\\ \mathfrak{f}\in L_2(\mu,\mathcal{H})\\ f=\Psi^*\mathfrak{f}}}\mathbb{E}_{t\sim\mu}\|\hat{\mathfrak{f}}(t)\|_{\mathcal{H}}^2\\ &= \inf_{\substack{\mathfrak{f}\in L_2(\mu,\mathcal{H})\\ f=\Psi^*\mathfrak{f}}}\mathbb{E}_{t\sim\mu}\|\Phi_t^*\mathfrak{f}(t)\|_{H_t}^2.\end{aligned}$$

To obtain an equality observe that for fixed $\mathfrak{f}\in L_2(\mu,\mathcal{H})$ we have $\|\Phi_t^*\mathfrak{f}(t)\|_{H_t}\le\|\mathfrak{f}(t)\|_{\mathcal{H}}$ so we conclude that $\mathbb{E}_{t\sim\mu}\|\Phi_t^*\mathfrak{f}(t)\|_{H_t}^2\le\mathbb{E}_{t\sim\mu}\|\mathfrak{f}(t)\|_{\mathcal{H}}^2=\|\mathfrak{f}\|_{L_2(\mu,\mathcal{H})}^2$. Equality then follows from the first line of the above displayed inequality, establishing the last assertion with an *infimum*. To obtain the expression with a *minimum*, observe that (8) and (9) imply that the infimum is attained in the first line. Let $\mathfrak{f}$ be a minimizer. Then the above discussion shows that $\hat{\mathfrak{f}}$ is also a minimizer. ■

Proof of Lemma 3.1: The first assertion follows from the proof of [29, Thm. 4.26]. Since Theorem 2.2 implies that $k_\mu(x,x) = \mathbb{E}_{t\sim\mu} k_t(x,x)$, and $k_t(x,x) \geq 0, t \in T, x \in X$, it follows from Tonelli's theorem that $\mathbb{E}_{x\sim\nu} k_\mu(x,x) = \mathbb{E}_{x\sim\nu}\mathbb{E}_{t\sim\mu} k_t(x,x) = \mathbb{E}_{(t,x)\sim\mu\otimes\nu} k_t(x,x)$. Therefore we obtain $\|\hat{k}_\mu\|_{L_1(\nu)} = \|\hat{k}\|_{L_1(\mu\otimes\nu)}$ so that the second assertion also follows from the proof of [29, Thm. 4.26]. ■

Proof of Theorem 3.2: Recall that the assumption that $t \mapsto I_t\Phi_t^*$ is simply measurable means that the function $t \mapsto I_t\Phi_t^* g$ is $L_2(\nu)$-measurable for all $g \in \mathcal{H}$. Consequently, [13, Prop. 1.13, Pg. 7] implies that the function $t \mapsto I_t\Phi_t^*\mathfrak{f}(t)$ is $L_2(\nu)$-measurable for all $\mathfrak{f} \in L_2(\mu,\mathcal{H})$. Since for all $t \geq 0$ we have $\|I_t\Phi_t^*\mathfrak{f}(t)\|_{L_2(\nu)} \leq \|I_t\|\|\Phi_t^*\mathfrak{f}(t)\|_{H_t} \leq \|I_t\|\|\mathfrak{f}(t)\|_{\mathcal{H}} \leq \|\hat{k}(t,\cdot)\|_{L_1(\nu)}^{\frac{1}{2}}\|\mathfrak{f}(t)\|_{\mathcal{H}}$, we conclude that

$$\mathbb{E}_{t\sim\mu}\|I_t\Phi_t^*\mathfrak{f}(t)\|_{L_2(\nu)} \leq \mathbb{E}_{t\sim\mu}\big(\|\hat{k}(t,\cdot)\|_{L_1(\nu)}^{\frac{1}{2}}\|\mathfrak{f}(t)\|_{\mathcal{H}}\big) \leq \|\hat{k}\|_{L_1(\mu\otimes\nu)}^{\frac{1}{2}}\|\mathfrak{f}\|_{L_2(\mu,\mathcal{H})}$$

and conclude that $t \mapsto I_t\Phi_t^*\mathfrak{f}(t)$ is integrable and so the integral operator $\mathcal{I}$ is well defined. Moreover, since

$$\|\mathcal{I}\mathfrak{f}\| \leq \mathbb{E}_{t\sim\mu}\|I_t\Phi_t^*\mathfrak{f}(t)\| \leq \|\hat{k}\|_{L_1(\mu\otimes\nu)}^{\frac{1}{2}}\|\mathfrak{f}\|_{L_2(\mu,\mathcal{H})},$$

we conclude that $\mathcal{I} : L_2(\mu,\mathcal{H}) \to L_2(\nu)$ is continuous and $\|\mathcal{I}\| \leq \|\hat{k}\|_{L_1(\mu\otimes\nu)}^{\frac{1}{2}}$.

To prove that $\mathcal{I} = I_\mu\Psi^*$, first observe that by the assumption that $T \times X \mapsto \Phi_t(x)$ is weakly $\mathcal{H}$-measurable, it follows from [13, Prop. 1.13, Pg. 7] that the function $T \times X \mapsto \langle \mathfrak{f}(t), \Phi_t(x)\rangle = \big(\Phi_t^*\mathfrak{f}(t)\big)(x)$ is measurable. Now let $\mathfrak{f} \in L_2(\mu,\mathcal{H})$ and consider $\mathcal{I}\mathfrak{f} = \mathbb{E}_{t\sim\mu} I_t\Phi_t^*\mathfrak{f}(t)$. Then the Dunford-Schwartz Theorem [14, Thm. 17, Pg. 198] states that there exists a measurable function $g : T\times X \to \mathbb{R}$, *uniquely* determined except for a set of $\mu\otimes\nu$-measure zero, such that $[g(t,\cdot)]_\sim = I_t\Phi_t^*\mathfrak{f}(t)$ for μ-almost all $t \in T$. Moreover, $g(\cdot,x)$ is μ-integrable for ν-almost all $x \in X$ and $\big[\mathbb{E}_{t\sim\mu} g(t,\cdot)\big]_\sim = \mathcal{I}\mathfrak{f}$. Consequently, since the function $(t,x) \mapsto \big(\Phi_t^*\mathfrak{f}(t)\big)(x)$ is measurable we conclude, by the uniqueness, that $\big[\mathbb{E}_{t\sim\mu}\big(\big(\Phi_t^*\mathfrak{f}(t)\big)(x)\big)\big]_\sim = \mathbb{E}_{t\sim\mu} I_t\Phi_t^*\mathfrak{f}(t) = \mathcal{I}\mathfrak{f}$. Since $(\Psi^*\mathfrak{f})(x) = \mathbb{E}_{t\sim\mu}\Big(\big(\Phi_t^*\mathfrak{f}(t)\big)(x)\Big)$ we conclude that $\big[\Psi^*\mathfrak{f}\big]_\sim = \mathcal{I}\mathfrak{f}$. That is, $\mathcal{I}\mathfrak{f} = I_\mu\Psi^*\mathfrak{f}$. Since $\mathfrak{f} \in L_2(\mu,\mathcal{H})$ was arbitrary we conclude that $\mathcal{I} = I_\mu\Psi^*$. ■

Proof of Theorem 3.4: First consider $\lambda > 0$. Suppose that $f \in H_\mu$. Then by Theorem 3.2, for all $\mathfrak{f} \in L_2(\mu,\mathcal{H})$ with $f = \Psi^*\mathfrak{f}$, we have $I_\mu f = I_\mu\Psi^*\mathfrak{f} = \mathcal{I}\mathfrak{f}$. In addition, Jensen's inequality for Bochner integrals, Theorem 6.3 (see [34, Sec. 4] for a more general result), implies that the integral of $t \mapsto \mathcal{R}(I_t\Phi_t^*\mathfrak{f}(t))$ exists and

$$\mathcal{R}(\mathcal{I}\mathfrak{f}) = \mathcal{R}(\mathbb{E}_{t\sim\mu} I_t\Phi_t^*\mathfrak{f}(t)) \leq \mathbb{E}_{t\sim\mu}\mathcal{R}(I_t\Phi_t^*\mathfrak{f}(t)).$$

Consequently, we have

$$\mathcal{R}_{\lambda,\mu}(f) = \lambda\|f\|_{H_\mu}^2 + \mathcal{R}(I_\mu f) \leq \lambda\mathbb{E}_{t\sim\mu}\|\mathfrak{f}(t)\|_{\mathcal{H}}^2 + \mathcal{R}(\mathcal{I}\mathfrak{f}) \leq \mathbb{E}_{t\sim\mu}\Big(\lambda\|\mathfrak{f}(t)\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^*\mathfrak{f}(t))\Big),$$

and conclude that

$$\mathcal{R}_{\lambda,\mu}^* \leq \inf_{\mathfrak{f}\in L_2(\mu,\mathcal{H})} \mathbb{E}_{t\sim\mu}\Big(\lambda\|\mathfrak{f}(t)\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^*\mathfrak{f}(t))\Big). \tag{18}$$

Now consider the function $\phi : T \times \mathcal{H} \to \mathbb{R}$ defined by

$$\phi(t,g) = \lambda\|g\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^* g).$$

Since the function $t \mapsto I_t\Phi_t^* g$ is $L_2(\nu)$-measurable for all g it follows from [13, Thm. 1.8, Pg. 5] that it is Borel measurable for all g. Since $\mathcal{R}$ is continuous, it follows that $\phi(\cdot,g)$ is measurable for all

$g \in \mathcal{H}$. On the other hand since $\|I_t\Phi_t^* g\|_{L_2(\nu)} \leq \|I_t\|\|\Phi_t^* g\|_{H_t} \leq \|I_t\|\|g\|_{\mathcal{H}}$ it follows for fixed t that the map $g \mapsto I_t\Phi_t^* g$ is continuous. Since $\mathcal{R}$ is continuous, it then follows that $\phi(t,\cdot)$ is continuous for all $t \in T$. Since $\mathcal{H}$ is separable and complete it is Polish. Therefore, it follows from Carathéodory's Lemma [12, Lem. III.39] that $\phi : T \times \mathcal{H} \to \mathbb{R}$ is measurable. Now, by the strict convexity of the Hilbert space norm, it is easy to see that for $t \in T$ there is a unique solution

$$\mathfrak{g}(t) \;:=\; \arg\min_{g\in\mathcal{H}} \phi(t,g) = \arg\min_{g\in\mathcal{H}} \Big(\lambda\|g\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^* g)\Big).$$

Moreover, since for fixed $g \in \mathcal{H}$ we have

$$\min_{\substack{g'\in\mathcal{H}\\ \Phi_t^* g' = \Phi_t^* g}} \|g'\|_{\mathcal{H}}^2 = \|\Phi_t^* g\|_{H_t}^2 ,$$

we have

$$\min_{\substack{g'\in\mathcal{H}\\ \Phi_t^* g' = \Phi_t^* g}} \Big(\lambda\|g\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^* g)\Big) = \lambda\|\Phi_t^* g\|_{H_t}^2 + \mathcal{R}(I_t\Phi_t^* g) = \mathcal{R}_{\lambda,t}(\Phi_t^* g)$$

and conclude that

$$\lambda\|\mathfrak{g}(t)\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^*\mathfrak{g}(t)) = \min_{g\in\mathcal{H}} \Big(\lambda\|g\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^* g)\Big) = \min_{g\in\mathcal{H}} \mathcal{R}_{\lambda,t}(\Phi_t^* g) = \min_{\hat{g}\in H_t} \mathcal{R}_{\lambda,t}(\hat{g}) = \mathcal{R}_{\lambda,t}^*.$$

Consequently, Aumann's selection principle [29, [Lem. III.39] implies that the function $\mathfrak{g} : t \mapsto \mathfrak{g}(t)$ is measurable and the function $t \mapsto \inf_{g\in\mathcal{H}} \Big(\lambda\|g\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^* g)\Big) = \mathcal{R}_{\lambda,t}^*$ is measurable. Moreover, since $\mathcal{R}_{\lambda,t}^* \leq \mathcal{R}(0)$, $t \in T$ and μ is finite, it follows the integral $\mathbb{E}_{t\sim\mu}\mathcal{R}_{\lambda,t}^*$ exists. Therefore we conclude that

$$\inf_{\mathfrak{f}\in L_2(\mu,\mathcal{H})} \mathbb{E}_{t\sim\mu}\Big(\lambda\|\mathfrak{f}(t)\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^*\mathfrak{f}(t))\Big) \leq \mathbb{E}_{t\sim\mu}\Big(\lambda\|\mathfrak{g}(t)\|_{\mathcal{H}}^2 + \mathcal{R}(I_t\Phi_t^*\mathfrak{g}(t))\Big) = \mathbb{E}_{t\sim\mu}\mathcal{R}_{\lambda,t}^*.$$

The assertion for $\lambda > 0$ then follows from (18).

For the case $\lambda = 0$, it follows from [29, Lem. A.6.4] that $\mathcal{R}_{\lambda,\mu}^*$ and $\mathcal{R}_{\lambda,t}^*$ are increasing and continuous functions of λ for each t. Since $\mathcal{R}_{\lambda,t}^* \leq \mathcal{R}(0), t \in T, \lambda \geq 0$, the extended monotone convergence theorem [4, Thm. 1.6.7] and the assertion for $\lambda > 0$ imply that

$$\mathcal{R}_{0,\mu}^* \leq \mathbb{E}_{t\sim\mu}\mathcal{R}_{0,t}^*.$$

■

Proof of Corollary 3.5: Theorem 3.4, $H_\mu \subset L_2(\nu)$, and the assumptions imply that

$$\mathcal{R}^* \leq \mathcal{R}_{0,\mu}^* \leq \mathbb{E}_{t\sim\mu}\mathcal{R}_{0,t}^* = \mathcal{R}^*$$

establishing the first assertion. Consequently Theorem 3.4 implies

$$A_\mu(\lambda) = \mathcal{R}_{\lambda,\mu}^* - \mathcal{R}_{0,\mu}^* \leq \mathbb{E}_{t\sim\mu}\mathcal{R}_{\lambda,t}^* - \mathcal{R}^* = \mathbb{E}_{t\sim\mu}\big(\mathcal{R}_{\lambda,t}^* - \mathcal{R}^*\big) = \mathbb{E}_{t\sim\mu}A_t(\lambda).$$

■

Proof of Theorem 4.1: Consider $t_1, t_2 > 0$, $x \in X$, and use the integral identity $\mathbb{E}e^{\sigma^{-1}\|\cdot\|_2^2} = (\pi\sigma)^{\frac{d}{2}}$, $\sigma > 0$, to obtain that

$$\begin{aligned}
\langle \Phi_{t_1}(x), \Phi_{t_2}(x)\rangle_{\mathcal{H}} &= \frac{2^d t_1^{\frac{d}{2}} t_2^{\frac{d}{2}}}{\pi^{\frac{d}{2}}} \int_{y\in\mathbb{R}^d} e^{-2(t_1^2+t_2^2)\|x-y\|_2^2} dy \\
&= \frac{2^d t_1^{\frac{d}{2}} t_2^{\frac{d}{2}}}{\pi^{\frac{d}{2}}} \int_{y\in\mathbb{R}^d} e^{-2(t_1^2+t_2^2)\|y\|_2^2} dy \\
&= \Big(\frac{2t_1t_2}{t_1^2+t_2^2}\Big)^{\frac{d}{2}} .
\end{aligned}$$

Therefore we conclude that

$$\begin{aligned}
\|\Phi_{t_1}(x) - \Phi_{t_2}(x)\|_{\mathcal{H}}^2 &= \|\Phi_{t_1}(x)\|_{\mathcal{H}}^2 + \|\Phi_{t_2}(x)\|_{\mathcal{H}}^2 - 2\langle \Phi_{t_1}(x), \Phi_{t_2}(x)\rangle_{\mathcal{H}} \\
&= 2 - 2\Big(\frac{2t_1t_2}{t_1^2+t_2^2}\Big)^{\frac{d}{2}} \\
&= 2 - 2\Big(1 - \frac{|t_1-t_2|^2}{t_1^2+t_2^2}\Big)^{\frac{d}{2}} .
\end{aligned}$$

That is,

$$\|\Phi_{t_1}(x) - \Phi_{t_2}(x)\|_{\mathcal{H}} = \sqrt{2 - 2\Big(1 - \frac{|t_1-t_2|^2}{t_1^2+t_2^2}\Big)^{\frac{d}{2}}} . \tag{19}$$

In addition, by invariance of integration under the translation $y \mapsto y + \frac{x_1+x_2}{2}$ we can show that

$$\begin{aligned}
\langle \Phi_t(x_1), \Phi_t(x_2)\rangle_{\mathcal{H}} &= \frac{2^d t^d}{\pi^{\frac{d}{2}}} \int_{y\in\mathbb{R}^d} e^{-2t^2\|x_1-y\|_2^2} e^{-2t^2\|x_2-y\|_2^2} dy \\
&= \frac{2^d t^d}{\pi^{\frac{d}{2}}} \int_{y\in\mathbb{R}^d} e^{-2t^2\left(\|\frac{x_1-x_2}{2}-y\|_2^2 + \|\frac{x_2-x_1}{2}-y\|_2^2\right)} dy \\
&= \frac{2^d t^d}{\pi^{\frac{d}{2}}} \int_{y\in\mathbb{R}^d} e^{-4t^2\left(\frac{\|x_1-x_2\|_2^2}{4} + \|y\|_2^2\right)} dy \\
&= e^{-t^2\|x_1-x_2\|_2^2} \frac{2^d t^d}{\pi^{\frac{d}{2}}} \int_{y\in\mathbb{R}^d} e^{-4t^2\|y\|_2^2} dy \\
&= e^{-t^2\|x_1-x_2\|_2^2} .
\end{aligned}$$

Consequently, we obtain

$$\|\Phi_t(x_1) - \Phi_t(x_2)\|_{\mathcal{H}} = \sqrt{2 - 2e^{-t^2\|x_1-x_2\|^2}} . \tag{20}$$

Therefore, using the identity $\Phi_{t_1}(x_1) - \Phi_{t_2}(x_2) = \Phi_{t_1}(x_1) - \Phi_{t_2}(x_1) + \Phi_{t_2}(x_1) - \Phi_{t_2}(x_2)$ we conclude that the map $\Phi : \mathbb{R}^+ \times X \to \mathcal{H}$ defined by $\Phi(t,x) := \Phi_t(x)$ is continuous and therefore measurable on $\{t > 0\} \times X$. Since Φ has the constant value z on $\{t = 0\} \times X$ it easily follows that Φ is measurable on $\mathbb{R}^+ \times X$. Since $\mathcal{H} = L_2(\mathbb{R}^d)$ is separable (see e.g. [2, Thm. 2.15]), it follows from [13, Thm. 1.8, Pg. 5] that that Φ is $\mathcal{H}$-measurable, and therefore it is weakly $\mathcal{H}$-measurable. Moreover, since X is a Borel set, it follows from [18, Thm. 1.7.9] that X is Suslin. Therefore, since $\mathcal{H}$ is separable, Lemma 2.1 applies. For the second assertion, observe that (19) implies that for $x \in X$

the function $\Psi_x : \mathbb{R}^+ \to \mathcal{H}$ defined by $\Psi_x(t) = \Phi_t(x)$ is continuous and therefore measurable on $(0, \infty)$. Consequently, Ψ_x is measurable on $\mathbb{R}^+$. Since $\mathcal{H}$ is separable it follows from [13, Thm. 1.8, Pg. 5] that Ψ_x is $\mathcal{H}$-measurable. Moreover, since μ is finite and $\|\Psi_x(t)\|_{\mathcal{H}} = 1, x \in X, t \geq 0$ it follows that $\Psi_x \in L_2(\mu, \mathcal{H}), x \in X$. Therefore, Theorem 2.2 applies.

For the third assertion, observe that since $k_t(x, x) = 1, t \in \mathbb{R}^+, x \in X$, and the measure ν on X is finite, it follows that Lemma 3.1 applies with $\hat{k}(t, \cdot) \equiv 1, t \in \mathbb{R}^+$, and so we also obtain the assertions on the bounds on the inclusions. To show that Theorem 3.2 applies we need to show that the map $\mathbb{R}^+ \to L(\mathcal{H}, L_2(\nu))$ defined by $t \mapsto I_t \Phi_t^*$ is simply $L_2(\nu)$-measurable. To that end, let $g \in \mathcal{H}$ and consider the function $\mathbb{R}^+ \to \mathcal{H}$ defined by $t \mapsto I_t \Phi_t^* g$. For fixed $t > 0$, observe that $\Phi_t(0) = \frac{t^{\frac{d}{2}} 2^{\frac{d}{2}}}{\pi^{\frac{d}{4}}} e^{-2t^2 \|\cdot\|_2^2}$, and hence

$$\big(\Phi_t^* g\big)(x) = \langle g, \Phi_t(x) \rangle_{\mathcal{H}} = \frac{t^{\frac{d}{2}} 2^{\frac{d}{2}}}{\pi^{\frac{d}{4}}} \int_{\mathbb{R}^d} e^{-2t^2 \|x-y\|_2^2} g(y) dy = \big(\Phi_t(0) * g\big)(x)\,,$$

where $*$ is the convolution operator. Therefore, we can write $\Phi_t^* g = \Phi_t(0) * g$. Since Young's inequality [15, Thm. 20.18] and (19) show that

$$\big\|\Phi_{t_1}^* g - \Phi_{t_2}^* g\big\|_\infty = \|(\Phi_{t_1}(0) - \Phi_{t_2}(0)) * g\|_\infty \leq \|\Phi_{t_1}(0) - \Phi_{t_2}(0)\|_{\mathcal{H}} \|g\|_{\mathcal{H}} \leq \sqrt{2 - 2\Big(1 - \frac{|t_1 - t_2|^2}{t_1^2 + t_2^2}\Big)^{\frac{d}{2}}} \|g\|_{\mathcal{H}}$$

we conclude that

$$\|I_{t_1} \Phi_{t_1}^* g - I_{t_2} \Phi_{t_2}^* g\|_{L_2(\nu)} \leq \big\|\Phi_{t_1}^* g - \Phi_{t_2}^* g\big\|_\infty \sqrt{\nu(X)} \leq \sqrt{2 - 2\Big(1 - \frac{|t_1 - t_2|^2}{t_1^2 + t_2^2}\Big)^{\frac{d}{2}}} \|g\|_{\mathcal{H}} \sqrt{\nu(X)}.$$

Consequently, $t \mapsto I_t \Phi_t^* g$ is continuous and therefore measurable on $(0, \infty)$. Moreover the continuity also implies its range is separable. It follows that $t \mapsto I_t \Phi_t^* g$ is measurable on $[0, \infty)$, and has a separable range. Consequently, by [13, Thm. 1.8, Pg. 5], it is $L_2(\nu)$-measurable. Since $g \in \mathcal{H}$ was arbitrary, it follows that $t \mapsto I_t \Phi_t^*$ is simply $L_2(\nu)$-measurable. Since we have shown in the proof of the first assertion that Φ is weakly $\mathcal{H}$-measurable, it follows that Theorem 3.2 applies.

For the fourth assertion, observe that the assumption $1 = k(x, x) = \int k_t(x, x) d\mu(t) = \mu(\mathbb{R}^+)$ implies that μ is a probability measure. Therefore, since $\mathcal{R} : L_2(\nu) \to \mathbb{R}$ is a continuous convex function, Theorem 3.4 applies.

For the last assertion, observe that it follows from [29, Lem. 2.17] that $\mathcal{R} : L_2(P_X) \to [0, \infty)$ is well defined, continuous and convex and therefore Theorem 3.4 applies. Moreover, by [29, Thm. 4.63], the Gaussian RKHSs are known to be dense in $L_2(P_X)$. Consequently, since $X \subset \mathbb{R}^d$ is measurable we have by [29, Thm. 5.31] and the discussion below it that $\mathcal{R}_{0,t}^* = \mathcal{R}^*$ for all $t > 0$. From the assumption $0 = \lim_{x \mapsto \infty} k(x, x')$, $x' \in X$ it follows from the Lebesgue dominated converge theorem that $0 = \lim_{x \mapsto \infty} k(x, x') = \mathbb{E}_{t \sim \mu} \lim_{x \mapsto \infty} k_t(x, x') = \mu(\{0\})$. Therefore Corollary 3.5 applies. ■

Proof of Corollary 4.2: From [1, Eq. 29.3.82] we have that $e^{-\alpha\sqrt{s}}$ is given by

$$e^{-\alpha\sqrt{s}} = \int_{\mathbb{R}^+} e^{-su} \frac{\alpha}{2\sqrt{\pi u^3}} e^{-\frac{\alpha^2}{4u}} du, \quad s \in \mathbb{R}^+.$$

Changing variables by $u = t^2$ we obtain

$$e^{-\alpha\sqrt{s}} = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}^+} e^{-st^2} t^{-2} e^{-\frac{\alpha^2}{4t^2}} dt$$

so that if we consider the Borel probability measure $\mu := \frac{\alpha}{\sqrt{\pi}} t^{-2} e^{-\frac{\alpha^2}{4t^2}} dt$ we have $\mu(\{0\}) = 0$ and $k = k_\mu$. Since the hinge loss function is a P-integrable Nemitski loss of order $p = 1$ we can apply Corollary 3.5 to obtain

$$A_k(\lambda) \leq \mathbb{E}_{t\sim\mu} A_t(\lambda).$$

Now [29, Thm. 8.18] implies that $A_t(\lambda) \leq C_{d,\beta}(\lambda t^d + t^{-\beta})$ where $C_{d,\beta}$ is a constant depending only on (d, β). However, we also have $A_k(\lambda) \leq 1, \lambda \geq 0$. Let

$$\Gamma(z, x) := \int_x^\infty t^{z-1} e^{-t} dt$$

denote the incomplete gamma function, which is well defined for all $z \in \mathbb{R}, x > 0$. We split up the domain of integration into $\mathbb{R}^+ = [0, \frac{\alpha}{2\sqrt{b}}) \cup [\frac{\alpha}{2\sqrt{b}}, \infty)$. Then, for $\kappa \in \mathbb{R}$, by the change of variables $\sigma := \frac{\alpha^2}{4t^2}$, we have

$$\int_{[0,\frac{\alpha}{2\sqrt{b}})} t^\kappa d\mu(t) = \frac{\alpha}{\sqrt{\pi}} \int_{[0,\frac{\alpha}{2\sqrt{b}})} t^{\kappa-2} e^{-\frac{\alpha^2}{4t^2}} dt = \frac{\alpha^\kappa}{2^\kappa \sqrt{\pi}} \int_{(b,\infty)} \sigma^{-\frac{1+\kappa}{2}} e^{-\sigma} ds = \frac{\alpha^\kappa}{2^\kappa \sqrt{\pi}} \Gamma\Big(\frac{1-\kappa}{2}, b\Big).$$

Since $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, we also have

$$\int_{[\frac{\alpha}{2\sqrt{b}},\infty)} d\mu(t) = \frac{\alpha}{\sqrt{\pi}} \int_{[\frac{\alpha}{2\sqrt{b}},\infty)} t^{-2} e^{-\frac{\alpha^2}{4t^2}} dt = \frac{1}{\sqrt{\pi}} \Big(\sqrt{\pi} - \Gamma\Big(\frac{1}{2}, b\Big)\Big).$$

Therefore, using each inequality $A_t(\lambda) \leq C_{d,\beta}(\lambda t^d + t^{-\beta})$ and $A_k(\lambda) \leq 1$ on different components of the split $\mathbb{R}^+ = [0, \frac{\alpha}{2\sqrt{b}}) \cup [\frac{\alpha}{2\sqrt{b}}, \infty)$, we obtain

$$\begin{aligned} A_k(\lambda) &\leq \mathbb{E}_{t\sim\mu} A_t(\lambda) \\ &\leq C_{d,\beta} \int_{[0,\frac{\alpha}{2\sqrt{b}})} \big(\lambda t^d + t^{-\beta}\big) d\mu(t) + \int_{(\frac{\alpha}{2\sqrt{b}},\infty)} d\mu(t) \\ &= C_{d,\beta} \left(\lambda \frac{\alpha^d}{2^d \sqrt{\pi}} \Gamma\Big(\frac{1-d}{2}, b\Big) + \frac{\alpha^{-\beta}}{2^{-\beta}\sqrt{\pi}} \Gamma\Big(\frac{1+\beta}{2}, b\Big) \right) + \frac{1}{\sqrt{\pi}} \Big(\sqrt{\pi} - \Gamma\Big(\frac{1}{2}, b\Big)\Big). \end{aligned}$$

Now consider that

$$\Gamma\Big(\frac{1+\beta}{2}, b\Big) \leq \Gamma\Big(\frac{1+\beta}{2}\Big),$$

$$\frac{1}{\sqrt{\pi}} \Big(\sqrt{\pi} - \Gamma\Big(\frac{1}{2}, b\Big)\Big) = \frac{1}{\sqrt{\pi}} \int_0^b e^{-\sigma} \sigma^{-\frac{1}{2}} d\sigma \leq \frac{1}{\sqrt{\pi}} \int_0^b \sigma^{-\frac{1}{2}} d\sigma \leq \frac{2}{\sqrt{\pi}} b^{\frac{1}{2}},$$

and

$$\Gamma\Big(\frac{1-d}{2}, b\Big) = \int_b^\infty e^{-\sigma} \sigma^{-\frac{d+1}{2}} d\sigma \leq \int_b^\infty \sigma^{-\frac{d+1}{2}} d\sigma = \frac{2}{d-1} b^{\frac{1-d}{2}} \leq 2b^{\frac{1-d}{2}}.$$

Therefore, we obtain

$$A_k(\lambda) \leq C_{d,\beta} \left(\lambda \frac{\alpha^d}{2^{d-1}\sqrt{\pi}} b^{\frac{1-d}{2}} + \frac{\alpha^{-\beta}}{2^{-\beta}\sqrt{\pi}} \Gamma\Big(\frac{1+\beta}{2}\Big) \right) + \frac{2}{\sqrt{\pi}} b^{\frac{1}{2}}.$$

Setting $b := \frac{1}{4}\lambda^{\frac{2}{d}}\alpha^2$, which amounts to the split $\mathbb{R}^+ + = [0, \lambda^{-\frac{1}{d}}) \cup [\lambda^{-\frac{1}{d}}, \infty)$, we see that $\lambda \frac{\alpha^d}{2^{d-1}\sqrt{\pi}} b^{\frac{1-d}{2}} = 2b^{\frac{1}{2}}$ and therefore obtain the assertion by adjusting the value of $C_{d,\beta}$. ■

Proof of Corollary 4.3: Consider the function $g(s) := (1+\frac{s}{m})^{-\alpha}$ so that $k(x,x') = g(\|x-x'\|_2^2)$. Then [1, Eq. 29.3.11] shows that g is given by

$$g(s) = \Big(1+\frac{s}{m}\Big)^{-\alpha} = \frac{m^\alpha}{\Gamma(\alpha)}\int_{\mathbb{R}^+} e^{-su}u^{\alpha-1}e^{-mu}du \quad s\in\mathbb{R}^+.$$

By the change of variable $u := t^2$ we obtain

$$\Big(1+\frac{s}{m}\Big)^{-\alpha} = \frac{2m^\alpha}{\Gamma(\alpha)}\int_{\mathbb{R}^+} e^{-st^2}t^{2\alpha-1}e^{-mt^2}dt.$$

Consequently, for the Borel probability measure $\mu := \frac{2m^\alpha}{\Gamma(\alpha)}t^{2\alpha-1}e^{-mt^2}dt$, we have $\mu(\{0\}) = 0$ and $k = k_\mu$. As in the proof of Corollary 4.2 we have $A_k(\lambda) \le \mathbb{E}_{t\sim\mu}A_t(\lambda)$ and $A_t(\lambda) \le C_{d,\beta}(\lambda t^d + t^{-\beta})$ where $C_{d,\beta}$ is a constant depending only on (d,β). Since for $\kappa > -2\alpha$ we have

$$\mathbb{E}_{t\sim\mu}t^\kappa = \frac{2m^\alpha}{\Gamma(\alpha)}\int_{\mathbb{R}^+} t^{\kappa+2\alpha-1}e^{-mt^2}dt = m^{-\frac{\kappa}{2}}\frac{\Gamma(\frac{\kappa}{2}+\alpha)}{\Gamma(\alpha)},$$

we conclude, for $2\alpha-\beta>0$, that

$$A_k(\lambda) \le \mathbb{E}_{t\sim\mu}A_t(\lambda) \le C_{d,\beta}(\lambda\mathbb{E}_{t\sim\mu}t^d + \mathbb{E}_{t\sim\mu}t^{-\beta}) = C_{d,\beta}\Big(\lambda m^{-\frac{d}{2}}\frac{\Gamma(\frac{d}{2}+\alpha)}{\Gamma(\alpha)} + m^{\frac{\beta}{2}}\frac{\Gamma(\alpha-\frac{\beta}{2})}{\Gamma(\alpha)}\Big).$$

Since $\Gamma(\alpha-\frac{\beta}{2})$ achieves small values near $2\alpha-\beta=1$, but gets large as $2\alpha-\beta\downarrow 0$, we will only use this inequality when $2\alpha-\beta\ge 1$. This establishes the first assertion. When $2\alpha-\beta<1$, we proceed as in the proof of Corollary 4.2. To that end, split the domain of integration into $\mathbb{R}^+ = [0,b)\cup[b,\infty)$. Then, since $\sigma_1^{\alpha-\frac{\beta}{2}-1} \le \sigma_2^{\alpha-\frac{\beta}{2}-1}$, for $\sigma_1\ge\sigma_2>0$, we have

$$\begin{aligned}
\int_{[b,\infty)} t^{-\beta}d\mu(t) = \frac{2m^\alpha}{\Gamma(\alpha)}\int_{[b,\infty)} t^{2\alpha-\beta-1}e^{-mt^2}dt &= \frac{m^{\frac{\beta}{2}}}{\Gamma(\alpha)}\int_{[mb^2,\infty)}\sigma^{\alpha-\frac{\beta}{2}-1}e^{-\sigma}d\sigma \\
&\le \frac{b^{2\alpha-\beta-2}m^{\alpha-1}}{\Gamma(\alpha)}\int_{[mb^2,\infty)} e^{-\sigma}d\sigma \\
&= \frac{b^{2\alpha-\beta-2}m^{\alpha-1}e^{-mb^2}}{\Gamma(\alpha)} \\
&\le \frac{b^{2\alpha-\beta-2}m^{\alpha-1}}{\Gamma(\alpha)},
\end{aligned}$$

and

$$\begin{aligned}
\int_{[0,b)} d\mu(t) = \frac{2m^\alpha}{\Gamma(\alpha)}\int_{[0,b)} t^{2\alpha-1}e^{-mt^2}dt &= \frac{1}{\Gamma(\alpha)}\int_{[0,mb^2)}\sigma^{\alpha-1}e^{-\sigma}d\sigma \\
&\le \frac{1}{\Gamma(\alpha)}\int_{[0,mb^2)}\sigma^{\alpha-1}d\sigma \\
&= \frac{b^{2\alpha}m^\alpha}{\alpha\Gamma(\alpha)}.
\end{aligned}$$

Therefore, using each inequality $A_t(\lambda) \le C_{d,\beta}(\lambda t^d + t^{-\beta})$ and $A_k(\lambda)\le 1$ on different components

of the split $\mathbb{R}^+ = [0,b) \cup [b,\infty)$, we obtain

$$\begin{aligned} A_k(\lambda) &\leq \mathbb{E}_{t\sim\mu} A_t(\lambda) \\ &\leq C_{d,\beta} \int_{[b,\infty)} \big(\lambda t^d + t^{-\beta}\big) d\mu(t) + \int_{[0,b)} d\mu(t) \\ &\leq \lambda C_{d,\beta} \mathbb{E}_{t\sim\mu} t^d + C_{d,\beta} \int_{[b,\infty)} t^{-\beta} d\mu(t) + \int_{[0,b)} d\mu(t) \\ &\leq \lambda C_{d,\beta} m^{-\frac{d}{2}} \frac{\Gamma(\frac{d}{2}+\alpha)}{\Gamma(\alpha)} + C_{d,\beta} \frac{b^{2\alpha-\beta-2} m^{\alpha-1}}{\Gamma(\alpha)} + \frac{\alpha^{-1} b^{2\alpha} m^{\alpha}}{\Gamma(\alpha)} \,. \end{aligned}$$

Setting $b := m^{-\frac{1}{\beta+2}}$ we obtain $b^{2\alpha} m^{\alpha} = b^{2\alpha-\beta-2} m^{\alpha-1} = m^{\frac{\alpha\beta}{\beta+2}}$ and so conclude that

$$A_k(\lambda) \leq C_{d,\beta} \Big(\lambda m^{-\frac{d}{2}} \frac{\Gamma(\frac{d}{2}+\alpha)}{\Gamma(\alpha)} + \frac{m^{\frac{\alpha\beta}{\beta+2}}}{\Gamma(\alpha)} \Big) + \frac{\alpha^{-1} m^{\frac{\alpha\beta}{\beta+2}}}{\Gamma(\alpha)} \,.$$

Adjusting the value of $C_{d,\beta}$ establishes the assertion. ■

Proof of Theorem 4.5: First observe that $\mathbf{1} \in H_0(\mathbb{R}^d)$, the RKHS associated with the kernel $k(x,x') = 1, x, x' \in \mathbb{R}^d$. To prove the assertion, assume to the contrary that $\mathbf{1} \in H_\mu(\mathbb{R}^d)$. Then by Theorem 6.1 there must exist a $\gamma > 0$ such that

$$\gamma^2 k_\mu - \mathbf{1} = \gamma^2 \mathbb{E}_{t\sim\mu} k_t - \mathbf{1} \gg 0 \tag{21}$$

where positive definiteness ($\gg 0$) is defined in (15). Now consider $c > 0$ and n points $x_i \in \mathbb{R}^d, i = 1,..,n$ such that $\|x_i - x_j\|_2 \geq c$, $i \neq j$. If we let $\eta_i := \frac{1}{n}, i = 1,..,n$ we obtain

$$\sum_{i,j=1,..,n} \Big(\gamma^2 k_\mu(x_i,x_j) - \mathbf{1}(x_i,x_j) \Big) \eta_i \eta_j = \gamma^2 \mathbb{E}_{t\sim\mu} \sum_{i,j=1,..,n} k_t(x_i,x_j) \eta_i \eta_j - 1 \,.$$

Let $t^* > 0$ and split the expectation on the right-hand side into

$$\mathbb{E}_{t\sim\mu} \sum_{i,j=1,..,n} k_t(x_i,x_j)\eta_i\eta_j = \int_{t<t^*} \Big(\sum_{i,j=1,..,n} k_t(x_i,x_j)\eta_i\eta_j \Big) d\mu(t) + \int_{t\geq t^*} \Big(\sum_{i,j=1,..,n} k_t(x_i,x_j)\eta_i\eta_j \Big) d\mu(t) \,. \tag{22}$$

Now observe that the integrand in the $t < t^*$ is bounded by 1. Moreover, in the $t \geq t^*$ term we observe that for $i \neq j$ we have

$$k_t(x_i,x_j) \leq e^{-c^2 (t^*)^2} \leq \frac{1}{e} c^{-2} (t^*)^{-2}.$$

Consequently, we obtain

$$\mathbb{E}_{t\sim\mu} \sum_{i,j=1,..,n} k_t(x_i,x_j)\eta_i\eta_j \leq \mu\big([0,t^*)\big) + \frac{\mu\big([t^*,\infty)\big)}{n} \Big(1 + (n-1) \frac{1}{e} c^{-2} (t^*)^{-2} \Big).$$

Therefore, setting $c := \frac{1}{t^*} \sqrt{\frac{n-1}{e}}$ we obtain

$$\sum_{i,j=1,..,n} \Big(\gamma^2 k_\mu(x_i,x_j) - \mathbf{1}(x_i,x_j) \Big) \eta_i \eta_j \leq \gamma^2 \Big(\mu\big([0,t^*)\big) + \frac{2\mu(\mathbb{R}^+)}{n} \Big) - 1$$

Since μ is a finite measure it follows (see e.g. [6, Thm. 3.2]) that $\lim_{t^*\to 0} \mu\big([0,t^*)\big) = \mu(\{0\}) = 0$. Consequently we can choose t^* small enough and n large enough to contradict (21). ■

Proof of Lemma 4.6: From (14) we know that, for all $0 < \alpha_1 \leq \alpha_2$ and $t > 0$, we have $H_{\alpha_1 t} \subset H_{\alpha_2 t}$ and that

$$\|\mathrm{id} : H_{\alpha_1 t} \to H_{\alpha_2 t}\| \leq \left(\frac{\alpha_2}{\alpha_1}\right)^{\frac{d}{2}}.$$

Moreover, it is trivial to observe that these relationships hold also for $t = 0$. Consequently, Theorem 6.1 implies that, for all $0 < \alpha_1 \leq \alpha_2$ and $t \geq 0$, we have

$$\left(\frac{\alpha_2}{\alpha_1}\right)^d k_{\alpha_2 t} - k_{\alpha_1 t} \gg 0.$$

It then follows from Lemma 6.2 that

$$\left(\frac{\alpha_2}{\alpha_1}\right)^d k_{\alpha_2^* \mu} - k_{\alpha_1^* \mu} = \mathbb{E}_{t\sim\mu}\left(\left(\frac{\alpha_2}{\alpha_1}\right)^d k_{\alpha_2 t} - k_{\alpha_1 t}\right) \gg 0.$$

The assertion then follows from Theorem 6.1. ■

Proof of Theorem 4.7: By (14) we know that for all $0 < t \leq t^*$ we have $H_t \subset H_{t^*}$ and that $\|\mathrm{id} : H_t \to H_{t^*}\| \leq \left(\frac{t^*}{t}\right)^{\frac{d}{2}}$ and so by Theorem 6.1 we have $\left(\frac{t^*}{t}\right)^d k_{t^*} - k_t \gg 0$. Consequently by, Lemma 6.2 we have that

$$\left((t^*)^d \mathbb{E}_{t\sim\mu} t^{-d}\right) k_{t^*} - k_\mu = \mathbb{E}_{t\sim\mu}\left(\left(\frac{t^*}{t}\right)^d k_{t^*} - k_t\right) \gg 0$$

and conclude the assertion from Theorem 6.1. ■

Proof of Theorem 4.8: The existence of such intervals follows by decomposing the positive half space into a countable sequence of non-overlapping intervals. For the last assertion, let $\mathcal{H} = L_2(\mathbb{R}^d)$ and consider the feature maps $(\Phi_t)_{t\in\mathbb{R}^+}$ defined at the beginning of Section 4. Let $[t_1, t_2]$ be any interval such that $\mu([t_1, t_2]) > 0$ and, for $s > 0$, let $W_s : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$ be the Gauss-Weierstraß operator defined by

$$\big(W_s g\big)(x) \;:=\; (\pi s)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{-s^{-1}\|y-x\|_2^2} g(y)dy, \quad x \in X,\, g \in L_2(\mathbb{R}^d).$$

Fix $f \in H_{t_1}$ and $g \in L_2(\mathbb{R}^d)$ such that $f \;:=\; \Phi_{t_1}^* g$. From [29, Prop. 4.46] we utilize the fact that $H_{t_1} \subset H_t$, $t \geq t_1$ and that the inclusion map $\mathrm{id}_{t_1,t} : H_{t_1} \to H_t$ satisfies

$$\mathrm{id}_{t_1,t} \circ \Phi_{t_1}^* = \Phi_t^* \left(\frac{t}{t_1}\right)^{\frac{d}{2}} W_{\frac{1}{2}(t^2 - t_1^2)}, \quad t > t_1\,. \tag{23}$$

Define the function $\mathfrak{f} : \mathbb{R}^+ \to L_2(\mathbb{R}^d)$ by

$$\mathfrak{f}(t) \;:=\; \begin{cases} g, & t = t_1 \\ \left(\frac{t}{t_1}\right)^{\frac{d}{2}} W_{\frac{1}{2}(t^2-t_1^2)} g\,, & t_1 < t \leq t_2 \\ 0\,, & t \notin [t_1, t_2]\,. \end{cases}$$

Since $\mathfrak{f}$ is continuous on $[t_1, t_2]$ and 0 elsewhere it follows that it is $L_2(\mathbb{R}^d)$-measurable. Moreover, Young's inequality [15, Thm. 20.18] implies that

$$\|\mathfrak{f}(t)\|_{L_2(\mathbb{R}^d)}^2 \leq \left(\frac{t}{t_1}\right)^d \|g\|_{L_2(\mathbb{R}^d)}^2, \quad t \in \mathbb{R}^+$$

and therefore

$$\|\mathfrak{f}\|^2_{L_2(\mu,\mathcal{H})} = \mathbb{E}_{t\sim\mu}\|\mathfrak{f}(t)\|^2_{L_2(\mathbb{R}^d)} \leq \|g\|^2_{L_2(\mathbb{R}^d)} t_1^{-d} \int_{[t_1,t_2]} t^d d\mu(t) \leq \|g\|^2_{L_2(\mathbb{R}^d)} \Big(\frac{t_2}{t_1}\Big)^d \mu([t_1,t_2]) < \infty. \quad (24)$$

Consequently, $\mathfrak{f} \in L_2(\mu,\mathcal{H})$. However, Equation (23) implies that for all $t_1 < t \leq t_2$ we have

$$\Phi_t^*\mathfrak{f}(t) = \Phi_t^*\Big(\frac{t}{t_1}\Big)^{\frac{d}{2}} W_{\frac{1}{2}(t^2-t_1^2)} g = \Phi_{t_1}^* g = f.$$

Since $\Phi_t^*\mathfrak{f}(t) = 0$, $t \notin [t_1,t_2]$, we conclude that, for all $x \in X$, we have

$$\big(\Psi^*\mathfrak{f}\big)(x) = \mathbb{E}_{t\sim\mu}\Big(\big(\Phi_t^*\mathfrak{f}(t)\big)(x)\Big) = \mu([t_1,t_2])f(x),$$

that is, $\Psi^*\mathfrak{f} = \mu([t_1,t_2])f$. Since $f \in H_{t_1}$ was arbitrary it follows that $H_{t_1} \subset H_\mu$. Moreover, using

$$\|\mu([t_1,t_2])f\|^2_{H_\mu} = \inf_{\substack{\mathfrak{f}\in L_2(\mu,\mathcal{H}) \\ \mu([t_1,t_2])f=\Psi^*\mathfrak{f}}} \|\mathfrak{f}\|^2_{L_2(\mu,\mathcal{H})}$$

and

$$\|f\|^2_{H_{t_1}} = \inf_{\substack{g\in L_2(\mathbb{R}^d) \\ f=\Phi_{t_1}^* g}} \|g\|^2_{L_2(\mathbb{R}^d)}$$

it follows from (24) that

$$\|\mu([t_1,t_2])f\|^2_{H_\mu} \leq t_1^{-d}\|f\|^2_{H_{t_1}} \int_{[t_1,t_2]} t^d d\mu(t).$$

establishing the bound on the inclusion. ■

Proof of Corollary 4.9: Let $k = k_\mu$ where μ is a finite representing measure guaranteed to exist by Theorem 1.1. The nonconstant assumption implies that $\mu(0,\infty) > 0$. It then follows from Theorem 4.8 that there exists a $t_* > 0$ such that $H_{t^*}(\mathbb{R}^d) \subset H_\mu(\mathbb{R}^d)$. Observe that [29, Thm. 4.63] implies that $H_{t^*}(\mathbb{R}^d)$ is dense in $L_p(\nu)$ for all $p \geq 1$ and all finite measures on $\mathbb{R}^d$. Consequently, it follows from $H_{t^*}(\mathbb{R}^d) \subset H_\mu(\mathbb{R}^d)$ that the same is true for $H_\mu(\mathbb{R}^d)$ thus establishing the first assertion. Now assume X is compact. Since [29, Cor. 4.58] implies that H_{t^*} is universal, the universality of H_μ follows from $H_{t^*} \subset H_\mu$. For the third assertion observe that [29, Prop. 4.46] implies that $H_{t_1} \subset H_{t_2}$ for all $t_1 \leq t_2$. This combined with the fact the assumptions imply we can choose t_1 in Theorem 4.8 to be as large as we like completes the proof. Finally, by considering the least squares loss in [29, Theorem 5.31 & Corollary 5.34], the denseness of H_ν in $L_2(\nu)$ for all finite measures ν on $\mathbb{R}^d$, implies that k_μ is strictly positive definite. ■

6 Appendix

We will use the following Theorem of Saitoh [24, Thm. 6, Pg. 37], based on the results of Aronszajn [3, Thms. I & II], connecting positive definiteness (15) and embedding constants.

Theorem 6.1 *Let k_1, k_2 be positive definite functions on X. Then*

$$H_{k_1} \subset H_{k_2}$$

if and only if there exists a constant $\gamma > 0$ such that

$$\gamma^2 k_2 \gg k_1$$

and the minimum of such constants is the norm of the inclusion

$$I : H_{k_1} \to H_{k_2}.$$

Lemma 6.2 *Let X be a set, and T measurable space equipped with a measure μ. Let $(k_t)_{t\in T}$ be a family of positive definite functions on X which is integrable with respect to μ. That is, $k_t \gg 0$, $t \in T$ and $t \mapsto k_t(x, x')$ is integrable with respect to μ for all $x, x' \in X$. Then*

$$\mathbb{E}_{t\sim\mu} k_t \gg 0\,.$$

Proof: Consider $n \in \mathbb{N}$, $x_i \in X$, $a_i \in \mathbb{R}, i = 1, .., n$. By assumption, for all $t \in T$, we have

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

Since the sum is finite, we find that

$$\sum_{i,j=1}^{n} a_i a_j \big(\mathbb{E}_{t\sim\mu} k_t\big)(x_i, x_j) = \int \Big(\sum_{i,j=1}^{n} a_i a_j k_t(x_i, x_j)\Big) d\mu(t) \geq 0\,.$$

■

Theorem 6.3 *Let E be a Banach space, μ a probability measure on a measurable space (T, Σ), and let $\mathfrak{f} : T \to E$ be a Bochner integrable function. Also let $F : E \to \mathbb{R}$ be a continuous convex function. Then the integral $\mathbb{E}_\mu(F \circ \mathfrak{f})$ exists (with possible value $+\infty$) and we have Jensen's inequality*

$$F(\mathbb{E}_\mu \mathfrak{f}) \leq \mathbb{E}_\mu(F \circ \mathfrak{f})$$

where, on the left, $\mathbb{E}_\mu$ denotes Bochner integration.

Proof: We follow the proof for real Borel functions in [28, Pg. 192]. The assumptions and [5, Cor. 2.1] imply that F is subdifferentiable everywhere. That is $\partial F(f) \neq \emptyset, f \in E$, where $\partial F(f)$ is the subdifferential of F at f. Then for $z_0 \in E, z^* \in \partial F(z_0)$ and for all $t \in T$ we have

$$\big(F \circ \mathfrak{f}\big)(t) = F(\mathfrak{f}(t)) \geq F(z_0) + z^*\big(\mathfrak{f}(t) - z_0\big).$$

Now since

$$E_{t\sim\mu} z^*\big(\mathfrak{f}(t) - z_0\big) = z^*\big(\mathbb{E}_{t\sim\mu}\mathfrak{f}(t) - z_0\big) > -\infty$$

it follows that the function $F \circ \mathfrak{f}$ is bounded below by an integrable function. Since F is continuous the function $F \circ \mathfrak{f}$ is Borel measurable. Consequently, we conclude from [4, Thm. 1.5.9] that $\mathbb{E}_\mu(F \circ \mathfrak{f})$ exists. Therefore we obtain

$$\mathbb{E}_\mu(F \circ \mathfrak{f}) \geq F(z_0) + \mathbb{E}_{t\sim\mu} z^*\big(\mathfrak{f}(t) - z_0\big) = F(z_0) + z^*\big(\mathbb{E}_\mu \mathfrak{f} - z_0\big)$$

and substituting $z_0 := \mathbb{E}_\mu \mathfrak{f}$ we obtain the assertion. ■

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1970.

[2] R. A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

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

[4] R. B. Ash. *Real Analysis and Probability*. Academic Press, New York, 1972.

[5] V. Barbu and Th. Precupanu. *Convexity and Optimization in Banach Spaces*. D. Reidel Publishing Company, Dordrecht, 1986.

[6] H. Bauer. *Measure and Integration Theory*. Walter de Gruyter, Berlin, 2001.

[7] C. Berg, J.P.R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer, New York, 1984.

[8] S. N. Bernstein. Sur les fonctions abolument monotones. *Acta Mathematica*, 52:1–66, 1928.

[9] S. Bochner. Monotone Funktionen, Stieltjes Integrale und harmonische Analyse. *Math. Annalen*, 108:378–410, 1933.

[10] S. Bochner and A. E. Taylor. Linear functionals on certain spaces of abstractly-valued functions. *Ann. of Math.*, 39:913–944, 1938.

[11] A. Cannon, J. Howse, D. Hush, and C. Scovel. Learning with the Neyman–Pearson and min–max criteria. `http://wwwc3.lanl.gov/~kelly/ml/pubs/2002_svm_bounds/paper.ps`, 2002.

[12] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*. Springer-Verlag, Berlin, 1977.

[13] N. Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*. Wiley Interscience, New York, 2000.

[14] N. Dunford and J. T. Schwartz. *Linear Operators*. Interscience, 1958.

[15] E. Hewitt and K. A. Ross. *Abstract Harmonic Analysis I*. Springer-Verlag, New York, 1997.

[16] D. Hush, P. Kelly, C. Scovel, and I. Steinwart. QP algorithms with guaranteed accuracy and run time for support vector machines. *J. Mach. Learn. Res.*, 7:733–769, 2006.

[17] D. Hush and C. Scovel. Polynomial-time decomposition algorithms for support vector machines. *Machine Learning*, 51:51–71, 2003.

[18] S. G . Krantz and H. R. Parks. *Geometric Integration Theory*. Birkhäuser, Boston, 2008.

[19] J. Lindenstrauss and L. Tzafriri. *Classical Banach Spaces I and II*. Springer, Berlin, 1977.

[20] N. List and H. U. Simon. General polynomial time decomposition algorithms. *J. Mach. Learn. Res.*, 8:303–321, 2007.

[21] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *J. Mach. Learn. Res.*, 7:2651–2667, 2006.

[22] E. H. Moore. On properly positive Hermitian matrices. *Bull. Amer. Math. Soc.*, 23:59, 1916.

[23] F. Qi. Bounds for the ratio of two gamma functions-from Wendel's and related inequalities to logarithmically completely monotone functions. `http://scholar.google.com/scholar?hl=en&q=fengqiboundsratio&um=1&ie=UTF-8&sa=N&tab=ws`, 2009.

[24] S. Saitoh. *Integral transforms, reproducing kernels and their applications.* Addison Wesley Longman, 1997.

[25] I. J. Schoenberg. Metric spaces and completely monotone functions. *Ann. of Math.*, 39:811–841, 1938.

[26] C. Scott. Performance measures for Neyman-Pearson classification. *IEEE Transactions on Information Theory*, 55(8):2852–2863, 2007.

[27] C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Signal Processing*, 52:2264–2274, 2005.

[28] A. N. Shiryaev. *Probability.* Springer-Verlag, New York, 2 edition, 1984.

[29] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer-Verlag, 2008.

[30] I. Steinwart, D. Hush, and C. Scovel. A new concentration result for regularized risk minimizers. In E. Giné, V. Koltchinskii, W. Li, and J. Zinn, editors, *High Dimensional Probability: Proceedings of the Fourth International Conference, Beachwood, Ohio, USA*, volume 51 of *IMS Lecture Notes Monograph Series*, pages 260–275, 2006.

[31] I. Steinwart, D. Hush, and C. Scovel. An oracle inequality for clipped regularized risk minimizers. *Neural Information Processing Systems(NIPS)*, 19:1321–1328, 2007.

[32] I. Steinwart and C. Scovel. Fast rates for support vector machines. In P. Auer and R. Meir, editors, *Proceedings of the 18th Conference on Learning Theory (COLT 2005), Bertinoro, Italy*, volume 3559 of *Lecture Notes in Computer Science*, pages 279–294, 2005.

[33] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35(2):575–607, 2007.

[34] T. O. To and Y. K. Wing. A generalized Jensen's inequality. *Pacific J. Math.*, 58:255–259, 1975.

[35] D. V. Widder. *The Laplace Transform.* Princeton Univ. Press, 1946.