Geometric Learning with Positively Decomposable Kernels

Nathaël Da Costa   NATHAEL.DACOSTA@GMAIL.COM
Cyrus Mostajeran   CYRUS.MOSTAJERAN@GMAIL.COM
Juan-Pablo Ortega   JUAN-PABLO.ORTEGA@NTU.EDU.SG
Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University
21 Nanyang Link, 637371, Singapore

Salem Said   SALEM.SAID@UNIV-GRENOBLE-ALPES.FR
Laboratoire Jean Kuntzman
Université Grenoble-Alpes
Grenoble, 38400, France

Editor: Jean-Philippe Vert

Abstract

Kernel methods are powerful tools in machine learning. Classical kernel methods are based on positive definite kernels, which enable learning in reproducing kernel Hilbert spaces (RKHS). For non-Euclidean data spaces, positive definite kernels are difficult to come by. In this case, we propose the use of reproducing kernel Krein space (RKKS) based methods, which require only kernels that admit a positive decomposition. We show that one does not need to access this decomposition to learn in RKKS. We then investigate the conditions under which a kernel is positively decomposable. We show that invariant kernels admit a positive decomposition on homogeneous spaces under tractable regularity assumptions. This makes them much easier to construct than positive definite kernels, providing a route for learning with kernels for non-Euclidean data. By the same token, this provides theoretical foundations for RKKS-based methods in general.

Keywords: kernel methods, geometric learning, Krein spaces, non-positive kernels, invariant kernels, Gaussian kernel, homogeneous spaces, symmetric spaces.

Contents

1 Introduction

Kernel methods have proved to be successful in machine learning. Classically, kernel methods rely on a positive definite (PD) kernel on the data space. Such a PD kernel gives an embedding of the data space into a reproducing kernel Hilbert space (RKHS), which is a space of functions over the data space. Learning problems are then often phrased as learning a function in the RKHS, which can be done effectively thanks to the Representer Theorem (Schölkopf et al., 2001).This theorem makes kernels a particularly attractive paradigm as, combined with universality results (Steinwart, 2001; Micchelli et al., 2006; Sriperumbudur et al., 2011; Simon-Gabriel and Schölkopf, 2018), it provides a learning guarantee. In addition, kernels allow for uncertainty quantification by using Gaussian processes (Rasmussen and Williams, 2005; Kanagawa et al., 2018).

Kernel methods on Euclidean data spaces have been thoroughly studied, and numerous families of PD kernels have been proposed in Euclidean spaces, many of which have universal approximation properties (Steinwart, 2001; Micchelli et al., 2006). However, in many applications, it is important to capture the geometry of the data and view it as lying on a non-Euclidean space, such as a manifold. There has been considerable recent interest in constructing PD kernels on manifolds (Jayasumana et al., 2013; Feragen et al., 2015; Borovitskiy et al., 2020; Azangulov et al., 2024a,b; Da Costa et al., 2023b). It has proved difficult to find closed-form PD kernels that are easy to implement and evaluate on general geometries.

Ong et al. (2004) proposed a different class of kernel algorithms that do not require positive definiteness of the kernel. Instead, they require the existence of a positive (PD) decomposition for the kernel, meaning that it can be written as a difference of PD kernels. Such a kernel embeds the data space into a reproducing kernel Krein space (RKKS), consisting of functions on the data space. As in the RKHS case, this allows for the solution of learning problems from data (Schleif and Tino, 2015; Loosli et al., 2016; Oglic and Gaertner, 2018; Liu et al., 2021a,b) via, for instance, an adapted representer theorem.

We are interested in the problem of characterizing kernels that admit a PD decomposition, both on Euclidean and non-Euclidean data spaces. Providing such a characterization promises to solve several open problems in the literature. More specifically, it achieves the following.

1. *It justifies the use of non-PD kernels for RKHS-based methods.* In applied areas, non-PD kernels have sometimes been used when restricted to a data set for which the Gram matrix corresponding to this kernel is PD. This has, for example, been done with the geodesic Gaussian kernel on non-Euclidean geometries (Calinon, 2020; Jousse et al., 2021), which is known not to be PD in general (Feragen et al., 2015; Da Costa et al., 2023a,b; Li, 2023). Classical RKHS learning theory does not provide theoretical justification for the use of such kernels, as they do not define an RKHS over the data space. However many RKHS-based methods (such as support vector

machines (Cristianini and Ricci, 2008)) may be generalized to RKKS-based methods (Loosli et al., 2016), such that the solution to the respective learning problems coincide when restricted to a finite data set on which the kernel is PD. Therefore applying an RKHS-based method to a non-PD kernel, which is PD on a given finite data set, can be justified by the existence of a PD decomposition for the kernel.

2. *It justifies the use of RKKS-based methods.* Analogous to how a kernel may be PD on a finite data set but not on the whole space, a kernel may be PD decomposable on a finite data set but not on the whole space. Except that in this case, this subtlety is even easier to ignore, as kernels are *always* PD decomposable on finite data sets. However, if a kernel does not admit a PD decomposition *on the whole* data space, then the solution to the learning problem has no guarantees to generalize to unseen data. RKKS-based methods have so far largely ignored the issue of checking PD decomposability of the kernels used. Only linear combinations of PD kernels were truly known to admit a PD decomposition. Recently, a theoretical advance was made by Liu et al. (2021a) by observing that invariant kernels on $X = \mathbb{R}^n$ admit a PD decomposition if and only if they are the inverse Fourier transform of a finite signed measure. We will generalize this result to many non-Euclidean data spaces X.

3. *It motivates the use of RKKS-based methods for non-Euclidean data spaces.* In the Euclidean case $X = \mathbb{R}^n$, many families of kernels are known to be PD. RKHS-based methods are, therefore, usually sufficient in this case. In contrast, for non-Euclidean data spaces X, PD kernels can be challenging to construct. For example, it has been shown that the generalization of some of the most widely used kernels in $\mathbb{R}^n$, such as the Gaussian kernel, are often not PD on non-Euclidean geometries (Feragen et al., 2015; Da Costa et al., 2023a,b; Li, 2023). The positive definiteness of a kernel is a stringent condition and heavily depends on the specific geometry of the data space in question. This makes it difficult to describe general families of kernels that are PD on general classes of geometries. Such families can sometimes be expressed as infinite series expansions on compact manifolds or as integrals on non-compact ones (Borovitskiy et al., 2020; Azangulov et al., 2024a,b). Closed-form PD kernels, however, have only been found on a handful of geometries (Feragen et al., 2015; Da Costa et al., 2023b). In contrast, we will see that PD decomposability is a much weaker condition to impose on a kernel, and it will be enough to require some symmetry and regularity assumptions. Therefore, in non-Euclidean settings, RKKS-based methods become particularly compelling.

1.1 Paper organization and contributions

This paper aims at providing verifiable sufficient conditions for kernels to admit a positive decomposition, thereby justifying their use for RKKS learning. We pay particular attention to *invariant* kernels on locally compact groups G and their coset spaces G/H, and then look into the special case where G is a Lie group, and G/H is a homogeneous space.

We start by briefly reviewing the theory of PD kernels and how they enable RKHS learning in Section 2.1. We then describe the analogous theory of PD decomposable kernels and RKKS learning in Section 2.2. In particular, we formalize a general representer

theorem for RKKS learning. Crucially, we show that, when the regularizer of the learning problem is linear in the squared indefinite inner product, one does not need access to the PD decomposition of the kernel in order to apply the representer theorem; one only needs to know that it exists. We then discuss the extent to which RKHS methods can be adapted to the RKKS framework. In Section 2.3 we review the few examples where PD decompositions are known to exist.

In Section 3 we focus on invariant kernels on homogeneous spaces of locally compact groups and prove the correspondence between (PD/PD decomposable) invariant kernels on G/H and (PD/PD decomposable) Hermitian functions on the double coset space $H\backslash G/H$. When $H = \{e\}$, this allows us to relate PD decomposable invariant kernels on G to real-valued functions in the Fourier-Stieltjes algebra $B_{\mathbb{C}}(G)$, which in turn allows us to leverage results from harmonic analysis.

In Section 4, we study the case where G is commutative, where the situation simplifies considerably. In particular, we review known sufficient conditions for functions to belong to the spaces $B_{\mathbb{C}}(\mathbb{R}^n)$ and $B_{\mathbb{C}}(S^1)$. As an example, we show that the Gaussian kernel, despite failing to be PD, for instance, on the torus $\mathbb{T}^n$, has a PD decomposition on any Abelian Lie group.

In Section 5, we move to the case where G is not assumed to be commutative. In this case the literature has focused on the space of functions $B_{\mathbb{C}}(G)$, and not the more general $B_{\mathbb{C}}(H\backslash G/H)$. To fill this gap, we show that any function that is invariant on (double) cosets and is PD decomposable has a PD decomposition into functions that are invariant on (double) cosets, i.e. $B_{\mathbb{C}}(G) \cap C_{\mathbb{C}}(G/H) = B_{\mathbb{C}}(G/H)$ and $B_{\mathbb{C}}(G) \cap C_{\mathbb{C}}(H\backslash G/H) = B_{\mathbb{C}}(H\backslash G/H)$. This allows us to show in Section 5.1 that on a homogeneous space G/H, with G a unimodular Lie group and H a compact Lie subgroup, smooth functions whose derivatives decay appropriately at infinity admit PD decompositions. As an example, we then show that the Gaussian kernel has a PD decomposition on non-compact symmetric spaces.

2 Kernels and Learning

In this section we review how positive definite and positively decomposable kernels can be used for learning.

2.1 Learning with positive definite kernels

Definition 1 *A kernel on a set X is a Hermitian map $k : X \times X \to \mathbb{C}$.*

In practice, we are interested in real-valued kernels $k : X \times X \to \mathbb{R}$, but allowing kernels to take complex values will be convenient in the following theoretical work.

Definition 2 *A kernel k on a set X is said to be positive definite (PD) if for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in X$ and all $c_1, \ldots, c_N \in \mathbb{C}$,*

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \overline{c_i} c_j k(x_i, x_j) \geq 0$$

i.e. the matrix $\big(k(x_i, x_j)\big)_{i,j}$, which we call the Gram matrix of $x_1, \ldots, x_N$, is Hermitian positive semidefinite.

We refer to Berg et al. (1984, Chapter 3) for the general theory of PD kernels.

In machine learning, a key reason for the importance of PD kernels is the following theorem (see Paulsen and Raghupathi (2016, Theorem 3.16)) that guarantees that a PD kernel has a natural associated reproducing kernel Hilbert space (RKHS):

Theorem 3 *Let k be a PD kernel on a set X. Then there is a complex Hilbert space $\mathcal{H}$ of complex-valued functions, which we call the reproducing kernel Hilbert space (RKHS) associated with k, with a map*

$$\begin{aligned} \Phi : X &\to \mathcal{H} \\ x &\mapsto k(x, \cdot) \end{aligned} \tag{1}$$

such that

1. *$\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = k(x, y)$ for all $x, y \in X$,*

2. *$\overline{\text{span}(\Phi(X))} = \mathcal{H}$.*

In the above theorem $\overline{\text{span}(\Phi(X))}$ stands for the completion of $\text{span}(\Phi(X))$ with respect to the RKHS norm.

The map Φ is called the *feature map* of the RKHS $\mathcal{H}$. The functions in the image of Φ in (1) are called *kernel sections*, or *canonical basis functions*, of k. In practice, kernels will be real-valued, so we will usually consider the induced *real* RKHS, defined as the subspace of real-valued functions of $\mathcal{H}$.

Suppose we are trying to learn a real or complex-valued function f over X from a finite set of observations. One can try to find a good approximation for this function in the RKHS associated with a particular kernel map k. This is particularly pertinent when the kernel in question has universality properties (Steinwart, 2001; Micchelli et al., 2006; Sriperumbudur et al., 2011; Simon-Gabriel and Schölkopf, 2018), that is, when its sections are dense, say, in the set of continuous functions on compact subsets of X. The representer theorem in RKHS (Schölkopf et al., 2001) tells us that, for a general class of learning problems, the best approximation f^* in the RKHS, given finitely many observations of the function, is a finite linear combination of the kernel sections $k(x_i, \cdot)$. More specifically, given

1. $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset (X \times \mathbb{R})^N$ a finite data set,

2. $g : [0, \infty) \to \mathbb{R}$ a strictly increasing function,

3. $L : \mathcal{H} \times (X \times \mathbb{R})^N \to \mathbb{R}$ a loss functional, determined exclusively through function evaluations, i.e. if $f, g \in \mathcal{H}$ are such that $f|_{\mathcal{D}} = g|_{\mathcal{D}}$, then $L(f, \mathcal{D}) = L(g, \mathcal{D})$,

4. $\Omega(\mathcal{D}) \subset \mathcal{H}$ a feasible set, determined exclusively through function evaluations, i.e. if $f, g \in \mathcal{H}$ are such that $f|_{\mathcal{D}} = g|_{\mathcal{D}}$, then $f \in \Omega(\mathcal{D})$ if and only if $g \in \Omega(\mathcal{D})$.

Then any solution to the minimization problem

$$\begin{aligned} \underset{f \in \mathcal{H}}{\text{minimize}} \ & L(f, \mathcal{D}) + g(\|f\|_{\mathcal{H}}) \\ \text{s.t. } & f \in \Omega(\mathcal{D}) \end{aligned} \tag{2}$$

has the form

$$f^* = \sum_{i=1}^{N} \alpha_i k(x_i, \cdot)$$

for some $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$.

The problem is therefore reduced to finding the coefficients α_i. We have thus transformed the infinite-dimensional problem of learning a function into a finite-dimensional linear algebra problem. In the simplest case, this linear algebra problem is essentially linear regression.

Example 4 Kernel ridge regression (KRR) is a learning algorithm for fitting a function in the RKHS to noisy observations. Given a PD kernel k on a set X, $\mathcal{H}$ the RKHS associated to k, and a finite data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset (X \times \mathbb{R})^N$, the learning problem solved by KRR is given by

$$\underset{f \in \mathcal{H}}{\text{minimize}} \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + c\|f\|_{\mathcal{H}}^2 \tag{3}$$

for some $c > 0$. The representer theorem tells us that the solution to this problem can be written in the form $f^* = \sum_{i=1}^{N} \alpha_i k(x_i, \cdot)$ for some $\alpha_i \in \mathbb{R}$. Plugging this in (3) and solving for the minimum we find the solution $\alpha = (K + NcI)^{-1} y$, where $K = \left(k(x_i, x_j)\right)_{i,j}$ is the Gram matrix and $y = (y_i)_i$.

Example 5 Support vector machines (SVM) (Cristianini and Ricci, 2008) are a class of learning algorithms for classification. Given a finite data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset (X \times \{-1, 1\})^N$, and k, X, $\mathcal{H}$ as above, the learning problem solved by binary SVM can be written as

$$\underset{f \in \mathcal{H}, \, b \in \mathbb{R}}{\text{minimize}} \|f\|_{\mathcal{H}}^2 \quad \text{s.t.} \quad \sum_{i=1}^{N} \max(0, 1 - y_i(f(x_i) + b)) \leq \tau \tag{4}$$

for some fixed parameter $\tau > 0$. Once again, the representer theorem gives us a closed-form solution to this problem. Note that to be precise, the parametric extension to the representer theorem (Schölkopf et al., 2001, Theorem 2) is applied here, since the minimization is over f and b, and not solely over f.

The field of positive definite kernel methods extends beyond the use of the representer theorem. Gaussian processes (GPs) are the Bayesian analogue to frequentist kernel methods, with the posterior mean of GP regression corresponding to the solution KRR being one of many correspondences between the two paradigms (Kanagawa et al., 2018). Kernels can also be used to define distances on the space of probability measures through maximum mean discrepancy (Gretton et al., 2012).

2.2 Learning with positively decomposable kernels

A natural question that can be posed is whether the assumption on the kernel k being positive definite is necessary. That is, can we still obtain a representer theorem while dropping this assumption? It turns out that we can. For this, we need to define the notion

of a Krein space. We refer to Bognár (1974) for a thorough presentation of Krein spaces and to Schwartz (1964) and Alpay (1991) for the study of their reproducing kernels. This last reference contains a good summary of Schwartz's contributions.

Definition 6 *An indefinite inner product on a real vector space $\mathcal{K}$ is a Hermitian bilinear map $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ which is non-degenerate, that is, for all $f \in \mathcal{K}$,*

$$\langle f, f \rangle = 0 \quad implies \ that \quad f = 0.$$

A complex vector space $\mathcal{K}$ equipped with an indefinite inner product $\langle \cdot, \cdot \rangle$ is called a Krein space if it can be written as the algebraic direct sum

$$\mathcal{K} = \mathcal{H}_+ \oplus \mathcal{H}_- \tag{5}$$

such that

1. *$\mathcal{H}_+$ equipped with $\langle \cdot, \cdot \rangle_+ := \langle \cdot, \cdot \rangle$ is a Hilbert space,*

2. *$\mathcal{H}_-$ equipped with $\langle \cdot, \cdot \rangle_- := -\langle \cdot, \cdot \rangle$ is a Hilbert space,*

3. *$\langle f_+, f_- \rangle = 0$ for all $f_+ \in \mathcal{H}_+$, $f_- \in \mathcal{H}_-$.*

In other words, there are complete inner products $\langle \cdot, \cdot \rangle_+$, $\langle \cdot, \cdot \rangle_-$ on $\mathcal{H}_+$, $\mathcal{H}_-$, respectively, such that

$$\langle f, g \rangle = \langle f_+, g_+ \rangle_+ - \langle f_-, g_- \rangle_-$$

for all $f = f_+ + f_-$, $g = g_+ + g_-$ with $f, g \in \mathcal{K}$, $f_+, g_+ \in \mathcal{H}_+$, $f_-, g_- \in \mathcal{H}_-$.

  $\mathcal{K}$ is equipped with the product topology on $\mathcal{H}_+ \times \mathcal{H}_-$, which can be shown to be independent of the choice of decomposition (5).

Definition 7 *Let k be a kernel on X. Then k is said to have a positive (PD) decomposition if it can be written as*

$$k = k_+ - k_-$$

where k_+ and k_- are PD kernels.

Remark 8 PD decompositions are also sometimes called *fundamental decompositions* (Bognár, 1974), or *Kolmogorov decompositions* (Mary, 2003).

With this notion, we obtain an analogous result to Theorem 3 (see Alpay (1991, Theorem 2.1)):

Theorem 9 *Let k be a PD decomposable kernel on a set X. Then there is a complex Krein space $\mathcal{K}$ of complex-valued functions, which we call the reproducing kernel Krein space (RKKS) associated to k, with a map*

$$\Phi : X \to \mathcal{K}$$
$$x \mapsto k(x, \cdot).$$

such that

7

1. $\langle \Phi(x), \Phi(y) \rangle_{\mathcal{K}} = k(x, y)$ *for all* $x, y \in X$,

2. $\overline{\text{span}(\Phi(X))} = \mathcal{K}$.

Remark 10 As opposed to the PD case where the RKHS associated to a kernel is unique (see Paulsen and Raghupathi (2016, Theorem 3.16 & following Remark)), the RKKS associated to a PD decomposable kernel k can, in some cases, be non-unique; see Schwartz (1964) and Alpay (1991). However, such counterexamples rely on non-trivial constructions, which we do not expect to be relevant for 'typical' kernels.

Theorem 9 can be made into an if and only if statement: k must admit a PD decomposition in order to give rise to an RKKS, in the same way that k must be PD in order to give rise to an RKHS. Indeed, a PD decomposition for k can be obtained given a decomposition of the RKKS as in (5).

As in the PD case, we will usually consider only the induced *real* RKKS of a real-valued PD decomposable kernel, defined as the subspace of real-valued functions of $\mathcal{K}$.

Now, as in the RKHS case, a representer theorem can be formulated in the context of RKKS. Such a representer theorem in RKKS was formulated in Ong et al. (2004, Theorem 11). The indefinite nature of the inner product implies that the optimization problem (2) is replaced by a critical point problem. More specifically, given

1. $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset (X \times \mathbb{R})^N$ a finite data set,

2. $g : \mathbb{R} \to \mathbb{R}$ a strictly monotonic differentiable function,

3. $L : \mathcal{K} \times (X \times \mathbb{R})^N \to \mathbb{R}$ a loss functional, determined exclusively through function evaluations, Fréchet differentiable in the first argument,

4. $I : \mathcal{K} \times (X \times \mathbb{R})^N \to \mathbb{R}^m$ and $E : \mathcal{K} \times (X \times \mathbb{R})^N \to \mathbb{R}^l$ functionals determined exclusively through function evaluations and Fréchet differentiable in the first argument,

under some appropriate additional regularity condition on I and E (see Appendix A), any solution to the stabilization (critical point) problem

$$\underset{f \in \mathcal{K}}{\text{stabilize}} \; L(f, \mathcal{D}) + g(\langle f, f \rangle_{\mathcal{K}})$$
$$\text{s.t. } I(f, \mathcal{D}) \leq 0 \qquad (6)$$
$$E(f, \mathcal{D}) = 0$$

has the form

$$f^* = \sum_{i=1}^{N} \alpha_i k(x_i, \cdot),$$

for some $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$. Such α_i may be found by stabilizing the Lagrangian

$$\mathcal{L}(f) = L(f, \mathcal{D}) + g(\langle f, f \rangle_{\mathcal{K}}) + \lambda^\top I(f, \mathcal{D}) + \mu^\top E(f, \mathcal{D})$$

with respect to α, that is, solving for $\nabla_\alpha \mathcal{L}(f^*) = 0$, and then finding the appropriate Lagrange multipliers $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^l$ such that $I(f^*, \mathcal{D}) \leq 0$ and $E(f^*, \mathcal{D}) = 0$.

Explicitly, f^* is given by

$$f^* = -\frac{1}{2g'(\langle f^*, f^* \rangle_{\mathcal{K}})} \sum_{i=1}^{N} \left(\partial_i L(f^*, \mathcal{D}) + \lambda^\top \partial_i I(f^*, \mathcal{D}) + \mu^\top \partial_i E(f^*, \mathcal{D}) \right) k(x_i, \cdot)$$

where the ∂_i denotes the derivative with respect to the ith evaluation. Importantly, note that in the generic case $g(x) = cx$ for some $c \in \mathbb{R}$ and all $x \in \mathbb{R}$, we have $g'(\langle f^*, f^* \rangle_{\mathcal{K}}) = c$, and hence the α_i do not depend on the choice of indefinite inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ associated to $\mathcal{K}$. This means that in this case *one does not need to access the positive decomposition of the kernel in order to compute the solution, one only needs to know of its existence.*

We emphasize that, in view of this result, the situation for RKKS is analogous to the RKHS case in the sense that the solution of the stabilization problem is *in the span of the data*, although this adapted representer theorem is designed for the study of critical points, as opposed to minima of a loss functional, and thus algorithms must be adapted to this setting (Hassibi et al., 1999).

The reason for stabilizing, as opposed to minimizing, the regularized loss functional $L(f, \mathcal{D}) + g(\langle f, f \rangle_{\mathcal{K}})$ is not that the proof of the RKKS representer theorem fails for minimization; it doesn't. Rather, the reason is that it is not expected that the regularized loss functional admits a minimum since $\langle f, f \rangle_{\mathcal{K}}$, as opposed to $\|f\|_{\mathcal{H}}$, is not bounded below for $f \in \mathcal{K}$, since $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ is indefinite. Thus, solutions to learning problems in RKKS may take the form of a saddle point instead.

Note that we have not described precisely what is meant by a critical point of the learning problem (6). This is a subject that has not received much attention in the RKKS learning literature, possibly due to the difficulty of defining a critical point of a constrained optimization problem in function space. In Appendix A, we thus expand on the work from Ong et al. (2004, Theorem 11) by providing a formal definition of such a critical point through Fréchet derivatives, and outline a proof for the resulting representer theorem.

Below we present a few examples of the use of the RKKS representer theorem.

Example 11 Kernel ridge regression (c.f. Example 4) can be generalized to PD decomposable kernels (Ong et al., 2004). Given a PD decomposable kernel k on a set X, $\mathcal{K}$ the RKKS associated to k, and a finite data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset (X \times \mathbb{R})^N$, the learning problem solved by KRR is given by

$$\underset{f \in \mathcal{H}}{\text{stabilize}} \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + c\langle f, f \rangle_{\mathcal{K}} \tag{7}$$

for some $c \in \mathbb{R}$. The representer theorem tells us that the solution to this problem can be written in the form $f^* = \sum_{i=1}^{N} \alpha_i k(x_i, \cdot)$ for some $\alpha_i \in \mathbb{R}$. Plugging this in (7) and solving for the critical point we find the solution $\alpha = (K + NcI)^{-1}y$, where $K = \left(k(x_i, x_j) \right)_{i,j}$ is the Gram matrix and $y = (y_i)_i$, assuming $-Nc$ is not an eigenvalue of K. We see that when $-Nc$ is an eigenvalue of K, there may be infinitely many solutions (if Ky is in the image of $K(K + NcI)$) or no solution at all (if Ky is not in the image of $K(K + NcI)$).

Example 12 Support vector machines (c.f. Example 5) have been generalized to PD decomposable kernels in Loosli et al. (2016). Given a finite data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$

$\subset (X \times \{-1,1\})^N$, and k, X, $\mathcal{K}$ as above, the analogue to the learning problem (4) can be written as

$$\underset{f \in \mathcal{K},\, b \in \mathbb{R}}{\text{stabilize}}\ \langle f, f \rangle_{\mathcal{K}} \quad \text{s.t.} \quad \sum_{i=1}^{N} \max(0, 1 - y_i(f(x_i) + b)) \leq \tau,$$

for some fixed parameter $\tau > 0$. The representer theorem tells us that any solution (f, b) to this problem has the form $f = \sum_{i=1}^{N} \alpha_i k(x_i, \cdot)$ for some $\alpha_1, \dots, \alpha_N \in \mathbb{R}$. We emphasize again that we are applying here the semiparametric extension to the representer theorem (Ong et al., 2004, Theorem 12), since the stabilization is over f and b, and not solely over f. See Figure 1 for an application.

Example 13 More generally, given an RKHS $\mathcal{H}$ and an RKKS $\mathcal{K}$, and a learning problem

$$\underset{f \in \mathcal{H}}{\text{minimize}}\ L(f, \mathcal{D}) + g(\|f\|_{\mathcal{H}}^2)$$
$$\text{s.t.}\ I(f, \mathcal{D}) \leq 0$$
$$E(f, \mathcal{D}) = 0,$$

one can consider the learning problem in RKKS

$$\underset{f \in \mathcal{K}}{\text{stabilize}}\ L(f, \mathcal{D}) + g(\langle f, f \rangle_{\mathcal{K}})$$
$$\text{s.t.}\ I(f, \mathcal{D}) \leq 0$$
$$E(f, \mathcal{D}) = 0.$$

This gives a general recipe for turning an RKHS problem into an RKKS problem. While Examples 11 and 12 follow this procedure, it appears to be an open question whether, starting from an arbitrary RKHS problem, the solutions to the corresponding RKKS problem behave as one would hope from solutions to the RKHS problem (eg. provide a good regression/classification rule).

Turning an RKHS problem into an RKKS problem achieves another goal: suppose that, as in Calinon (2020); Jousse et al. (2021), we use a non-PD kernel, which has a PD Gram matrix when restricted to a given data set, and consider a representer solution, as if the kernel was PD on the whole space. Through PD kernel theory, this has no theoretical backing: the kernel has no underlying RKHS over the whole space, therefore one cannot write an RKHS minimization problem that the representer solution solves. However, assuming that the kernel has a PD decomposition and that the appropriate regularity conditions on L, I, E are satisfied, *the representer solution is actually a critical point of the corresponding RKKS problem!*

The vast majority of optimization research has focused on minimization or maximization and rarely on stabilization. Stabilization problems are nonetheless common in fields such as Lagrangian and Hamiltonian mechanics where the least action principles (Abraham and Marsden, 1978; Marsden and Ratiu, 1999) that are at the core of the formulation of various physical theories adopt a stabilization rather than a minimization form. It is in contexts of this type, which are rarely Euclidean, where the RKKS framework could be particularly appropriate.
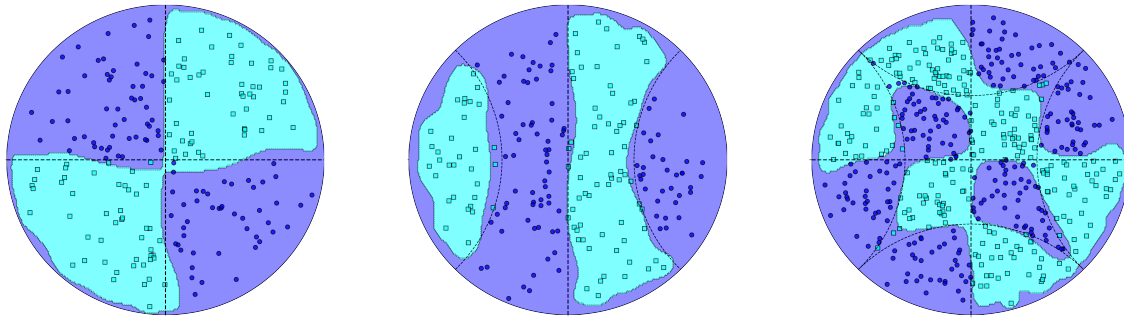
Figure 1: The Krein SVM algorithm (Loosli et al., 2016) applied on the hyperbolic plane $\mathbb{H}^2$, with the geodesic Gaussian kernel $k = \exp(-\lambda d(\cdot, \cdot)^2)$. The data is sampled from a Riemannian Gaussian distribution (Said et al., 2018, 2022, 2023) centered at the origin of the Poincare disc and is split into two classes according to geodesic decision boundaries (dotted curves in the figure). The number of sampled data points is 200, 200, and 500, respectively. The results of the classification are displayed in the Poincare disc model of $\mathbb{H}^2$. We will show in Corollary 56 that the Gaussian kernel admits a PD decomposition on $\mathbb{H}^2$, justifying its use in this scenario.

Nevertheless, designing minimization problems in RKKS could allow one to leverage existing optimization and RKHS learning literature more easily. This is possible, by replacing the regularizer $g(\langle \cdot, \cdot \rangle_{\mathcal{K}}) = g(\langle \cdot, \cdot \rangle_+ - \langle \cdot, \cdot \rangle_-)$ in (6) by a regularizer $g_+(\langle \cdot, \cdot \rangle_+) + g_-(\langle \cdot, \cdot \rangle_-)$ for some strictly increasing g_+, g_-. With respect to such a regularizer, one is then able to solve minimization problems. However, this requires a choice of inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ for the RKKS, i.e. a choice of PD decomposition for the kernel. Constructive choices of PD decompositions for kernels are beyond the scope of the present work.

Example 14 Krein kernel ridge regression introduced in Oglic and Gaertner (2018), is an alternative generalization of kernel ridge regression to Krein spaces (c.f. Example 11). It refers to the learning problem

$$\underset{f \in \mathcal{K}}{\text{minimize}} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + c_+ \langle f, f \rangle_+ + c_- \langle f, f \rangle_-$$

$$\text{s.t.} \sum_{i=1}^{N} \frac{1}{N} \left(f(x_i) - \sum_{j=1}^{N} \frac{1}{N} f(x_j) \right)^2 = r^2$$

for some $r > 0$.

We conclude this section by noting that it is unclear at present whether other PD kernel non-representer-based approaches, such as Gaussian processes and maximum mean discrepancy, have analogues in the PD decomposable context. This appears to be an interesting avenue for future research.

11

2.3 Positive decompositions

Without a PD decomposition, we do not have a representer theorem. In that case, it is not so much that the proof of the representer theorem fails; rather, we cannot state it in the first place. Indeed, the kernel does not have a reproducing property as in Theorem 9, so there is no obvious concept of a space of functions we can stabilize over. Therefore, a solution found, for instance, in Example 12, may not come with a guarantee to provide a good classification rule for unseen data. This gives rise to the following fundamental question:

Problem 15 When does a kernel k admit a PD decomposition?

Remark 16 Problem 15 was first studied by Schwartz (1964). His setting was more general than ours, and only abstract conditions for PD decomposability were obtained. For instance, it is stated in that reference that k has a PD decomposition if and only if there is a PD kernel k_+ such that $k_+ - k$ is PD (see Schwartz (1964, Proposition 38 & Proposition 23)). However, if one wants to recognize PD decomposable kernels in practice, this condition does not appear to be any easier to work with than Definition 7 itself.

Example 17 If $k_1, \ldots, k_n$ are PD kernels and $a_1, \ldots, a_n \in \mathbb{R}$, then the kernel

$$\sum_{i=1}^{n} a_i k_i$$

has the PD decomposition

$$\sum_{i:a_i \geq 0} a_i k_i - \sum_{i:a_i < 0} (-a_i) k_i.$$

In fact, observe that the space of kernels that admit a PD decomposition is exactly the real span of the PD kernels.

Example 18 The pointwise limit of PD kernels is PD. It follows that the pointwise limit of PD decomposable kernels is PD decomposable.

Remark 19 PD kernels form a convex cone that is closed under pointwise convergence. Now, Example 17 and Example 18 show that PD decomposable kernels are the real linear span of these and thus form a real vector space closed under pointwise convergence.

Example 20 Similarly, since the pointwise product of PD kernels is PD (Berg et al., 1984, Chapter 3 Theorem 1.12), so it follows that the pointwise product of PD decomposable kernels is PD decomposable.

Example 21 Suppose X is finite. Then a kernel k can be viewed as a Hermitian matrix K. By Sylvester's law of inertia, we can write

$$K = A^{\dagger} E A$$

for some invertible real matrix A, where $A^\dagger$ denotes the conjugate transpose of A, and E is of the form

$$E = \begin{bmatrix} I_{N_+} & & \\ & -I_{N_-} & \\ & & 0_M \end{bmatrix}$$

for some $N_+, N_-, M \in \mathbb{Z}_{\geq 0}$, where I_{N+}, I_{N-} denote the identity matrices with the corresponding sizes, and 0_M is the zero matrix of size $M \times M$. Then $E = E_+ - E_-$ where

$$E_+ = \begin{bmatrix} 0_{N_+} & & \\ & -I_{N_-} & \\ & & 0_M \end{bmatrix}, \; E_- = \begin{bmatrix} I_{N_+} & & \\ & 0_{N_-} & \\ & & 0_M \end{bmatrix}.$$

This implies that

$$K = A^\dagger E_+ A - A^\dagger E_- A,$$

and hence k has a PD decomposition. However, when X is infinite, one should not expect all kernels to admit a PD decomposition.

Example 22 We present an example of a finite-dimensional RKKS associated with a non-finite X, namely $X = \mathbb{H}^n$, the n-dimensional hyperbolic space. To motivate our construction, we recall that when $X = \mathbb{R}^n$, the standard inner product kernel $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is PD and gives rise to the RKHS $\mathcal{H} = \{\langle x, \cdot \rangle : x \in \mathbb{R}^n\}$, which consists of all linear maps $\mathbb{R}^n \to \mathbb{R}$. So for $f \in \mathcal{H}$, $b \in \mathbb{R}$, the set $\{x \in \mathbb{R}^n : f(x) = b\}$ is a hyperplane in $\mathbb{R}^n$.

Taking $X = \mathbb{H}^n$, it is therefore natural to ask if there is a kernel that gives rise to an RKHS of functions f such that for any $b \in \mathbb{R}$, the sets $\{x \in \mathbb{H}^n : f(x) = b\}$ are geodesic hyperplanes. This has practical relevance, as it would allow us to perform large-margin classification with geodesic decision boundaries, as in Cho et al. (2019). This is done naturally with a PD decomposable kernel giving rise to an RKKS. First consider the Minkowski inner product on $\mathbb{R}^{n+1}$:

$$(x, y) = x_0 y_0 - x_1 y_1 - \cdots - x_n y_n$$

for all $x = (x_0, \ldots, x_n), y = (y_0, \ldots, y_n) \in \mathbb{R}^{n+1}$. Then, we can view $\mathbb{H}^n$ as a hypersurface in this Minkowski space:

$$\mathbb{H}^n = \{x \in \mathbb{R}^{n+1} : (x, x) = 1, x_0 > 0\}.$$

This is the hyperboloid model of $\mathbb{H}^n$. The geodesic hyperplanes in $\mathbb{H}^n$ are given by

$$\{x \in \mathbb{H}^n : (y, x) = b\}$$

for some $y \in \mathbb{R}^{n+1}$, where the above is non-empty if and only if $(y, y) < 0$. These are the intersection of planes through the origin with the hyperboloid model of $\mathbb{H}^n$ (see Figure 2). So the kernel $k : \mathbb{H}^n \times \mathbb{H}^n \to \mathbb{R}$ of interest can be taken to be $k = (\cdot, \cdot)$. It has a PD decomposition by definition of the Minkowski inner product, and the associated real RKKS is the Minkowski space $\mathbb{R}^{n+1}$ equipped with $(\cdot, \cdot)$.
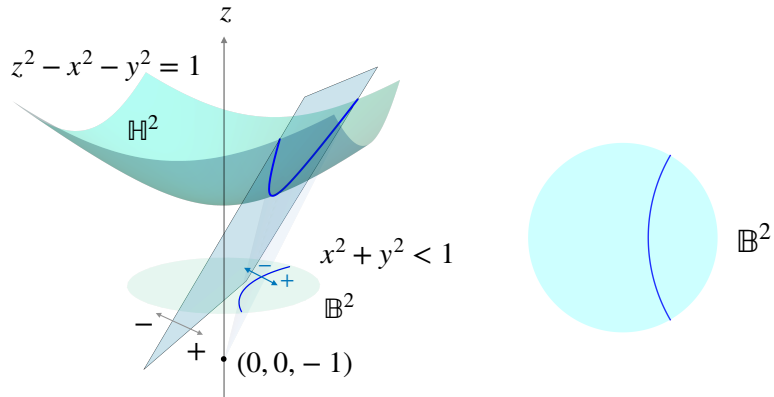
Figure 2: A geodesic in the hyperboloid model of the hyperbolic plane $\mathbb{H}^2$, obtained as the intersection of a plane with the hyperboloid. The corresponding geodesic on the Poincare disc model $\mathbb{B}^2$ of the hyperbolic plane is obtained by stereographic projection from the point $(0, 0, -1)$.

Example 23 We now present a kernel that does not admit a PD decomposition. Let B be a real reflexive Banach space whose norm does not arise from an inner product (for instance, an L^p space with $1 < p < \infty$, $p \neq 2$) and B^* be its dual. Let $X = B \times B^*$ and the kernel $k : X \times X \to \mathbb{R}$ be given by

$$k(x, \varphi; y, \psi) := \varphi(y) + \psi(x).$$

Then k has no PD decomposition. This example was first presented in Schwartz (1964, Page 243). A self-contained proof can be found in Alpay (1991, Theorem 2.2).

3 Group Actions and Invariant Kernels

To have at our disposal additional structure to address Problem 15, we shall restrict our attention to kernels that possess certain symmetries, which we shall refer to as *invariant kernels*. In the rest of the paper, G will be a locally compact (always assumed Hausdorff) topological group, acting on the left on the set X.

Definition 24 *A kernel $k : X \times X \to \mathbb{C}$ is called invariant with respect to the action of G if*

$$k(g \cdot x, g \cdot y) = k(x, y)$$

for all $g \in G$ and all $x, y \in X$, $\cdot$ denotes a left action of the elements of G on the elements of X.

Remark 25 For notational convenience, we will, at times, drop the $\cdot$ all together and write $g \cdot x = gx$.

PD invariant kernels were first studied by Yaglom (1961), and more recently by Azangulov et al. (2024a,b) in the context of Gaussian processes, and by Da Costa et al. (2023b).

Example 26 Taking $X = G = \mathbb{R}^n$ equipped with addition and acting on itself, an invariant kernel on $\mathbb{R}^n$ is a *translation-invariant* or *stationary* kernel:

$$k(x + g, y + g) = k(x, y)$$

for all $x, y, g \in \mathbb{R}^n$. Standard invariant theory arguments show that translation-invariant kernels are exactly those that depend only on the difference $x - y$, i.e. kernels k on $\mathbb{R}^n$ for which there is a function f on $\mathbb{R}^n$ such that

$$k(x, y) = f(x - y)$$

for all $x, y \in \mathbb{R}^n$.

In this paper, we shall study a generalization of this example, namely when the action of G is transitive on X. In that case, if we pick a distinguished element o of X, the orbit-stabilizer theorem gives us the bijection of sets

$$G/H \to X$$
$$gH \mapsto g \cdot o$$

where $H = \text{Stab}(o)$. In what follows, we will write X as G/H, where H is a subgroup of G. This notation has the advantage of specifying both the group G acting on X and the action (the canonical left action of G on G/H), so from here on, we will call an invariant kernel on G/H one that is invariant with respect to that action. We will further make the assumption that H is closed, providing us with a locally compact Hausdorff quotient topology for $X = G/H$, and making the action of G on X proper (see Folland (2015, Section 2.6) for the theory of such homogeneous spaces).

Invariant kernels under transitive actions are particularly interesting for us because the symmetries they respect allow us to view them as functions of a single variable instead of two, as in Example 26. To see this, let us first give the analogues to Definitions 1 and 7 for functions defined on the relevant spaces.

Definition 27 *Write $H \backslash G / H := \{HgH : g \in G\}$ for the double coset space. A complex-valued function $f : H \backslash G / H \to \mathbb{C}$ is said to be positive definite (PD) if for all $N \in \mathbb{N}$ and $g_1, \ldots, g_N \in G$, the matrix*

$$\left(f(Hg_i^{-1}g_j H) \right)_{i,j}$$

is positive semidefinite. f is said to have a positive (PD) decomposition if it can be written as

$$f = f_+ - f_-$$

where f_+ and f_- are PD functions on $H \backslash G / H$.

In particular, when $H = \{e\}$, $f : G \to \mathbb{C}$ is PD if the matrices $\left(f(g_i^{-1}g_j) \right)_{i,j}$ are positive semidefinite for all N and all $g_i, g_j \in G$. Functions on $H \backslash G / H$ should be considered as H-invariant functions on G/H. The relevance of double coset spaces in the kernel context will become clear later in Proposition 32.
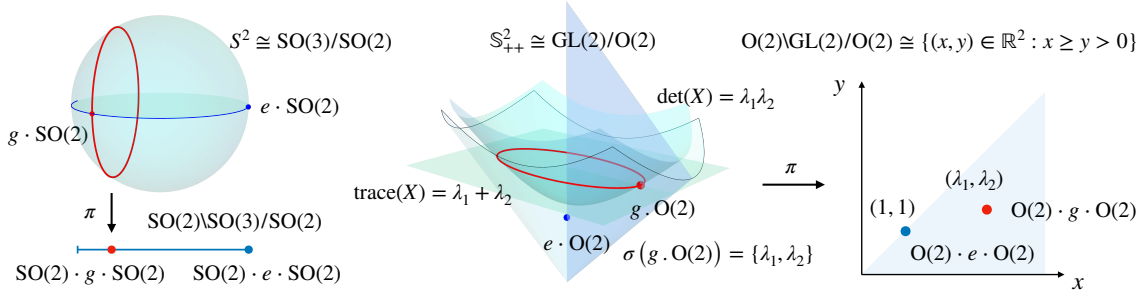
Figure 3: A visualisation of the projections to the double coset spaces π : $S^2 \cong SO(3)/SO(2) \to SO(2)\backslash SO(3)/SO(2)$ and π : $\mathbb{S}^2_{++} \cong GL(2)/O(2) \to O(2)\backslash GL(2)/O(2)$. In each case, we have in red an orbit HgH. In the case of S^2, it is obtained by rotating $gSO(2)$ around a horizontal axis. In the case of $\mathbb{S}^2_{++}$, it is the set of matrices with eigenvalues equal to the ones of $gO(2)$, obtained as the intersection of the constant-trace plane and the constant-determinant surface. In each case, we have in dark blue a set transversal to such orbits, and homeomorphic to $H\backslash G/H$.

Example 28 The n-dimensional sphere S^n may be viewed as the coset space

$$S^n \cong SO(n+1)/SO(n).$$

Moreover, viewed as a submanifold $S^n \subset \mathbb{R}^{n+1}$, $SO(n+1)$ acts transitively on S^n by rotations. Picking $o = (1, 0, \ldots, 0) \in S^n$, the stabilizer of o is the subgroup of $SO(n+1)$ leaving the first coordinate untouched, which is isomorphic to $SO(n)$. Now the double coset space $SO(n)\backslash SO(n+1)/SO(n)$ is obtained by quotienting out each orbit (see Figure 3), so we see

$$SO(n)\backslash SO(n+1)/SO(n) \cong [-\pi, \pi].$$

Functions on $SO(n)\backslash SO(n+1)/SO(n)$ are functions of the geodesic distance on S^n, or equivalently of the inner product of $\mathbb{R}^{n+1}$ restricted to S^n.

Example 29 The space of $n \times n$ real symmetric positive definite matrices $\mathbb{S}^n_{++}$ may be viewed as the coset space

$$\mathbb{S}^n_{++} \cong GL(n)/O(n).$$

Indeed, $GL(n)$ acts transitively on $\mathbb{S}^n_{++}$ by $G \cdot X := GXG^\top$ for $G \in GL(n)$ and $X \in \mathbb{S}^n_{++}$. The stabilizer of the the identity I consists of the matrices G such that $GG^\top = I$, i.e. $\mathrm{Stab}(I) = O(n)$. Now, for $X \in \mathbb{S}^n_{++}$, $O(n)XO(n) \in O(n)\backslash GL(n)/O(n)$ consists of $\{HXH^\top : H \in O(n)\}$, which is fully determined by the n eigenvalues of X. So

$$O(n)\backslash GL(n)/O(n) \cong \{(\lambda_1, \ldots, \lambda_n) : \lambda_i > 0 \,\forall\, i\}/\sim$$

where $(\lambda_1, \ldots, \lambda_n) \sim (\lambda_1', \ldots, \lambda_n')$ if and only if $\{\lambda_1, \ldots, \lambda_n\} = \{\lambda_1', \ldots, \lambda_n'\}$ (see Figure 3). Functions on $O(n)\backslash GL(n)/O(n)$ are functions of the eigenvalues on $\mathbb{S}^n_{++}$.

16

Let us generalize the concept of Hermitian functions to complex-valued functions on double coset spaces:

Definition 30 $f : H\backslash G/H \to \mathbb{C}$ *is called Hermitian if*

$$f(Hg^{-1}H) = \overline{f(HgH)}$$

for all $g \in G$.

Since G is a topological group, G/H and $H\backslash G/H$ are equipped with natural quotient topologies. In the rest of the paper, we shall make the following continuity assumption: we will consider continuous kernels k, and ask whether they admit PD decompositions into continuous PD kernels $k = k_+ - k_-$. We call such a decomposition a *continuous PD decomposition*.

Remark 31 The continuity assumption is non-trivial: see Schaback and Wendland (2003, Section 3) for a construction of an everywhere discontinuous PD kernel. Note, however, that at least for functions on groups, a continuous function on G has a continuous PD decomposition if and only if it has a PD decomposition (Kaniuth and Lau, 2018, Corollary 2.2.2).

The reason for considering functions on the *double* coset space $H\backslash G/H$ as opposed to the regular coset space G/H is made apparent in the following important result.

Proposition 32 *There is a bijection between the continuous Hermitian functions on $H\backslash G/H$ and the continuous invariant kernels on G/H. This bijection restricts to:*

1. *A bijection between the continuous Hermitian PD functions and the continuous PD invariant kernels.*

2. *An injection between the continuous Hermitian PD functions with continuous PD decompositions and the continuous kernels with continuous PD decompositions.*

Proof Given a continuous invariant kernel $k : G/H \times G/H \to \mathbb{C}$, define the function $f : H\backslash G/H \to \mathbb{C}$ by

$$f(HgH) := k(gH, eH)$$

for all $g \in G$, where e is the identity of G. Note that

$$f(Hg^{-1}H) = k(g^{-1}H, e) = k(g^{-1}H, eH) = k(eH, gH) = \overline{k(gH, eH)} = \overline{f(HgH)}$$

i.e. if k is Hermitian then f is necessarily Hermitian. Finally, the continuity of f follows from the continuity of k. Conversely, given a Hermitian function $f : H\backslash G/H \to \mathbb{C}$, define the invariant kernel $k : G/H \times G/H \to \mathbb{C}$ by

$$k(gH, g'H) := f(Hg^{-1}g'H)$$

for all $g, g' \in G$ where the continuity of and the Hermitian character of f imply the same properties for k.

These correspondences are inverses of each other and thus provide a bijection. Moreover, we can see that k is PD if and only if f is PD. Also, this correspondence is linear, which implies that if f admits a continuous PD decomposition, so does k. ∎

Remark 33 In the case $H = \{e\}$, the correspondence above is simply the bijection between Hermitian functions on G and invariant kernels on G. Example 26 is an instance of this case.

Remark 34 Note that in the second part of Proposition 32, we have only proved the injectivity of the correspondence between continuous Hermitian PD functions with continuous PD decompositions and the continuous kernels with continuous PD decompositions. The difficulty in proving surjectivity is that k may not admit a continuous PD decomposition $k = k_+ - k_-$ with k_+ and k_- invariant. In Appendix B we provide a partial solution by showing that the correspondence in part 2 is indeed surjective when G is compact. In any case, the injectivity in this part will be enough for us to provide sufficient conditions for a kernel k to have a continuous PD decomposition.

Being able to handle invariant kernels as functions will allow us to leverage tools from harmonic analysis and representation theory to address Problem 15.

In the sequel, we shall use the following spaces in relation with the locally compact group G and the closed subgroup H:

$$
\begin{aligned}
B_+(G) &:= \{\text{continuous PD functions } G \to \mathbb{C}\} \\
B_\mathbb{R}(G) &:= \operatorname{span}_\mathbb{R}(B_+(G)) \\
&= \{\text{continuous functions with continuous PD decompositions } G \to \mathbb{C}\} \\
B_\mathbb{C}(G) &:= \operatorname{span}_\mathbb{C}(B_+(G))
\end{aligned}
$$

and, more generally,

$$
\begin{aligned}
B_+(H\backslash G/H) &:= \{\text{PD functions } H\backslash G/H \to \mathbb{C}\} \\
B_\mathbb{R}(H\backslash G/H) &:= \operatorname{span}_\mathbb{R}(B_+(H\backslash G/H)) \\
&= \{\text{continuous functions with continuous PD decompositions } H\backslash G/H \to \mathbb{C}\} \\
B_\mathbb{C}(H\backslash G/H) &:= \operatorname{span}_\mathbb{C}(B_+(H\backslash G/H)).
\end{aligned}
$$

Remark 35 While $B_\mathbb{R}(G)$ and $B_\mathbb{R}(H\backslash G/H)$ are defined as the real span of $B_+(G)$ and $B_+(H\backslash G/H)$, note that the latter contain complex-valued functions that are not real-valued, hence so do the former.

The space $B_\mathbb{C}(G)$ is called the *Fourier-Stieltjes algebra* of G in the literature and has been studied thoroughly; see, for example, Kaniuth and Lau (2018). We will present and extend some of these results to $B_\mathbb{C}(H\backslash G/H)$.

Motivated by Proposition 32, our goal is to characterize $B_\mathbb{R}(H\backslash G/H)$ and consider Hermitian elements in this space. In fact, since we are interested in real-valued kernels and hence in real-valued functions, the following proposition shows that it suffices to characterize $B_\mathbb{C}(H\backslash G/H)$ and consider real-valued even ($f(HgH) = f(Hg^{-1}H)$ for all g) elements in this space.

Proposition 36 *If* $f \in B_\mathbb{C}(G)$ *or* $f \in B_\mathbb{C}(H\backslash G/H)$ *is real-valued, then* $f \in B_\mathbb{R}(G)$ *or* $f \in B_\mathbb{R}(H\backslash G/H)$, *respectively.*

Proof Observe that if $f \in B_+(G)$, then $\operatorname{Re} f = \frac{f + \bar{f}}{2} \in B_+(G)$, and $\operatorname{Im} f = \frac{f - \bar{f}}{2} \in B_\mathbb{R}(G)$. If $f \in B_\mathbb{C}(G)$ we can write

$$f = f_+ - f_- + if_{+i} - if_{-i}$$

with $f_+, f_-, f_{+i}, f_{-i} \in B_+(G)$. If f is real-valued,

$$f = \operatorname{Re} f = \operatorname{Re} f_+ - \operatorname{Re} f_- + \operatorname{Re}(if_{+i}) - \operatorname{Re}(if_{-i}) = \operatorname{Re} f_+ - \operatorname{Re} f_- - \operatorname{Im} f_{+i} + \operatorname{Im} f_{-i}$$

so $f \in B_\mathbb{R}(G)$. The proof is identical for functions in $B_\mathbb{C}(H \backslash G / H)$. ∎

4 Positive Decompositions for Invariant Kernels: Commutative Case

In this section, we focus on invariant kernels that are invariant with respect to the action of an Abelian locally compact group G. The commutativity hypothesis will simplify the theory considerably.

First, observe that in this case, any closed subgroup H of G is automatically a normal subgroup, so $X = G/H$ is itself a group. Equipped with the quotient topology, G/H is actually a locally compact group (Folland, 2015, Proposition 2.2). Therefore without loss of generality, we can set $X = G$.

In view of the results in the previous section, we shall focus in this case on $B_\mathbb{C}(G)$, which is a well-studied object in standard commutative abstract harmonic analysis. For the sake of completeness, we recall here a few facts in relation to this theory that are needed in our developments. We refer to Folland (2015) for a more detailed exposition and proofs. Folland (2015, Chapter 4) is particularly relevant to this section.

G has a *dual group*, $\widehat{G}$, which is defined as the group of continuous irreducible unitary representations of G (see Definition 59 in Appendix C where the concept of representations becomes much more central). G being Abelian implies that irreducible unitary representations of G are one-dimensional, so $\widehat{G}$ consists of the continuous group homomorphisms $\xi : G \to S^1 \subset \mathbb{C}$, i.e. $\widehat{G} = \operatorname{Hom}(G, S^1)$. Equipped with the open-compact topology, $\widehat{G}$ is itself a locally compact group. Moreover, we have a natural topological group isomorphism $\widehat{\widehat{G}} \cong G$. This is the *Pontryagin Duality Theorem*.

Now define

$$M_+(G) := \{\text{finite positive Radon measures on } G\}$$
$$M_\mathbb{R}(G) := \operatorname{span}_\mathbb{R}(M_+(G))$$
$$= \{\text{finite signed Radon measures on } G\}$$
$$M_\mathbb{C}(G) := \operatorname{span}_\mathbb{C}(M_+(G))$$
$$= \{\text{finite complex Radon measures on } G\}$$

and $M_+(\widehat{G})$, $M_\mathbb{R}(\widehat{G})$, $M_\mathbb{C}(\widehat{G})$ analogously. Importantly for us, we can define the *Fourier transform* $\mathscr{F} : M_\mathbb{C}(G) \to C_\mathbb{C}(\widehat{G})$ by

$$\mathscr{F}\mu(\xi) = \int_G \overline{\xi(g)} d\mu(g)$$

19

for all $\xi \in \widehat{G}$. Here $C_{\mathbb{C}}(\widehat{G})$ is the space of continuous complex-valued functions on $\widehat{G}$. Similarly, we can define the inverse Fourier transform $\mathscr{F}^{-1} : M_{\mathbb{C}}(\widehat{G}) \to C_{\mathbb{C}}(G)$ by

$$\mathscr{F}^{-1}\nu(g) = \int_G \xi(g) d\nu(\xi)$$

for all $g \in G$. $\mathscr{F}$ and $\mathscr{F}^{-1}$ are injective linear maps.

Fourier transforms are often thought of as acting on functions. Our description of Fourier transforms encompasses this: note that G comes equipped with a Haar measure μ. This is a translation-invariant Radon measure on G, which is unique up to multiplication by a positive constant. Then, for f a μ-integrable function on G, $f\mu \in M_{\mathbb{C}}(G)$, so we can define the Fourier transform of f as $\mathscr{F}(f\mu)$.

Remark 37 $\mathscr{F}$ and $\mathscr{F}^{-1}$ will only be inverse to each other with the right choice of Haar measures on G and $\widehat{G}$. Indeed, as described above, the actions of the operators $\mathscr{F}$, $\mathscr{F}^{-1}$ on $L^1(G)$, $L^1(\widehat{G})$ depend on the choice of Haar measures. See Folland (2015, Theorem 4.22).

Example 38 When $G = \mathbb{R}^n$, we have $\widehat{G} \cong \mathbb{R}^n$, and the Fourier transform can be written as

$$\mathscr{F}\mu(\xi) = \int_{\mathbb{R}^n} e^{-i\langle \xi, x\rangle} d\mu(x)$$

for $\mu \in M_{\mathbb{C}}(\mathbb{R}^n)$ and $\xi \in \mathbb{R}^n$. The inverse Fourier transform can be written as

$$\mathscr{F}^{-1}\nu(x) = \int_{\mathbb{R}^n} e^{i\langle \xi, x\rangle} d\nu(\xi)$$

for $\nu \in M_{\mathbb{C}}(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$.

Example 39 When $G = S^1$, we have $\widehat{G} \cong \mathbb{Z}$, and the Fourier transform corresponds to the discrete Fourier transform, and can be written as

$$\mathscr{F}\mu(k) = \int_{S^1} e^{-ik[x]} d\mu(x)$$

for $\mu \in M_{\mathbb{C}}(S^1)$ and $k \in \mathbb{Z}$, where $[x] \in [-\pi, \pi)$ is the argument of x. The inverse Fourier transform can be written as

$$\mathscr{F}\nu(x) = \int_{S^1} e^{ik[x]} d\nu(x)$$

for $\nu \in M_{\mathbb{C}}(S^1)$ and $x \in S^1$.

We are now able to state the key theorem of this section (Folland, 2015, Theorem 4.19).

Theorem 40 (Bochner's Theorem) $f \in B_+(G)$ *is PD if and only if there is* $\nu \in M_+(\widehat{G})$ *such that* $f = \mathscr{F}^{-1}\nu$. *Moreover, such* ν *is unique.*

In other words, $\mathscr{F}^{-1}$ is a linear isomorphism between $M_+(\widehat{G})$ and $B_+(G)$ (which are convex cones, not vector spaces). Therefore, tensoring by $\mathbb{R}$ and then by $\mathbb{C}$ in Bochner's Theorem,

20

we also get that $\mathscr{F}^{-1}$ is a linear isomorphism between $M_{\mathbb{R}}(\widehat{G})$, $M_{\mathbb{C}}(\widehat{G})$ and $B_{\mathbb{R}}(G)$, $B_{\mathbb{C}}(G)$, respectively. In other words, we have the commutative diagram

$$
\begin{array}{ccc}
M_+(\widehat{G}) & \xleftrightarrow{\ \sim\ } & B_+(G) \\
\downarrow & & \downarrow \\
M_{\mathbb{R}}(\widehat{G}) & \xleftrightarrow{\ \sim\ } & B_{\mathbb{R}}(G) \\
\downarrow & & \downarrow \\
M_{\mathbb{C}}(\widehat{G}) & \xleftrightarrow{\ \sim\ } & B_{\mathbb{C}}(G) \\
& {}_{\mathscr{F}^{-1}}\searrow & \downarrow \\
& & C_{\mathbb{C}}(G)
\end{array}
$$

where $\sim$ denote linear isomorphisms and $\hookrightarrow$ denote the inclusions. In other words, PD decomposable functions on G are the functions that are the inverse Fourier transforms of finite signed measures on $\widehat{G}$. This description of $B_{\mathbb{R}}(G)$ and $B_{\mathbb{C}}(G)$ is in some sense the best we have, as being the inverse Fourier transform of a finite measure does not appear to be a reducible property. However, what we can do is provide *sufficient* conditions for a function to be in $B_{\mathbb{R}}(G)$ and $B_{\mathbb{C}}(G)$.

First, let us state two important results.

Theorem 41 *The map $\mathscr{F}^{-1} : M(\widehat{G}) \to C_{\mathbb{C}}(G)$ is surjective if and only if G is finite. So, $B_{\mathbb{C}}(G) = C_{\mathbb{C}}(G)$ if and only if G is finite.*

See Graham and McGehee (1979). We showed the *if* direction in Example 21, and this proposition provides a converse. It also provides a glimpse at the fact that characterizing $B_{\mathbb{C}}(G)$ is a deep problem.

We have, however, the following positive result:

Theorem 42 $B_{\mathbb{C}}(G) \cap C_{\mathbb{C}}^c(G)$ *is dense in* $C_{\mathbb{C}}^c(G)$*, the space of compactly supported complex-valued continuous functions, with the uniform norm. As a consequence, $B_{\mathbb{C}}(G) \cap L^p(G)$ is dense in $L^p(G)$ for all $1 \le p < \infty$.*

See Folland (2015, Proposition 3.33). Theorem 48 below is the general version of this result for the non-commutative case.

For the commutative case, there are, in practice, two examples of interest: $\mathbb{R}$ and S^1. This is reflected by the fact that any Abelian Lie group G is a product of these. In Example 38 and Example 39, we have already given expressions for the Fourier transform in these cases. Let us study the implications in each case.

Example 43 $G = \mathbb{R}^n$, $\widehat{G} = \mathbb{R}^n$. In this case, it is well-known that $\mathscr{F}$ is an automorphism of the Schwartz space $\mathscr{S}(\mathbb{R}^n)$. So, in particular, $\mathscr{S}(\mathbb{R}^n) \subset B_{\mathbb{C}}(\mathbb{R}^n)$, where we recall that

$$
\mathscr{S}(\mathbb{R}^n) := \left\{ f \in C_{\mathbb{C}}^\infty(\mathbb{R}^n) : \sup_{x \in \mathbb{R}^n} \left| x^\alpha \left(D^\beta f \right)(x) \right| < \infty \ \forall \, \alpha, \beta \in \mathbb{N}^n \right\}.
$$

In other words, $\mathscr{S}(\mathbb{R}^n)$ is the space of complex-valued smooth functions on $\mathbb{R}^n$ whose values and derivatives decay faster at infinity than any reciprocal of a polynomial.

Example 44 We can view S^1 as $G = \mathbb{R}/\mathbb{Z}$, and we have $\widehat{G} = \mathbb{Z}$. As described in Example 39, the Fourier transform in this case is the discrete Fourier transform. We have that $M_{\mathbb{C}}(\mathbb{Z})$ is $l^1_{\mathbb{C}}(\mathbb{Z})$, and $B_{\mathbb{C}}(\mathbb{R}/\mathbb{Z})$ is the space of absolutely continuous Fourier series. This is called the Wiener algebra, and we have that the Hölder continuous functions $C^\alpha_{\mathbb{C}}(\mathbb{R}/\mathbb{Z}) \subset B_{\mathbb{C}}(\mathbb{R}/\mathbb{Z})$ for $\alpha > 1/2$. This is Bernstein's Theorem, see for example Katznelson (2004, Theorem 6.3). Recall that for $\alpha < 1$,

$$C^\alpha_{\mathbb{C}}(\mathbb{R}/\mathbb{Z}) := \left\{ f \in C_{\mathbb{C}}(\mathbb{R}/\mathbb{Z}) : \sup_{x,y \in \mathbb{R}/\mathbb{Z}, \, x \neq y} \frac{|f(x) - f(y)|}{|x - y|^\alpha} < \infty \right\},$$

where the absolute value $|x - y|$ in the denominator is taken modulo 1.

Example 45 The Gaussian kernel on the circle S^1

$$k(x,y) = \exp(-\lambda |x - y|^2)$$

for all $x, y \in \mathbb{R}/\mathbb{Z} \cong S^1$, where $\lambda > 0$ and the absolute value is taken modulo 1, is not PD for any $\lambda > 0$ (see Da Costa et al. (2023a)). However, note that it is invariant, and under Proposition 32 corresponds to the Gaussian function $f(x) = \exp(-\lambda |x|^2)$ for all $x \in \mathbb{R}/\mathbb{Z} \cong S^1$, where the absolute value is taken modulo 1. f is Lipschitz, so by Example 44, $f \in B_{\mathbb{C}}(S^1)$, and since it is real-valued, $f \in B_{\mathbb{R}}(S^1)$ by Proposition 36. So k, despite not being PD, has a PD decomposition.

From Example 45, we obtain the following result.

Proposition 46 *The Gaussian kernel $k = \exp(-\lambda d(\cdot,\cdot)^2)$ has a PD decomposition on any Abelian Lie group (with appropriate choice of the metric d) for any $\lambda > 0$.*

Proof Any Abelian Lie group G is a product of copies of $\mathbb{R}$ and S^1, say $G \cong \mathbb{R}^n \times (S^1)^m$. Taking the product Riemannian metric on the right-hand side of this induced by the standard Riemannian metrics on $\mathbb{R}$ and S^1, we can write the Gaussian kernel k on G as

$$k(x,y) = \exp(-\lambda d(x,y)^2) = \prod_{i=1}^{n+m} \exp(-\lambda |x_i - y_i|^2)$$

for all $x, y \in G \cong \mathbb{R}^n \times (S^1)^m$, where the last m absolute values are taken modulo 1. By positive definiteness of the Gaussian kernel on $\mathbb{R}$ and PD decomposability of the Gaussian kernel on S^1 (Example 45), k is a product of PD decomposable kernels, so has a PD decomposition by Example 20. ∎

We conclude this section with the following example.

Example 47 Suppose $\tilde{f} \in C^\alpha_{\mathbb{C}}(\mathbb{R})$ for some $\alpha > 1/2$ and that $\tilde{f}$ is periodic, with period T, say. Then, if $\pi : \mathbb{R} \to \mathbb{R}/T\mathbb{Z}$ is the projection, there is a function $f \in C^\alpha_{\mathbb{C}}(\mathbb{R}/T\mathbb{Z})$ such that $f \circ \pi = \tilde{f}$. Then, $f \in B_{\mathbb{R}}(\mathbb{R}/T\mathbb{Z})$, assuming for simplicity that $\tilde{f}$ is real-valued. Writing its PD decomposition $f = f_+ - f_-$, we have that $\tilde{f} = f_+ \circ \pi - f_- \circ \pi$ is a PD decomposition, so that $\tilde{f} \in B_{\mathbb{R}}(\mathbb{R})$.

PD decompositions behave well with respect to quotients. We will prove a general version of this phenomenon, as well as its converse, in Section 5.

5 Positive Decompositions for Invariant Kernels: Non-Commutative Case

Armed with an understanding of the commutative case, we are now in a position to tackle the non-commutative case. Recall that we aim to characterize sufficient conditions for an invariant kernel to admit a PD decomposition, which by Proposition 32 and Proposition 36 comes down to characterizing sufficient conditions for a function to belong to $B_{\mathbb{C}}(H\backslash G/H)$. Here the only assumptions are that G is a locally compact group and H is a closed subgroup of G.

We begin by observing that, just as in the commutative case (Theorem 42), PD decomposable functions on G are plentiful and are, in fact, dense in the space of compactly supported continuous functions:

Theorem 48 $B_{\mathbb{C}}(G) \cap C_{\mathbb{C}}^c(G)$ *is dense in* $C_{\mathbb{C}}^c(G)$, *the space of compactly supported continuous functions, with the uniform norm. As a consequence,* $B_{\mathbb{C}}(G) \cap L^p(G)$ *is dense in* $L^p(G)$ *for all* $1 \leq p < \infty$.

See Kaniuth and Lau (2018, Corollary 2.3.5) for a proof.

To obtain more specific descriptions of $B_{\mathbb{C}}(G)$ and $B_{\mathbb{C}}(H\backslash G/H)$, there are multiple different possible approaches. Firstly, Fourier transforms can be defined on general compact groups G analogously to the circle: they can be described as Fourier series, see Folland (2015, Chapter 5). The non-compact case is more difficult. Additional regularity assumptions must be placed on the group, for example, that it is of type I, see for instance Yaglom (1961). This latter reference provides Bochner-like theorems for the homogeneous spaces of type I groups, which include symmetric spaces. The most general version of Bochner's Theorem is known as the Bochner-Godement Theorem, see Faraut and Harzallah (1974, Theorem 3.1), which describes PD H-invariant functions on G/H, i.e. PD functions on $H\backslash G/H$, with some additional mild regularity assumptions on G/H.

All of these references are quite technical. In this paper, we will take a different route. Describing a general Bochner Theorem for non-Abelian groups will not get us much closer to characterizing $B_{\mathbb{C}}(H\backslash G/H)$. What we are really after are concrete sufficient conditions for a function to belong to $B_{\mathbb{C}}(H\backslash G/H)$ on spaces of interest, for instance, when G is a Lie group and H is a Lie subgroup. Therefore, we will proceed as follows.

As exemplified by Theorem 48, $B_{\mathbb{C}}(G)$ has been more extensively studied than the more general $B_{\mathbb{C}}(H\backslash G/H)$. We will, therefore, start by describing how the latter relates to the former. After this, we will be in a position in Section 5.1 to describe sufficient conditions for a function to belong to $B_{\mathbb{C}}(H\backslash G/H)$ for general classes of Lie groups G and Lie subgroups H.

Suppose that $f \in C_{\mathbb{C}}(H\backslash G/H)$. If $f \in B_{\mathbb{R}}(H\backslash G/H)$ and $\pi : G \to H\backslash G/H$ is the projection, write

$$f = f_+ - f_-$$

with $f_+, f_- \in B_+(H\backslash G/H)$. Then

$$f \circ \pi = f_+ \circ \pi - f_- \circ \pi$$

with $f_+ \circ \pi, f_- \circ \pi \in B_+(H\backslash G/H)$, so $f \circ \pi \in B_{\mathbb{R}}(G)$.

If we can characterize $B_{\mathbb{R}}(G)$ well, then it is the converse we would like to show: that if $f \in C_{\mathbb{C}}(H\backslash G/H)$ and $f \circ \pi \in B_{\mathbb{R}}(G)$, then $f \in B_{\mathbb{R}}(H\backslash G/H)$. This is *not* obvious: from

the fact that f admits a PD decomposition with functions on G and is invariant on double cosets, we would like to show that we can choose a PD decomposition for f for which the terms are themselves invariant on double cosets.

One idea may be, given $f \in C_{\mathbb{C}}(H\backslash G/H)$ with $f \circ \pi \in B_{\mathbb{R}}(G)$ and a PD decomposition $f \circ \pi = \tilde{f}_+ - \tilde{f}_-$, to 'average out' values of $\tilde{f}_+$ and $\tilde{f}_-$ along each double coset, in order to find PD $f_+, f_- \in C_{\mathbb{C}}(H\backslash G/H)$ such that $f = f_+ - f_-$. This argument works if H is compact.

For the general case, the situation is more difficult, as we cannot simply average over the non-compact double cosets. We will only need the case where H is compact in Section 5.1, thus we only prove this case here. Moreover, the case where H is compact is sufficient for Riemannian homogeneous spaces by the Myers-Steenrod Theorem (Gallot et al., 2004, Theorem 2.35). The reader may refer to Appendix C for a full proof of Theorem 49, using the language of representation theory.

In the statement and the proof of Theorem 49, we will view $C_{\mathbb{C}}(H\backslash G/H)$ as a subset of $C_{\mathbb{C}}(G)$, which will simplify notation.

Theorem 49 *We have*

1. $B_{\mathbb{R}}(G) \cap C_{\mathbb{C}}(H\backslash G/H) = B_{\mathbb{R}}(H\backslash G/H)$,

2. $B_{\mathbb{C}}(G) \cap C_{\mathbb{C}}(H\backslash G/H) = B_{\mathbb{C}}(H\backslash G/H)$.

Proof when H is compact $\supseteq$: For part 1, if $f \in B_{\mathbb{R}}(H\backslash G/H)$ then $f \in C_{\mathbb{C}}(H\backslash G/H)$ by definition, and f has a PD decomposition, so $f \in B_{\mathbb{R}}(G)$ (or see the argument at the beginning of this section). Part 2 is similar.

$\subseteq$: For $g \in G$, define L_g and R_g to be the operators acting on functions $f : G \to \mathbb{C}$ by

$$L_g f(a) = f(ga)$$

and

$$R_g f(a) = f(ag)$$

for all $a \in G$. Suppose $f \in B_{\mathbb{R}}(G) \cap C_{\mathbb{C}}(H\backslash G/H)$. Then, by invariance of f on double cosets,

$$f(g) = \frac{1}{\nu(H)^2} \int_H \int_H R_{h'} L_h f(g) d\nu(h) d\nu(h')$$

for $g \in G$, where ν is a translation bi-invariant Haar measure on H, which exists and is finite by compactness of H. So writing the PD decomposition of f as $f = f_+ - f_-$, we get

$$f = \frac{1}{\nu(H)^2} \int_H \int_H R_{h'} L_h f_+ d\nu(h) d\nu(h') - \frac{1}{\nu(H)^2} \int_H \int_H R_{h'} L_h f_- d\nu(h) d\nu(h').$$

Both the first term and minus the second term are PD, and they are invariant on double cosets by bi-invariance of the Haar measures. Continuity of the terms is easy to check, and the proof for $B_{\mathbb{C}}(G)$ and $B_{\mathbb{C}}(H\backslash G/H)$ is similar. ∎

5.1 Geometry and sufficient conditions for positive decomposability

Assuming some geometry on G and H, we are now in a position to describe sufficient conditions for a function to be in $B_{\mathbb{C}}(H\backslash G/H)$. Specifically, we assume in this section that G is a connected Lie group. Furthermore, we assume that G is unimodular, meaning it possesses a Haar measure that is bi-invariant. This includes all Abelian and all compact Lie groups, but also many non-compact Lie groups of interest, such as $GL_{\mathbb{R}}(n)$ and $GL_{\mathbb{C}}(n)$.

Then, we have the following powerful characterization.

Theorem 50 *Fix a left-invariant Riemannian metric on the connected unimodular Lie group G, and let Δ be the Laplace-Beltrami operator with respect to this metric. Suppose $f \in C_{\mathbb{C}}^{\infty}(G)$ is such that*

$$\Delta^k f = \underbrace{\Delta \cdots \cdot \Delta}_{k \text{ times}} f$$

is L^2-integrable with respect to the Haar measure on G for any integer $k \geq 0$. Then $f \in B_{\mathbb{C}}(G)$.

Proof When G is compact, the assumption can be reduced to $f \in C_{\mathbb{C}}^{\infty}(G)$, and the result follows from Sugiura (1971), or from Gaudry and Pini (1986). For non-compact G, the result is a special case of Gaudry et al. (1990, Theorem 1), which proves the inclusion of certain Besov spaces in Fourier algebras. See also the remark after Gaudry et al. (1990, Equation 12), as well as the containment of the Fourier algebra $A_2(G)$ in the Fourier-Stieltjes algebra, for example in Kaniuth and Lau (2018, Proposition 2.3.3). We should note that the description of the Laplacian in Gaudry et al. (1990) as the sum of the squares of vector fields generating the Lie algebra $\mathfrak{g}$ of G is non-standard, but we prove in Appendix D that the Laplace-Beltrami operator arising from a left-invariant metric on a unimodular Lie group can be formulated in these terms. ∎

Remark 51 For simplicity, Theorem 50 relaxes the results from the works referred to in its proof. For instance, by Gaudry and Pini (1986) we can show that when G is compact, $C_{\mathbb{C}}^k(G) \subset B_{\mathbb{C}}(G)$ for $k > n/2$, where $n = \dim G$. This still does not capture the full strength of the result in this reference, which is a generalization of Bernstein's Theorem (see Example 44). On the other hand, when G is non-compact, Gaudry et al. (1990) shows, in particular, that it suffices that $\Delta^k f$ is weakly in L^2 for $k \leq (n^2 + n)/4$ (see also Varopoulos (1988, Equation 2.6)).

Remark 52 Compare this with Example 44. In both cases, for a function f to belong to $B_{\mathbb{C}}(G)$ it suffices that

1. f is smooth enough,

2. f and its derivatives decay fast enough at infinity.

We would now like to generalize Theorem 50 from $B_{\mathbb{C}}(G)$ to $B_{\mathbb{C}}(H\backslash G/H)$ with the help of Theorem 49. For this, we assume H is a compact Lie subgroup of G. Then, G/H is a homogeneous space, which inherits the structure of a manifold from the one on G. Moreover,

observe that if the left-invariant Riemannian metric on G is also right-H-invariant, then it gives rise to a natural Riemannian metric on the homogeneous space G/H, making the projection map $\pi_R : G \to G/H$ into a Riemannian submersion (see Besse (1987, Chapter 9)). The 'R' in π_R stands for 'right', and we will write $\pi_L : G/H \to H\backslash G/H$ for this other projection map, and $\pi = \pi_L \circ \pi_R : G \to H\backslash G/H$.

Remark 53 While G/H inherits the structure of a manifold from the one on G, $H\backslash G/H$ does not. See Example 28 and Example 29.

Theorem 54 *Suppose that the connected unimodular Lie group G is equipped with a left-invariant right-H-invariant Riemannian metric, where H is a compact Lie subgroup of G. Equip G/H with the Riemannian metric induced by the one on G. Suppose $\tilde{f} \in C_{\mathbb{C}}^{\infty}(G/H)$ is such that there is $f \in C_{\mathbb{C}}(H\backslash G/H)$ with $\tilde{f} = f \circ \pi_L$, i.e. $\tilde{f}$ is H-invariant. If $\Delta^k \tilde{f}$ is L^2-integrable with respect to the Haar measure on G/H for any integer $k \geq 0$, then $f \in B_{\mathbb{C}}(H\backslash G/H)$.*

Proof $f \in C_{\mathbb{C}}(H\backslash G/H)$, so by Theorem 49, to show that $f \in B_{\mathbb{C}}(H\backslash G/H)$ we only need to prove that $f \circ \pi = \tilde{f} \circ \pi_R \in B_{\mathbb{C}}(G)$. Noting that compact groups are unimodular, we see by Folland (2015, Equation 2.52) that $\Delta^k \tilde{f}$ is L^2-integrable with respect to the Haar measure on G/H if and only $\Delta^k \tilde{f} \circ \pi_R$ is L^2-integrable with respect to the Haar measure on G. Now, we show in Appendix D that $(\Delta^k \tilde{f}) \circ \pi_R = \Delta^k(\tilde{f} \circ \pi_R)$, so Theorem 50 tells us that $\tilde{f} \circ \pi_R \in B_{\mathbb{C}}(G)$. ∎

Example 55 By Corollary 54, any kernel $k = f(\langle \cdot, \cdot \rangle)$ that is a smooth function of the inner product has a PD decomposition on S^n (see Example 28 for the description of functions on the double coset space of the sphere $S^n \cong SO(n+1)/SO(n)$, i.e. of invariant kernels on $S^n \cong SO(n+1)/SO(n)$). For instance, the hyperbolic tangent kernel $k = \tanh(a\langle \cdot, \cdot \rangle + b)$ on a sphere S^n, for $a, b \in \mathbb{R}$, has a PD decomposition, despite the fact that it is not PD in general (see Smola et al. (2000, Example 5)).

We now consider the case where $M = G/H$ is a symmetric space of the non-compact type. We may choose G and H such that G is connected, semisimple, has a finite center, and is non-compact, and H is compact. The Riemannian metric on such a non-compact symmetric space is G-invariant and always arises from a left-invariant right-H-invariant metric on G (Helgason, 1962). In Corollary 56 and its proof below, we will use a shorthand and say a function 'decays like' or 'grows like' another function if it decays at least as fast or grows at most as fast as another function.

Corollary 56 *Suppose G/H is a non-compact semisimple symmetric space, with G and H chosen as above. Suppose $f \in C_{\mathbb{C}}(H\backslash G/H)$ is a function of the Riemannian distance squared, i.e.*

$$f \circ \pi_L = \tilde{f} = g(d(eH, \cdot)^2).$$

If $g \in C_{\mathbb{C}}^{\infty}([0, \infty))$ is such that it and its derivatives decay exponentially at infinity, i.e.

$$g^{(k)}(x) = o(e^{-c_k x}) \text{ as } x \to \infty$$

for all $k \geq 0$ and some $c_k > 0$ possibly depending on k, then $f \in B_{\mathbb{C}}(H \backslash G / H)$. In particular, the Gaussian function $\exp(-\lambda d(eH, \cdot)^2) \in B_{\mathbb{R}}(H \backslash G / H)$, and the corresponding Gaussian kernel $k = \exp(-\lambda d(\cdot, \cdot)^2)$ has a PD decomposition for any $\lambda > 0$.

Proof This follows from Theorem 54 combined with two of Helgason's formulas for non-compact semisimple symmetric spaces, the first one for integration and the other for the radial part of the Laplace-Beltrami operator. Observe that since G is semisimple, it is also unimodular (Helgason, 1962, Chapter IV Proposition 1.4), so we are in a position to apply Theorem 54. Also note that since the metric on G/H is left-invariant, any function of the distance squared $g(d(eH, \cdot))^2$ defines a function in $C_{\mathbb{C}}(H \backslash G / H)$.

Using the notation from Said et al. (2018, Equation 13), for $\tilde{f}$ an integrable function on G/H,

$$\int_{G/H} \tilde{f}(x) d\nu(x) = C \int_H \int_{\mathfrak{a}} \tilde{f}(x(a, h)) D(a) da \, d\mu(h)$$

for some constant C independent of $\tilde{f}$. Here $\mathfrak{a}$ is an Abelian subalgebra of $\mathfrak{g}$ resulting from the Iwasawa decomposition $\mathfrak{g} = \mathfrak{h} + \mathfrak{a} + \mathfrak{n}$ where $\mathfrak{g}$, $\mathfrak{h}$ are the Lie algebras of G, H, respectively, $\mathfrak{n}$ is a nilpotent subalgebra of $\mathfrak{g}$, da is the Lebesgue measure on $\mathfrak{a}$, $d\nu$ is a Haar measure on G/H, $d\mu$ is the normalized Haar measure of H, $x(a, h) := \exp(\mathrm{Ad}(h)a)H$ and

$$D(a) := \prod_{\lambda > 0} \sinh^{m_\lambda}(|\lambda(a)|)$$

where the product is over positive roots $\lambda : \mathfrak{a} \to \mathbb{R}$ and m_λ is the dimension of the root space corresponding to λ. See Helgason (1962, Chapter IV Proposition 1.17). Few details will be important for us, only the form of the integral and the growth rate of D at infinity.

Let $\tilde{f}$ and g be as in the statement of the corollary. We first show that $\tilde{f}$ is L^2-integrable. We have

$$\int_{G/H} |\tilde{f}(x)|^2 d\nu(x) = A \int_{\mathfrak{a}} |\tilde{f}(x(a, e))|^2 D(a) da = A \int_{\mathfrak{a}} |g(\|a\|^2)|^2 D(a) da$$

for some constant A, where we have noted that the integrand does not depend on h and have integrated out that variable. The integrand is continuous, so to show that it is finite we only need to consider its behavior at infinity. Since g decays exponentially and D grows exponentially (by using that the roots are linear maps), the integrand decays like a Gaussian at infinity. We deduce that $\tilde{f}$ is L^2-integrable.

Now since $\tilde{f}$ is a function of distance only, $\Delta \tilde{f} = \Delta_r \tilde{f}$ where

$$\Delta_r = \sum_{i=1}^r \frac{\partial^2}{\partial a_i^2} + \sum_{\lambda > 0} m_\lambda \coth(\lambda(a)) \frac{\partial}{\partial a_\lambda} \tag{8}$$

is the radial part of the Laplace-Beltrami operator, where the a_i form an orthonormal basis of $\mathfrak{a}$, the second sum is over positive roots λ, and $\lambda = B(a_\lambda, \cdot)$ where $B : \mathfrak{g} \times \mathfrak{g} \to \mathbb{R}$ is the Killing form. See Helgason (1984, Chapter II Theorem 5.24).

Since g and $d(eH, \cdot)^2$ are smooth, so is $\tilde{f}$. Thus, to show that for $k \geq 0$, $\Delta^k \tilde{f}$ is L^2-integrable, it suffices to consider its behavior at infinity. It is not too hard to see from (8) that since $\tilde{f}(x(a, h)) = g(\|a\|^2) = g(a_1^2 + \cdots + a_r^2)$ and the derivatives of g and of coth decay

exponentially at infinity, $\Delta^k \tilde{f}$ decays like a Gaussian at infinity. Hence, arguing as above we deduce that $\Delta^k \tilde{f}$ is L^2-integrable. So we are done by Theorem 54. ∎

Remark 57 We can show the PD decomposability of the Gaussian kernel for reductive non-compact symmetric spaces too. For instance, the space of $n \times n$ symmetric positive definite matrices $\mathbb{S}_{++}^n \cong GL(n)/O(n)$ is not a semisimple symmetric space. However, we can also write it as $\mathbb{S}_{++}^n \cong SL(n)/SO(n) \times \mathbb{R}_{>0}$, with the diffeomorphism

$$\begin{aligned} \mathbb{S}_{++}^n &\to SL(n)/SO(n) \times \mathbb{R}_{>0} \\ X &\mapsto (X/\det(X)^n, \det(X)) \end{aligned} \tag{9}$$

where $SL(n)/SO(n)$ is viewed as the space of symmetric positive definite matrices with unit determinant. Then, if we write d for the Riemannian distance on $\mathbb{S}_{++}^n \cong GL(n)/O(n)$, sometimes called the affine-invariant Riemannian distance, and d' for the Riemannian distance on $SL(n)/SO(n)$, we can view the map (9) as an isometry:

$$d(X,Y) = d'(X/\det(X)^n, Y/\det(Y)^n) + |\log(\det(X)) - \log(\det(Y))|$$

for all $X,Y \in \mathbb{S}_{++}^n$. Therefore, the Gaussian kernel on $\mathbb{S}_{++}^n$ can be viewed as

$$\begin{aligned} &\exp(-\lambda d(X,Y)^2) \\ &= \exp(-\lambda d'(X/\det(X)^n, Y/\det(Y)^n)^2) \exp(-\lambda |\log(\det(X)) - \log(\det(Y))|^2) \end{aligned}$$

a product of a PD decomposable kernel by Corollary 56 with a PD kernel, which is PD decomposable by Example 20. See also Azangulov et al. (2024b).

6 Conclusion

We have seen that kernel methods can be applied with non-PD kernels that admit a PD decomposition and that knowledge of the specific form of the decomposition is not necessary for RKKS learning. We have argued that RKKS-based methods are particularly interesting in non-Euclidean settings where PD kernels are challenging to construct. We have then related PD decomposable invariant kernels on locally compact quotient spaces G/H to Hermitian PD decomposable functions on the double coset space $H\backslash G/H$. This allowed us to leverage the extensive harmonic analytic literature on the Fourier-Stieltjes algebra $B_{\mathbb{C}}(G)$ to describe the invariant kernels that admit a PD decomposition. In particular, assuming some geometry on G/H, we showed that smoothness and appropriate decay of the derivatives at infinity is sufficient for the PD decomposability of a kernel, providing weak and verifiable sufficient conditions for a kernel to admit a PD decomposition. This work provides theoretical foundations for kernel methods with non-PD kernels and for their application to non-Euclidean data spaces.

Acknowledgments and Disclosure of Funding

Appendix A.

In this appendix we expand on the RKKS representer theorem described in Section 2.2.
Recall that we are given

1. $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset (X \times \mathbb{R})^N$ a finite data set,

2. $g : \mathbb{R} \to \mathbb{R}$ a strictly monotonic differentiable function,

3. $L : \mathcal{K} \times (X \times \mathbb{R})^N \to \mathbb{R}$ a loss functional, determined exclusively through function evaluations, Fréchet differentiable in the first argument,

4. $I : \mathcal{K} \times (X \times \mathbb{R})^N \to \mathbb{R}^m$ and $E : \mathcal{K} \times (X \times \mathbb{R})^N \to \mathbb{R}^l$ functionals determined exclusively through function evaluations and Fréchet differentiable in the first argument.

For any hope at formulating, let alone proving, an RKKS representer theorem, a key additional requirement is that the set defined by the constraints

$$\mathcal{M} := \{f \in \mathcal{K} : I(f, \mathcal{D}) \leq 0 \text{ and } E(f, \mathcal{D}) = 0\}$$

is an infinite-dimensional submanifold (with boundary) of $\mathcal{K}$. By infinite-dimensional submanifold we mean here a Hilbert manifold with Fréchet differentiable transition maps, since $\mathcal{K}$ is given the topology of the Hilbert space $\mathcal{H}_+ \times \mathcal{H}_-$ (see Definition 6). Given this, we have a well-defined notion of a solution to the problem

$$\underset{f \in \mathcal{K}}{\text{stabilize}} \, L(f, \mathcal{D}) + g(\langle f, f \rangle_{\mathcal{K}})$$

$$\text{s.t. } f \in \mathcal{M}.$$

Namely a solution f^* is one for which the Fréchet derivative of the regularized loss $D[L(\cdot, \mathcal{D}) + g(\langle \cdot, \cdot \rangle_{\mathcal{K}})]|_{\mathcal{M}}(f_*)$ vanishes on any tangent vector to the manifold $\mathcal{M}$.

Moreover, we assume that the map $(I \times E)(\cdot, \mathcal{D}) : \mathcal{K} \to \mathbb{R}^{m+l}$ is a submersion. Note that when there are no inequality constraints I, this automatically implies that $\mathcal{M}$ is a manifold (Abraham et al., 1988, Theorem 3.5.4).

Given these assumptions, we can apply (Abraham et al., 1988, Corollary 3.5.29) to conclude that for each critical point $f^* \in \mathcal{K}$ of the restriction of $L(\cdot, \mathcal{D}) + g(\langle \cdot, \cdot \rangle_{\mathcal{K}})$ to $\mathcal{M}$, there exist constants $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^l$ such that f^* is a critical point of the corresponding Lagrange function defined by

$$\mathcal{L}(f) = L(f, \mathcal{D}) + g(\langle f, f \rangle_{\mathcal{K}}) + \lambda^\top I(f, \mathcal{D}) + \mu^\top E(f, \mathcal{D}),$$

that is the Fréchet derivative $D\mathcal{L}(f^*) = 0$. We recall that $\lambda_i I_i(f^*, \mathcal{D}) = 0$ for all $i \in \{1, \ldots, m\}$, which implies that λ_i is non-zero only when the corresponding inequality constraint is active, that is $I_i(f^*, \mathcal{D}) < 0$.

By the Fréchet differentiability of the various functions, we can apply the Fréchet derivative chain rule:

$$0 = D\mathcal{L}(f^*)$$
$$= \sum_{i=1}^{N} \partial_{f(x_i)}(L(f^*, X) + \lambda^T I(f^*, \mathcal{D}) + \mu^T E(f^*, \mathcal{D}))k(x_i, \cdot) + 2f^* g'(\langle f^*, f^* \rangle_{\mathcal{K}}) \tag{10}$$

where on the RHS, the derivatives are classical derivatives. So we get

$$f^* = -\frac{1}{2g'(\langle f^*, f^* \rangle_{\mathcal{K}})} \sum_{i=1}^{N} \partial_{f(x_i)}(L(f^*, \mathcal{D}) + \lambda^T I(f^*, \mathcal{D}) + \mu^T E(f^*, \mathcal{D}))k(x_i, \cdot)$$

where we used the strict monotonicity of g to deduce $g'(\langle f^*, f^* \rangle_{\mathcal{K}}) \neq 0$. In particular, when $g(x) = cx$ for all $x \in \mathbb{R}$ we get

$$f^* = -\frac{1}{c} \sum_{i=1}^{N} \partial_{f(x_i)}(L(f^*, \mathcal{D}) + \lambda^T I(f^*, \mathcal{D}) + \mu^T E(f^*, \mathcal{D}))k(x_i, \cdot).$$

To find f^* in practice, observe from (10) that for any $f \in \mathcal{V} := \mathrm{span}\{k(x_i, \cdot) : 1 \leq i \leq N\}$ we have $D\mathcal{L}(f) \in \mathcal{V}$. This implies that, to find $f^* = \sum_{i=1}^{N} \alpha_i k(\cdot, x_i)$, it suffices to consider $D\mathcal{L}(f)$ restricted to $\mathcal{V}$. That is, it suffices to solve the finite dimensional critical point problem $\nabla_\alpha \mathcal{L}(f^*, \mathcal{D}) = 0$, and then find $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^l$ such that $I(f^*, \mathcal{D}) \leq 0$ and $E(f^*, \mathcal{D}) = 0$.

Appendix B.

In this appendix, we show that when G is compact, the correspondence from Proposition 32 between Hermitian functions on $H\backslash G/H$ with continuous PD decompositions and kernels on G/H with continuous PD decompositions is a surjection. As explained in Remark 34, this requires PD decomposable invariant kernels to admit invariant PD decompositions.

Theorem 58 *If the locally compact group G is compact and H is a closed subgroup of G, then an invariant kernel k on G/H has a PD decomposition if and only if it has a PD decomposition into invariant kernels.*

Proof For $g \in G$, define T_g to be the operator acting on maps $k : G/H \times G/H \to \mathbb{C}$ by

$$T_g k(aH, bH) = k(gaH, gbH)$$

for all $a, b \in G$. Let k be an invariant PD decomposable kernel on G/H. Then, by invariance of k,

$$k(aH, bH) = \frac{1}{\mu(G)} \int_G T_g k(aH, bH) d\mu(g)$$

for $a, b \in G$, where μ is a Haar measure for G, which is finite by compactness of G. So writing the PD decomposition of k as $k = k_+ - k_-$, we get

$$k = \frac{1}{\mu(G)} \int_G T_g k_+ d\mu(g) - \frac{1}{\mu(G)} \int_G T_g k_- d\mu(g).$$

Both the first term and minus the second term are PD, and they are invariant by left-invariance of the Haar measure, so this is a PD decomposition into invariant kernels. ∎

Appendix C.

The aim of this appendix is to provide an alternative general proof of Theorem 49 using the language of representation theory.

Definition 59 *A representation of G is a group homomorphism*

$$\pi : G \to GL_{\mathbb{C}}(V)$$

for some complex vector space V. A representation is called unitary if V is a Hilbert space and the range of π consists of unitary operators, and it is called continuous if π is continuous.

Proposition 60 *$f \in B_+(G)$ if and only if there is a continuous unitary representation $\pi : G \to GL_{\mathbb{C}}(V)$ and a vector $v \in V$ such that*

$$f = \langle \pi(\cdot)v, v \rangle. \tag{11}$$

Proof $\Longleftarrow$: For f as in (11), $g_1, \ldots, g_N \in G$ and $c_1, \ldots, c_N \in \mathbb{C}$,

$$
\begin{aligned}
\sum_{i=1}^{N} \sum_{j=1}^{N} \overline{c_i} c_j f(g_i^{-1} g_j) &= \sum_{i=1}^{N} \sum_{j=1}^{N} \overline{c_i} c_j \langle \pi(g_i^{-1} g_j) v, v \rangle \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \langle c_j \pi(g_j) v, c_i \pi(g_i) v \rangle \\
&= \left\langle \sum_{j=1}^{N} c_j \pi(g_j) v, \sum_{i=1}^{N} c_i \pi(g_i) v \right\rangle \geq 0
\end{aligned}
$$

using the fact that π is a unitary representation. Moreover, f is continuous and bounded since π is, so $f \in B_+(G)$.

$\Longrightarrow$: It can be obtained by combining Theorem 3.20 and Proposition 3.35 in Folland (2015). ∎

Proposition 61 *$f \in B_{\mathbb{C}}(G)$ if and only if there is a continuous unitary representation $\pi : G \to GL_{\mathbb{C}}(V)$ and vectors $v, w \in V$ such that*

$$f = \langle \pi(\cdot)v, w \rangle. \tag{12}$$

Proof $\Longrightarrow$: We show that functions f of the form (12) form a complex linear space. Closure under scalar multiplication can be obtained by scaling v. For closure under addition, observe that

$$\langle \pi_1(\cdot)v_1, w_1 \rangle + \langle \pi_2(\cdot)v_2, w_2 \rangle = \langle (\pi_1 \oplus \pi_2)(\cdot)(v_1 \oplus v_2), w_1 \oplus w_2 \rangle$$

and $\pi_1 \oplus \pi_2$ is a continuous unitary representation if π_1 and π_2 are. Thus, functions of the form (12) form a complex linear space, and contain $B_+(G)$ by Proposition 60, so they contain $B_{\mathbb{C}}(G)$.

$\Longleftarrow$: See Kaniuth and Lau (2018, Lemma 2.1.4). ∎

We have described PD and PD decomposable functions on G in the language of representation theory. What about their analogues on $H \backslash G / H$?

Lemma 62 *For $f \in B_+(G)$,*
$$f = \langle \pi(\cdot)v, v \rangle$$
as in (11), we have that $f \in C_{\mathbb{C}}(H\backslash G/H)$ if and only if $\pi(h)v = v$ for all $h \in H$.

Proof $\Longleftarrow$: If $\pi(h)v = v$ for all $h \in H$, then
$$f(hgh') = \langle \pi(hgh')v, v \rangle = \langle \pi(g)\pi(h')v, \pi(h^{-1})v \rangle = \langle \pi(g)v, v \rangle = f(g)$$
for all $g \in G$ and $h, h' \in H$. Thus, $f \in C_{\mathbb{C}}(H\backslash G/H)$.
$\Longrightarrow$: If $f \in C_{\mathbb{C}}(H\backslash G/H)$, for $h \in H$,
$$\langle \pi(h)v, \pi(h)v \rangle = \langle v, v \rangle = f(e) = f(h) = \langle \pi(h)v, v \rangle$$
where the leftmost equality uses the fact that $\pi(h)$ is unitary. Thus,
$$\langle \pi(h)v, v \rangle = \sqrt{\langle \pi(h)v, \pi(h)v \rangle \langle v, v \rangle}$$
and by the equality condition of Cauchy-Schwarz's inequality, we deduce that $\pi(h)v = v$. ∎

For π a linear representation of G into $GL_{\mathbb{C}}(V)$, write
$$W := \{w \in V : \pi(h)w = w \ \forall \, h \in H\} \tag{13}$$
and let $p : V \to W$ denote the orthogonal projection. In Lemma 62, we have shown $p(v) = v$ for the corresponding π and v. For the more general setting $f \in B_{\mathbb{C}}(G)$, we do not necessarily have $p(v) = v$ or $p(w) = w$, but rather the following weaker lemma.

The proof for the forward direction of this lemma is technical, requires quite a bit of work to prove, and relies on Zorn's lemma. The reason for this difficulty is the generality of the setting we are working in, and specifically the generality of the locally compact group G. With such generality, there are few results that we can use about its continuous unitary representations. However, this will pay off with the generality that we shall obtain in Theorem 49.

Lemma 63 *For $f \in B_{\mathbb{C}}(G)$,*
$$f = \langle \pi(\cdot)v, w \rangle$$
as in (12), we have that $f \in C_{\mathbb{C}}(H\backslash G/H)$ if and only if we can also write f as
$$f = \langle \pi(\cdot)p(v), p(w) \rangle$$
with p as defined above.

Proof $\Longleftarrow$: If $g \in G$ and $h, h' \in H$,
$$f(hgh') = \langle \pi(hgh')p(v), p(w) \rangle = \langle \pi(g)\pi(h')p(v), \pi(h^{-1})p(w) \rangle = \langle \pi(g)p(v), p(w) \rangle = f(g)$$
so $f \in C_{\mathbb{C}}(H\backslash G/H)$.
$\Longrightarrow$: fix $g \in G$ and consider the set
$$\mathcal{P} := \{Z : W \le Z \le V \text{ s.t. } Z = \overline{Z} \text{ and } \langle \pi(g)v, w \rangle = \langle \pi(hgh')v, q(w) \rangle$$
$$\forall \, h, h' \in H \text{ where } q : V \to Z \text{ is the orthogonal projection}\}.$$

$\mathcal{P}$ is partially ordered by inclusion. We want to apply Zorn's lemma to $\mathcal{P}$. So we break down our proof into two steps: we first show in step I that every chain—i.e. totally ordered subset—of $\mathcal{P}$ has a lower bound in $\mathcal{P}$. Zorn's lemma then tells us that $\mathcal{P}$ has a minimal element. Then we show in step II that this minimal element is W. It may seem obvious, but it requires some care since a chain may be uncountable and need not be well-ordered. II, on the other hand, will require us to look more closely at the representation π. We will expand it in irreducible components along a cyclic subgroup of H.

I : Let $\mathcal{C} \subset \mathcal{P}$ be a chain. We want to show that $\mathcal{C}$ has a lower bound in $\mathcal{P}$. This lower bound will, of course, be $Y := \bigcap_{Z \in \mathcal{C}} Z$, so we only need to show $Y \in \mathcal{P}$. Note that for $Z \in \mathcal{C}$, $Y \subset Z$ so $Y^\perp \supset Z^\perp$, thus taking finite sums of elements and closure over such Z we get $Y^\perp \supset \overline{\sum_{Z \in \mathcal{C}} Z^\perp}$. We can also see that for $Z' \in \mathcal{C}$, $Z' \supset \left(\sum_{Z \in \mathcal{C}} Z^\perp\right)^\perp$, so taking the intersection over such Z' we get $Y \supset \left(\sum_{Z \in \mathcal{C}} Z^\perp\right)^\perp$ and taking orthogonal complements $Y^\perp \subset \overline{\sum_{Z \in \mathcal{C}} Z^\perp}$. So we have shown

$$Y^\perp = \overline{\sum_{Z \in \mathcal{C}} Z^\perp}. \tag{14}$$

Let $r : V \to Y$, $r^\perp : V \to Y^\perp$ be the orthogonal projections, and for $Z \in \mathcal{C}$, let $q_Z : V \to Z$, $q_Z^\perp : V \to Z^\perp$ be the orthogonal projections. We have for such Z and $h, h' \in H$,

$$\langle \pi(hgh')v, q_Z(w) \rangle = \langle \pi(g)v, w \rangle = \langle \pi(hgh')v, w \rangle = \langle \pi(hgh')v, q_Z(w) \rangle + \langle \pi(hgh')v, q_Z^\perp(w) \rangle$$

so $\langle \pi(hgh')v, q_Z^\perp(w) \rangle = 0$. Similarly, to show $Y \in \mathcal{P}$, we only need to show $\langle \pi(hgh')v, r^\perp(w) \rangle = 0$. Now by (14), $r^\perp(w)$ can be written as a series of countably many elements of the $Z^\perp$ for $Z \in \mathcal{C}$, so $q_Z^\perp(r^\perp(w))$ gets arbitrarily close to $r^\perp(w)$ as Z decreases (we don't use the notion of limit here as $\mathcal{C}$ may be uncountable). Now $q_Z^\perp(w) = q_Z^\perp(r^\perp(w))$, so $\langle \pi(hgh')v, q_Z^\perp(w) \rangle = 0$ for all $Z \in \mathcal{C}$ implies $\langle \pi(hgh')v, r^\perp(w) \rangle = 0$, and we are done.

II : By Zorn's lemma, $\mathcal{P}$ has a minimal element, which we denote by Z and write $q : V \to Z$ for the corresponding orthogonal projection. If $q(w) = p(w)$, then by minimality of Z we have $Z = W$. Otherwise, suppose first $q(w) \neq p(w)$. Then there is $h \in H$ such that $\pi(h)q(w) \neq q(w)$. Let C_h be the cyclic group generated by h. For conciseness, we will treat the cases C_h finite and C_h infinite together and only use the fact that C_h is Abelian. Indeed, Folland (2015, Theorem 4.45) tells us that we can decompose $\pi|_{C_h}$ into irreducible components in the form

$$\pi(h^k) = \int_{\widehat{C}_h} \xi(h^k) dP(\xi)$$

for all $k \in \mathbb{Z}$, for some regular V-projection-valued measure P on $\widehat{C}_h$ (see Folland (2015, Theorem 1.38) and the following discussion for the definition of a projection-valued measure). Then

$$\langle \pi(g)v, w \rangle = \langle \pi(h'gh'')v, \pi(h^k)q(w) \rangle$$

$$= \left\langle \pi(h'gh'')v, \int_{\widehat{C}_h} \xi(h^k) dP(\xi)q(w) \right\rangle$$

$$= \int_{\widehat{C}_h} \xi(h^k) dP_{\pi(h'gh'')v, q(w)}(\xi)$$

$$= \mathscr{F}^{-1} P_{\pi(h'gh'')v, q(w)}(h^k)$$

for all $k \in \mathbb{Z}$, where $P_{\pi(h'gh'')v,q(w)}$ is a finite complex Radon measure, i.e. $P_{\pi(h'gh'')v,q(w)} \in M_{\mathbb{C}}(\widehat{C}_h)$, and $\mathscr{F}^{-1}$ is the inverse Fourier transform. Then by injectivity of $\mathscr{F}^{-1}$ we must have

$$P_{\pi(h'gh'')v,q(w)} = \langle \pi(g)v, w \rangle \delta_1$$

where δ_1 is the Dirac measure at $\xi = 1$.

Now consider

$$Y := \{z \in Z : \pi(h)z = z\}$$

where h is as before. So for all $h', h'' \in H$,

$$
\begin{aligned}
\langle \pi(g)v, w \rangle &= P_{\pi(h'gh'')v,q(w)}(1) \\
&= \int_{\{1\}} dP_{\pi(h'gh'')v,q(w)}(\xi) \\
&= \left\langle \pi(h'gh'')v, \int_{\{1\}} dP(\xi)q(w) \right\rangle \\
&= \langle \pi(h'gh'')v, P(1)q(w) \rangle \\
&= \langle \pi(h'gh'')v, r(w) \rangle
\end{aligned}
$$

where $r : V \to Y$ is the orthogonal projection. Hence, $Y \in \mathcal{P}$. But $\pi(h)q(w) \neq q(w)$ so $r(w) \neq q(w)$, which implies $Y < Z$. This contradicts the minimality of Z. Thus, $Z = W$, and hence $W \in \mathcal{P}$.

Now by noting that

$$\langle \pi(hgh')v, p(w) \rangle = \langle q(v), \pi((hgh')^{-1})p(w) \rangle = \overline{\langle \pi(h'^{-1}g^{-1}h^{-1})p(w), v \rangle},$$

for all $h, h' \in H$, we can apply the same argument on v to obtain

$$\langle \pi(g)v, w \rangle = \langle \pi(h'gh'')p(v), p(w) \rangle$$

for all $h', h'' \in H$. Finally, $g \in G$ was arbitrary, so we are done.. ∎

We are now in a position to prove Theorem 49 in its full generality, which we restate here for convenience.

Theorem 49 *We have*

1. *$B_{\mathbb{R}}(G) \cap C_{\mathbb{C}}(H \backslash G / H) = B_{\mathbb{R}}(H \backslash G / H)$,*

2. *$B_{\mathbb{C}}(G) \cap C_{\mathbb{C}}(H \backslash G / H) = B_{\mathbb{C}}(H \backslash G / H)$.*

Proof $\supseteq$: For part 1, if $f \in B_{\mathbb{R}}(H \backslash G / H)$ then $f \in C_{\mathbb{C}}(H \backslash G / H)$ by definition, and f has a PD decomposition, so $f \in B_{\mathbb{R}}(G)$. Part 2 is similar.

$\subseteq$: For part 1, let $f \in B_{\mathbb{R}}(G) \cap C_{\mathbb{C}}(H \backslash G / H)$. By Proposition 60, there are continuous unitary representations $\pi_{\pm} : G \to GL_{\mathbb{C}}(V_{\pm})$, and $v_{\pm} \in V_{\pm}$ such that

$$f = \langle \pi_+(\cdot)v_+, v_+ \rangle - \langle \pi_-(\cdot)v_-, v_- \rangle = \langle (\pi_+ \oplus \pi_-)(\cdot)(v_+ \oplus v_-), v_+ \oplus (-v_-) \rangle.$$

Define $W_\pm \leq V_\pm$ as

$$W_\pm := \{w \in V_\pm : \pi_\pm(h)w = w \ \forall \ h \in H\}$$

and $p_\pm : V_\pm \to W_\pm$ the orthogonal projections. For W as in (13) and $p : V \to W$ the orthogonal projection, observe that $W = W_+ \oplus W_-$ and $p = p_+ \oplus p_-$. Then by Lemma 63 we have

$$
\begin{aligned}
f &= \langle (\pi_+ \oplus \pi_-)(\cdot)p(v_+ \oplus v_-), p(v_+ \oplus (-v_-)) \rangle \\
&= \langle (\pi_+ \oplus \pi_-)(\cdot)(p_+(v_+) \oplus p_-(v_-)), p_+(v_+) \oplus p_-(-v_-) \rangle \\
&= \langle \pi_+(\cdot)p_+(v_+), p_+(v_+) \rangle - \langle \pi_-(\cdot)p_-(v_-), p_-(v_-) \rangle.
\end{aligned}
$$

Now $\langle \pi_\pm(\cdot)p_\pm(v_\pm), p_\pm(v_\pm) \rangle \in B_+(H\backslash G/H)$ by Lemma 62, so $f \in B_\mathbb{R}(H\backslash G/H)$ and we are done.

For part 2, the proof follows in a similar way by first writing f as

$$f = \langle \pi_+(\cdot)v_+, v_+ \rangle - \langle \pi_-(\cdot)v_-, v_- \rangle + i\langle \pi_{+i}(\cdot)v_{+i}, v_{+i} \rangle - i\langle \pi_{-i}(\cdot)v_{-i}, v_{-i} \rangle.$$

$\blacksquare$

Remark 64 Note that the proof of Lemma 63 also gives us the results $B_\mathbb{R}(G) \cap C_\mathbb{C}(G/H) = B_\mathbb{R}(G/H)$ and $B_\mathbb{C}(G) \cap C_\mathbb{C}(G/H) = B_\mathbb{C}(G/H)$.

Appendix D.

In this appendix we investigate properties of the Laplace-Beltrami operator on the Lie group G equipped with a left-invariant Riemannian metric, which are needed in Section 5.1. We will use the following definition of the Laplace-Beltrami operator:

$$\Delta := \sum_{i=1}^{n} (X_i^2 - \nabla_{X_i} X_i)$$

locally around a point $g \in G$, where $X_1, \ldots, X_n$ is a local orthonormal frame of vector fields around g and ∇ is the Levi-Civita connection with respect to the metric on G. This definition of Δ is independent of the choice of local orthonormal frame.

In the proof of Theorem 50, we implicitly use the following:

Proposition 65 *Let G be a connected unimodular Lie group. If $X_1, \ldots, X_n$ is a basis of left-invariant vector fields for G, and we equip G with the left-invariant Riemannian metric making $X_1, \ldots, X_n$ orthonormal, then the Laplace-Beltrami operator Δ can be written as*

$$\Delta = \sum_{i=1}^{n} X_i^2.$$

Proof We have

$$\Delta = \sum_{i=1}^{n} (X_i^2 - \nabla_{X_i} X_i),$$

so we need to show that $\sum_{i=1}^{n} \nabla_{X_i} X_i = 0$.

G being unimodular is equivalent to the adjoint representation $\mathrm{ad} : \mathfrak{g} \to \mathrm{Lie}(\mathrm{Aut}(\mathfrak{g}))$ of the Lie algebra $\mathfrak{g}$ of G having vanishing trace at every $X \in \mathfrak{g}$ (Milnor, 1976, Lemma 6.3), where $\mathrm{Lie}(\mathrm{Aut}(\mathfrak{g}))$ is the Lie algebra of $\mathrm{Aut}(\mathfrak{g})$, the Lie group of linear automorphisms of $\mathfrak{g}$. So for all i,

$$
\begin{aligned}
0 &= \mathrm{tr}(\mathrm{ad}(X_i)) \\
&= \sum_{j=1}^{n} \langle [X_i, X_j], X_j \rangle_{\mathfrak{g}} \\
&= \sum_{j=1}^{n} (\underbrace{\langle \nabla_{X_i} X_j, X_j \rangle_{\mathfrak{g}}}_{=0} - \langle \nabla_{X_j} X_i, X_j \rangle_{\mathfrak{g}}) \\
&= -\sum_{j=1}^{n} \langle \nabla_{X_j} X_i, X_j \rangle_{\mathfrak{g}} \\
&= \sum_{j=1}^{n} \langle \nabla_{X_j} X_j, X_i \rangle_{\mathfrak{g}} \\
&= \left\langle \sum_{j=1}^{n} \nabla_{X_j} X_j, X_i \right\rangle_{\mathfrak{g}}
\end{aligned}
$$

where we used the orthonormality of the X_j in the fourth and fifth equality. So $\sum_{i=1}^{n} \nabla_{X_i} X_i = 0$ at e, and thus everywhere on G by left-invariance. ∎

Now in the proof of Theorem 54, we need the following:

Proposition 66 *Let G be a Lie group and let H be a compact Lie subgroup of G. Suppose G is equipped with a left-invariant, right-H-invariant Riemannian metric. Equip G/H with the Riemannian metric induced by the one on G. Then, for $\tilde{f} \in C_{\mathbb{C}}^{\infty}(G/H)$,*

$$
\Delta(\tilde{f} \circ \pi_R) = \Delta \tilde{f} \circ \pi_R
$$

where $\pi_R : G \to G/H$ is the projection, and Δ is the Laplace-Beltrami operator with respect to the respective metrics.

Proof This follows many of the steps of Bergery and Bourguignon (1982, proof of Theorem 1.5). Write n and m for the dimensions of G and H, respectively. For $g \in G$, take a local orthonormal frame $\tilde{X}_1, \ldots, \tilde{X}_{n-m}$ for $T_{gH} G/H$ around gH. This lifts to a local orthonormal frame $X_1, \ldots, X_{n-m}$ for $\mathrm{Ker}(\pi_{R*})^{\perp}$ around g, subbundle of TG. So $\pi_{R*} X_i = \tilde{X}_i$ for all i. Also take a local orthonormal frame $U_1, \ldots, U_m$ for $\mathrm{Ker}(\pi_{R*})$ around g. Then, the Laplace-Beltrami operator on G can be locally written as

$$
\Delta = \sum_{i=1}^{n-m} (X_i^2 - \nabla_{X_i} X_i) + \sum_{i=1}^{m} (U_i^2 - \nabla_{U_i} U_i)
$$

and the Laplace-Beltrami operator on G/H can be locally written as

$$
\Delta = \sum_{i=1}^{n-m} (\tilde{X}_i^2 - \nabla_{\tilde{X}_i} \tilde{X}_i)
$$

36

where ∇ are the Levi-Civita connections induced by the respective metrics. By the chain rule, $U(\tilde{f} \circ \pi_R) = \pi_{R*}U(\tilde{f}) \circ \pi_R = 0$ for all $U \in \text{Ker}(\pi_{R*})$. Similarly, the chain rule gives $X_i^2(\tilde{f} \circ \pi_{R*}) = \pi_{R*}X_i\pi_{R*}X_i(\tilde{f}) \circ \pi_R = \tilde{X}_i^2(\tilde{f}) \circ \pi_R$ for all i. Now by Besse (1987, Theorem 9.80) if $U \in \text{Ker}(\pi_{R*})$ then $\nabla_U U \in \text{Ker}(\pi_{R*})$ (we use here the left-invariance and right-H-invariance of the metric on G, and the compactness of H). Finally, by Besse (1987, Definition 9.23 & following discussion), $\nabla_{X_i}X_i(\tilde{f} \circ \pi_R) = \pi_{R*}(\nabla_{X_i}X_i)(\tilde{f}) \circ \pi_R = \nabla_{\tilde{X}_i}\tilde{X}_i(\tilde{f}) \circ \pi_R$ for all i. Thus, we obtain

$$\Delta(\tilde{f} \circ \pi_R) = \sum_{i=1}^{n-m}(X_i^2 - \nabla_{X_i}X_i)(\tilde{f} \circ \pi_R) + \sum_{i=1}^{m}(U_i^2 - \nabla_{U_i}U_i)(\tilde{f} \circ \pi_R)$$
$$= \sum_{i=1}^{n-m}(\tilde{X}_i^2 - \nabla_{\tilde{X}_i}\tilde{X}_i)(\tilde{f}) \circ \pi_R$$
$$= \Delta\tilde{f} \circ \pi_R.$$

∎

References

Ralph Abraham and Jerrold E. Marsden. *Foundations of Mechanics*. Addison-Wesley, Reading, MA, 2nd edition, 1978.

Ralph Abraham, Jerrold E. Marsden, and Tudor S. Ratiu. *Manifolds, Tensor Analysis, and Applications*, volume 75. Applied Mathematical Sciences. Springer-Verlag, 1988.

Daniel Alpay. Some remarks on reproducing kernel Krein spaces. *Rocky Mountain Journal of Mathematics*, 21(4):1189–1205, December 1991. ISSN 0035-7596. doi: 10.1216/rmjm/1181072903. Publisher: Rocky Mountain Mathematics Consortium.

Iskander Azangulov, Andrei Smolensky, Alexander Terenin, and Viacheslav Borovitskiy. Stationary kernels and Gaussian processes on Lie groups and their homogeneous spaces I: the compact case. *Journal of Machine Learning Research*, 25(280):1–52, 2024a.

Iskander Azangulov, Andrei Smolensky, Alexander Terenin, and Viacheslav Borovitskiy. Stationary kernels and Gaussian processes on Lie groups and their homogeneous spaces II: non-compact symmetric spaces. *Journal of Machine Learning Research*, 25(281):1–51, 2024b.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*, volume 100 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1984. ISBN 978-1-4612-7017-1 978-1-4612-1128-0. doi: 10.1007/978-1-4612-1128-0.

Lionel Bérard Bergery and Jean-Pierre Bourguignon. Laplacians and Riemannian submersions with totally geodesic fibres. *Illinois Journal of Mathematics*, 26(2):181–200, June 1982. ISSN 0019-2082, 1945-6581. doi: 10.1215/ijm/1256046790. Publisher: Duke University Press.

Arthur L. Besse. *Einstein Manifolds.* Springer, Berlin, Heidelberg, 1987. ISBN 978-3-540-74120-6 978-3-540-74311-8. doi: 10.1007/978-3-540-74311-8.

János Bognár. *Indefinite Inner Product Spaces.* Springer, Berlin, Heidelberg, 1974. ISBN 978-3-642-65569-2 978-3-642-65567-8. doi: 10.1007/978-3-642-65567-8.

Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Matérn Gaussian processes on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 33, pages 12426–12437. Curran Associates, Inc., 2020.

S. Calinon. Gaussians on Riemannian manifolds: Applications for robot learning and adaptive control. *IEEE Robotics and Automation Magazine (RAM)*, 27(2):33–45, June 2020. doi: 10.1109/MRA.2020.2980548.

Hyunghoon Cho, Benjamin DeMeo, Jian Peng, and Bonnie Berger. Large-margin classification in hyperbolic space. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1832–1840. PMLR, April 2019. ISSN: 2640-3498.

Nello Cristianini and Elisa Ricci. Support Vector Machines. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*, pages 928–932. Springer US, Boston, MA, 2008. ISBN 978-0-387-30162-4. doi: 10.1007/978-0-387-30162-4_415.

Nathaël Da Costa, Cyrus Mostajeran, and Juan-Pablo Ortega. The Gaussian kernel on the circle and spaces that admit isometric embeddings of the circle. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric Science of Information*, pages 426–435, Cham, 2023a. Springer Nature Switzerland. ISBN 978-3-031-38271-0.

Nathael Da Costa, Cyrus Mostajeran, Juan-Pablo Ortega, and Salem Said. Invariant kernels on Riemannian symmetric spaces: a harmonic-analytic approach, October 2023b. arXiv:2310.19270 [cs, math, stat].

Jacques Faraut and Khelifa Harzallah. Distances hilbertiennes invariantes sur un espace homogène. *Annales de l'Institut Fourier*, 24(3):171–217, 1974. ISSN 1777-5310. doi: 10.5802/aif.525.

Aasa Feragen, François Lauze, and Søren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3032–3042, 2015. doi: 10.1109/CVPR.2015.7298922.

Gerald B. Folland. *A Course in Abstract Harmonic Analysis.* Chapman and Hall/CRC, New York, 2 edition, December 2015. ISBN 978-0-429-15469-0. doi: 10.1201/b19172.

Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian Geometry.* Universitext. Springer, Berlin, Heidelberg, 2004. ISBN 978-3-540-20493-0 978-3-642-18855-8. doi: 10.1007/978-3-642-18855-8.

G. I. Gaudry, S. Meda, and R. Pini. A heat semigroup version of Bernstein's theorem on Lie groups. *Monatshefte für Mathematik*, 110(2):101–114, June 1990. ISSN 1436-5081. doi: 10.1007/BF01302779.

Garth I. Gaudry and Rita Pini. Bernstein's theorem for compact, connected Lie groups. *Mathematical Proceedings of the Cambridge Philosophical Society*, 99(2):297–305, March 1986. ISSN 1469-8064, 0305-0041. doi: 10.1017/S0305004100064215. Publisher: Cambridge University Press.

Colin C. Graham and O. Carruth McGehee. The Behavior of Transforms. In Colin C. Graham and O. Carruth McGehee, editors, *Essays in Commutative Harmonic Analysis*, Grundlehren der mathematischen Wissenschaften, pages 1–47. Springer, New York, NY, 1979. ISBN 978-1-4612-9976-9. doi: 10.1007/978-1-4612-9976-9_1.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25): 723–773, 2012. ISSN 1533-7928.

Babak Hassibi, Ali H. Sayed, and Thomas Kailath. *Indefinite-Quadratic Estimation and Control*. Studies in Applied and Numerical Mathematics. Society for Industrial and Applied Mathematics, January 1999. ISBN 978-0-89871-411-1. doi: 10.1137/1. 9781611970760.

Sigurdur Helgason. *Differential Geometry and Symmetric Spaces*. Academic Press, 1962.

Sigurdur Helgason. Groups and Geometric Analysis. volume 83 of *Mathematical Surveys and Monographs*, Providence, Rhode Island, October 1984. American Mathematical Society. ISBN 978-0-8218-2673-7 978-1-4704-1310-1. doi: 10.1090/surv/083.

Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, Portland, OR, USA, June 2013. IEEE. ISBN 978-0-7695-4989-7. doi: 10.1109/ CVPR.2013.17.

Florent Jousse, Xavier Pennec, Hervé Delingette, and Matilde Gonzalez. Geodesic squared exponential kernel for non-rigid shape registration. pages 1–8. IEEE Computer Society, December 2021. ISBN 978-1-66543-176-7. doi: 10.1109/FG52635.2021.9666997.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences, July 2018. arXiv:1807.02582 [cs, stat].

Eberhard Kaniuth and Anthony Lau. *Fourier and Fourier-Stieltjes Algebras on Locally Compact Groups*, volume 231 of *Mathematical Surveys and Monographs*. American Mathematical Society, July 2018. ISBN 978-0-8218-5365-8 978-1-4704-4768-7 978-1-4704-4769-4. doi: 10.1090/surv/231. ISSN: 0076-5376, 2331-7159.

Yitzhak Katznelson. *An Introduction to Harmonic Analysis*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 3 edition, 2004. ISBN 978-0-521-83829-0. doi: 10.1017/CBO9781139165372.

Siran Li. Gaussian kernels on nonsimply connected closed Riemannian manifolds are never positive definite. *Bulletin of the London Mathematical Society*, 2023. ISSN 1469-2120. doi: 10.1112/blms.12928.

Fanghui Liu, Xiaolin Huang, Yingyi Chen, and Johan Suykens. Fast learning in reproducing kernel Krein spaces via signed measures. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 388–396. PMLR, 13–15 Apr 2021a.

Fanghui Liu, Lei Shi, Xiaolin Huang, Jie Yang, and Johan A. K. Suykens. Analysis of regularized least-squares in reproducing kernel Krein spaces. *Machine Learning*, 110(6): 1145–1173, June 2021b. ISSN 1573-0565. doi: 10.1007/s10994-021-05955-2.

Gaëlle Loosli, Stéphane Canu, and Cheng Soon Ong. Learning SVM in Krein spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(6):1204–1216, 2016. doi: 10.1109/TPAMI.2015.2477830.

Jerrold E. Marsden and Tudor S. Ratiu. *Introduction to mechanics and symmetry*. Springer-Verlag, New York, second edition, 1999.

Xavier Mary. Hilbertian subspaces, subdualities and applications. *Doctoral dissertation, Institut National des Sciences apliquées de Rouen*, 2003.

Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal Kernels. *Journal of Machine Learning Research*, 7(95):2651–2667, 2006. ISSN 1533-7928.

John Milnor. Curvatures of left invariant metrics on lie groups. *Advances in Mathematics*, 21(3):293–329, September 1976. ISSN 00018708. doi: 10.1016/S0001-8708(76)80002-3.

Dino Oglic and Thomas Gaertner. Learning in reproducing kernel Krein spaces. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3859–3867. PMLR, 10–15 Jul 2018.

Cheng Soon Ong, Xavier Mary, Stéphane Canu, and Alexander J. Smola. Learning with non-positive kernels. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, page 81, New York, NY, USA, July 2004. Association for Computing Machinery. ISBN 978-1-58113-838-2. doi: 10.1145/1015330.1015443.

Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 1 edition, April 2016. ISBN 978-1-107-10409-9 978-1-316-21923-2. doi: 10.1017/CBO9781316219232.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, November 2005. ISBN 978-0-262-25683-4. doi: 10.7551/mitpress/3206.001.0001.

Salem Said, Hatem Hajri, Lionel Bombrun, and Baba C. Vemuri. Gaussian distributions on Riemannian symmetric spaces: Statistical learning with structured covariance matrices. *IEEE Transactions on Information Theory*, 64(2):752–772, 2018. doi: 10.1109/TIT.2017. 2713829.

Salem Said, Cyrus Mostajeran, and Simon Heuveline. Chapter 10 - Gaussian distributions on Riemannian symmetric spaces of nonpositive curvature. In Frank Nielsen, Arni S.R. Srinivasa Rao, and C.R. Rao, editors, *Geometry and Statistics*, volume 46 of *Handbook of Statistics*, pages 357–400. Elsevier, 2022. doi: 10.1016/bs.host.2022.03.004.

Salem Said, Simon Heuveline, and Cyrus Mostajeran. Riemannian statistics meets random matrix theory: Toward learning from high-dimensional covariance matrices. *IEEE Transactions on Information Theory*, 69(1):472–481, 2023. doi: 10.1109/TIT.2022.3199479.

Robert Schaback and Holger Wendland. Approximation by positive definite kernels. In Martin D. Buhmann and Detlef H. Mache, editors, *Advanced Problems in Constructive Approximation*, ISNM International Series of Numerical Mathematics, pages 203–222, Basel, 2003. Birkhäuser. ISBN 978-3-0348-7600-1. doi: 10.1007/978-3-0348-7600-1_15.

Frank-Michael Schleif and Peter Tino. Indefinite Proximity Learning: A Review. *Neural Computation*, 27(10):2039–2096, October 2015. ISSN 0899-7667. doi: 10.1162/NECO_a_00770.

Laurent Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (Noyaux reproduisants). *Journal d'Analyse Mathématique*, 13(1):115–256, December 1964. ISSN 1565-8538. doi: 10.1007/BF02786620.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A Generalized Representer Theorem. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, Lecture Notes in Computer Science, pages 416–426, Berlin, Heidelberg, 2001. Springer. ISBN 978-3-540-44581-4. doi: 10.1007/3-540-44581-1_27.

Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

Alex Smola, Zoltán Óvári, and Robert C Williamson. Regularization with dot-product kernels. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert R G Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

Mitsuo Sugiura. Fourier series of smooth functions on compact Lie groups. *Osaka Journal of Mathematics*, 8(1):33–47, January 1971. ISSN 0030-6126. Publisher: Osaka University and Osaka Metropolitan University, Departments of Mathematics.

N. Th Varopoulos. Analysis on Lie groups. *Journal of Functional Analysis*, 76(2):346–410, February 1988. ISSN 0022-1236. doi: 10.1016/0022-1236(88)90041-9.

A. M. Yaglom. Second-order homogeneous random fields. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory*, volume 4.2, pages 593–623. University of California Press, January 1961.