Санкт-Петербургский государственный университет

АЗАНГУЛОВ Искандер Фаритович

Выпускная квалификационная работа

Алгоритмы для гауссовских процессов на некоторых геометрических структурах

Уровень образования: магистратура
Направление 01.04.01 «Математика»
Основная образовательная программа ВМ.5832.2020 «Современная математика»

Научный руководитель:
профессор, д.ф.-м.н.,
Факультет математики и
компьютерных наук СПбГУ
Тихомиров Сергей Борисович

Рецензент:
старший научный сотрудник, Ph.D,
Университетский колледж Лондона
Со Такао

Санкт-Петербург
2022

St Petersburg State University

Azangulov Iskander

Graduation work

Algorithms for Gaussian Processes on Some Geometric Structures

Level of education: Master
Direction 01.04.01 “Mathematics”
The main educational program BM.5832.2020 “Advanced Mathematics”

Supervisor:
Professor, Dr. Sci. (Phys.-Math.),
St Petersburg State University
Sergey Tikhomirov

Reviewer:
Senior Research Fellow, Ph.D.,
University College London,
So Takao

St Petersburg
2022

Contents

1 Introduction

Gaussian process regression based on stationary Gaussian processes provides a powerful framework for data efficient learning in a relatively low dimension. One of the key features of the framework is the ability to quantify uncertainty associated to the predictions. This is often used in applications involving automatic decision making, including optimization [24], reinforcement learning [8] and more [21].

In some applications, inputs of the unknown function lie in a non-Euclidean space like a manifold or a graph. Although one can often model a function like this by embedding these inputs into a Euclidean space, the inner structure of the input space, which is an important modeling assumption, is lost. In practice this hinders data efficiency and impairs overall modeling quality. It is thus important to study Gaussian process regression with inputs on such spaces directly.

Recent developments on that account include [4] where the general case of compact Riemannian manifolds is studied and [3] where the object of consideration are Gaussian processes on graph-structured finite sets.

There exists, however, a number of examples of noncompact manifolds of great significance for applications for which the theory and the corresponding computational techniques are yet to be developed. Arguably, the most important ones are the manifold of postive definite matrices and the hyperbolic space — examples of the class of noncompact *symmetric spaces*.

In this work we study stationary Gaussian processes (most notably, based on heat and Matèrn kernels) on such spaces and computational approaches for Gaussian process regression on them.

These techniques include: efficient (approximate) algorithms for point-wise kernel evaluation and differentiation with respect to parameters; efficient algorithms for sampling, conditioning and sampling from the conditional Gaussian process.

In the next three parts of the introduction we give an overview of Gaussian processes regression on Euclidean spaces. In the forth and final part of the introduction we give a brief but more specific account of goals of the thesis and of the structure of the further text.

Notation We use lowercase bold to denote vectors (e.g. $\boldsymbol{x}$) and uppercase upright bold to denote matrices (e.g. $\mathbf{A}$).

For a function $f(\cdot)$ on $\mathcal{X}$ and $\boldsymbol{x} \in \mathcal{X}^n$ we put $f(\boldsymbol{x}) = [f(x_1), \dots, f(x_n)]$. Similarly, for a function $f(\cdot,\cdot)$ on $\mathcal{X} \times \mathcal{Y}$ and $\boldsymbol{x} \in \mathcal{X}^n, \boldsymbol{y} \in \mathcal{Y}^m$ we put $f(\boldsymbol{x}, \boldsymbol{y}) = [f(x_i, y_j)]_{i \le n, j \le m}$

1.1 Gaussian Processes Regression

We start with definition of *Gaussian processes*. Let us fix an arbitrary set $\mathcal{X}$. A random process $f(\cdot)$ over $\mathcal{X}$ is called *Gaussian* if all its finite-dimensional distributions are multivariate Gaussian: for all $n \in \mathbb{N}$ and $\boldsymbol{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$

$$f(x_1), f(x_2), \dots, f(x_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{1}$$

If we denote by $m(x) = \mathbb{E}\, f(x)$ the *mean function* and by $k(x, x') = \operatorname{Cov}(f(x), f(x'))$ the *covariance function* or *kernel* of the Gaussian process f, then

$$\boldsymbol{\mu} = [m(x_1), m(x_2), \dots, m(x_n)], \qquad \boldsymbol{\Sigma} = [k(x_i, x_j)]_{1 \le i,j \le n}. \tag{2}$$

This illustrates the fact that the mean and covariance functions fully determine the distribution of the process f. The fact that f is a Gaussian process with mean function $m(\cdot)$ and kernel $k(\cdot,\cdot)$ we will denote, following the common practice, by

$$f \sim GP(m, k). \tag{3}$$

Since a covariance matrix is always non-negative definite, a covariance function k must also be non-negative definite. The converse is also true, that any pair of functions $m(\cdot)$ and $k(\cdot,\cdot)$, where m is arbitrary and k is non-negative definite defines a Gaussian process.

The core of applications of Gaussian processes in machine learning is the *Gaussian processes regression*. Suppose we are given noisy measurements $\boldsymbol{y} \in \mathbb{R}^n$ at n inputs $\boldsymbol{x} \in \mathcal{X}^n$ of an unknown function f

$$y_i = f(x_i) + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_n^2), \tag{4}$$

and we want to learn (interpolate) function f.

Gaussian processes regression, solves this problem by conditioning some prior Gaussian process $GP(m(\cdot), k(\cdot,\cdot))$ by observations $\boldsymbol{y}$. It is usually assumed that $m \equiv 0$ or $m \equiv const$ and that k is stationary. The conditional (posterior) process $(f|\boldsymbol{y})$ is also Gaussian with mean m' and covariance k', where

$$m'(\cdot) = m(\cdot) + k(\cdot, \boldsymbol{x})(k(\boldsymbol{x}, \boldsymbol{x}) + \sigma_n^2 \mathbf{I}_n)^{-1}(\boldsymbol{y} - m(\boldsymbol{x})), \tag{5}$$

$$k'(\cdot,\cdot) = k(\cdot,\cdot) - k(\cdot, \boldsymbol{x})(k(\boldsymbol{x}, \boldsymbol{x}) + \sigma_n^2 \mathbf{I}_n)^{-1} k(\boldsymbol{x}, \cdot) \tag{6}$$

and $\mathbf{I}_n$ is the identity matrix of size n.

For input x_* the conditional (posterior) mean $m'(x_*)$ is used as a prediction and the conditional (posterior) variance $k'(x_*, x_*)$ as a measure of uncertainty associated to this prediction.

In many applications [30, 18] conditional (posterior) samples are of interest. The following trick called Matheron's rule [30, 7] helps reduce the problem of sampling the posterior to the problem of sampling the prior and then performing a data dependent update:

$$(f|\boldsymbol{y})(\cdot) = f(\cdot) + k(\cdot, \boldsymbol{x})(k(\boldsymbol{x}, \boldsymbol{x}) + \sigma_n^2 \mathbf{I}_n)^{-1}(\boldsymbol{y} - f(\boldsymbol{x}) - \boldsymbol{\varepsilon}), \tag{7}$$

where the equality is understood in the sense of distributions.

Computations required for evaluating both Eq. (5), Eq. (6) and Eq. (7) are similar and expensive. Algorithmically, they require a time consuming inversion of the matrix $k(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \mathbf{I}_n$ which entails the computational complexity of $O(n^3)$. In Section 1.3 we introduce computational techniques which are used to get approximate answers in a more computationally efficient manner.

Now we turn to the kernels used to define prior Gaussian processes in practical applications.

1.2 Stationary Kernels on Euclidean Spaces

Let us consider the space $\mathbb{R}^d$ of dimension $d \in \mathbb{N}$. Most often the prior Gaussian processes are picked from the class of stationary Gaussian processes, i.e. $GP(0, k(\cdot,\cdot))$ with zero or constant mean (hereinafter we assume that the mean is zero) and a *stationary* kernel. Recall that a kernel $k(\cdot,\cdot)$ on $\mathbb{R}^d \times \mathbb{R}^d$ is called stationary if it is transitionary invariant:

$$k(x + c, y + c) = k(x, y) \qquad \forall x, y, c \in \mathbb{R}^d. \tag{8}$$

In this case $k(x, y)$ depends only on difference $x - y$, so there is a function $k' : \mathbb{R}^d \to \mathbb{R}$

$$k(x, y) = k'(x - y). \tag{9}$$

By Bochner's theorem, stationary kernels are in one-to-one correspondence with non-negative finite measures on $\mathbb{R}^d$:

$$k(x, y) = \int_{\mathbb{R}^d} e^{2\pi i w^T (x-y)} \, \mathrm{d}S(w), \tag{10}$$

where S is a non-negative finite measure called spectral measure. This correspondence gives us the universal way of making stationary kernels from non-negative measures on $\mathbb{R}^d$.

People often consider a more restricted class of *isotropic* kernels i.e. kernels invariant under all isometries of $\mathbb{R}^d$ both translations and (proper and improper) rotations. In this case kernels only depend on the distance between points, so there is a function $k' : \mathbb{R} \mapsto \mathbb{R}$

$$k(x, y) = k'(|x - y|). \tag{11}$$

In this case, the spectral measure S is invariant to rotations and representation (10) can be rewritten in terms of *Fourier–Bessel (Hankel) transform.*

1.2.1 Matérn and Heat Kernels

One family of stationary and isotropic kernels is most often used in applications, the *Matérn* family that we will define in this section. We start this by defining the heat kernel[1] — the single most popular isotropic kernel and the (limiting) member of the Matèrn family.

Recall that heat kernel equation on $\mathbb{R}^d$ is

$$\frac{\partial \mathcal{P}}{\partial t}(t, x, y) = \Delta_x K(t, x, y), \tag{12}$$

The heat kernel is the solution of this equation under the following initial condition

$$\lim_{t \to 0} \mathcal{P}(t, x, y) = \delta_x(y), \tag{13}$$

where where $t > 0$, Δ is the Laplace operator and the limit is taken in sense of distributions.

In applications related to Gaussian process regression $\mathcal{P}$ is reparametrized like this:

$$k_{\infty,\kappa,\sigma^2}(x, y) = \frac{\sigma^2}{C_\kappa} \mathcal{P}(\kappa^2/2, x, y), \tag{14}$$

where C_k normalizes process in such a way that $\mathcal{P}(\kappa^2/2, x, x)$ has variance 1.

The *Matérn* kernel k_{ν,κ,σ^2} may be defined by the following integral

$$k_{\nu,\kappa,\sigma^2}(x, y) = \frac{(2\nu)^\nu}{\Gamma(\nu)\kappa^{2\nu}} \int_0^\infty u^{\nu-1} e^{-\frac{2\nu}{\kappa^2}u} k_{\infty,\sqrt{2u},\sigma^2}(x, y)\, \mathrm{d}u. \tag{15}$$

There are closed form expressions both for $k_{\infty,\kappa,\sigma^2}$ and k_{ν,κ,σ^2}

$$k_{\infty,\kappa,\sigma^2}(x, y) = \sigma^2 e^{-\frac{\|x-y\|^2}{2\kappa^2}}, \tag{16}$$

$$k_{\nu,\kappa,\sigma^2}(x, y) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|x - y\|}{\kappa} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{\|x - y\|}{\kappa} \right), \tag{17}$$

where K_ν is the Bessel function of the second kind.

Spectral measures $S_{\infty,\kappa,\sigma^2}$ and S_{ν,κ,σ^2} of Matérn and heat kernels are also explicit:

$$s_{\infty,\kappa,\sigma^2}(w) = \sigma^2 C_{\infty,\kappa} e^{-2\pi^2\kappa^2|w|^2}; \tag{18}$$

$$s_{\nu,\kappa,\sigma^2}(w) = \sigma^2 C_{\nu,\kappa} \left(\frac{2\nu}{\kappa^2} + 4\pi^2|w|^2 \right)^{-\nu-d/2}. \tag{19}$$

Note that up to re-normalization these are the densities of normal and Student t-distributions.

[1] Also referred to as squared exponential, Gaussian or RBF kernel.

1.3 Computational Algorithms

As it was mentioned before, exact evaluation of posterior mean and variance (Eq. (5) and Eq. (6)) have computational complexity of order $O(n^3)$, where n is the number of observations. Moreover, we did not yet discuss any techniques for sampling the prior Gaussian process.

In this section we consider the commonly used approximate computational techniques to address these challenges.

1.3.1 Sampling and Conditioning via Finite-dimensional Feature Maps

We start with the abstract setting. Consider an arbitrary set of inputs $\mathcal{X}$ and a Gaussian process $f \sim GP(m(\cdot), k(\cdot,\cdot))$ on $\mathcal{X}$.

If there is a vector valued function $\mathbf{\Phi}(\cdot) = [\varphi_1(\cdot), \varphi_2(\cdot), \dots, \varphi_L(\cdot)]^\top$ such that

$$k(x, y) \approx \sum_{l=1}^{L} \varphi_l(x)\varphi_l(y), \tag{20}$$

then we immediately have an approximation of the Gaussian process itself:

$$f(\cdot) \approx \sum_{l=1}^{L} w_l \varphi_l(\cdot), \qquad \text{where } w_l \overset{i.i.d.}{\sim} \mathcal{N}(0, 1). \tag{21}$$

This approximation may be easily used to sample from the prior in $O(L)$ time.

For a vector $\boldsymbol{x} \in \mathcal{X}^n$ let $\mathbf{\Phi}(\boldsymbol{x}) = [\mathbf{\Phi}(x_1), \dots, \mathbf{\Phi}(x_n)]$, then for any $\boldsymbol{x} \in \mathcal{X}^n, \boldsymbol{y} \in \mathcal{X}^m$

$$k(\boldsymbol{x}, \boldsymbol{y}) \approx \mathbf{\Phi}(\boldsymbol{x})^\top \mathbf{\Phi}(\boldsymbol{y}). \tag{22}$$

Then the matrix $k(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \mathbf{I}_n$ from equations (5) – (7) can be approximated by $\mathbf{\Phi}(\boldsymbol{x})^\top \mathbf{\Phi}(\boldsymbol{x}) + \sigma_n^2 \mathbf{I}_n$ and then, using the Woodbury matrix identity we get an approximate conditioning technique with computational complexity $O(nL^2)$. Note that for $L \ll n$ this method scales linearly with respect to n.

1.3.2 Getting Finite-dimensional Feature Maps

Let us show how to find finite-dimensional approximations for stationary kernels on Euclidean spaces. Consider $\mathcal{X} = \mathbb{R}^d$ and a real stationary Gaussian processes $f \sim GP(0, k)$ on $\mathcal{X}$.

Recall that by Bochner's theorem

$$k(x, y) = \int_{\mathbb{R}^d} e^{2\pi i w^\top (x-y)} \, \mathrm{d}S(w) \tag{23}$$

where S denotes the spectral measure.

This decomposition leads to a representation of the kernel as an inner product:

$$k(x, y) = \int_{\mathbb{R}^d} e^{2\pi i w^\top x} \, \overline{e^{2\pi i w^\top y}} \, \mathrm{d}S(w). \tag{24}$$

Using the Monte Carlo approximation of integral and the representation $e^{ix} = \cos(x) + i \sin(x)$ while assuming that k is real-valued we get that $k(x, y) \approx \mathbf{\Phi}^\top(x)\mathbf{\Phi}(y)$ with

$$\mathbf{\Phi}(x) = \sigma \sqrt{\frac{2}{L}} [\cos(w_1^\top x), \sin(w_1^\top x), \dots, \cos(w_L^\top x), \sin(w_L^\top x)]^\top, \qquad w_l \overset{i.i.d.}{\sim} S \tag{25}$$

where $\sigma^2 = k(0,0)$ is the variance of k and L responsible for the quality of the approximation. This approximation is known as the *Random Fourier features* technique [23].

Note that the feature map given by Eq. (25) will always have an even-dimensional output space. To remove this restriction, people also consider a similar feature map called *random phase Fourier features*: $\mathbf{\Phi} = [\varphi_1(\cdot), \dots, \varphi_L(\cdot)]$ with

$$\varphi_i(x) = \sigma\sqrt{\frac{2}{L}}\cos(2\pi w_i^\top x + \theta_i), \qquad \theta_i \overset{i.i.d.}{\sim} U(0, 2\pi), \qquad w_i \overset{i.i.d.}{\sim} S. \tag{26}$$

For approximation error analysis of such techniques see [25].

1.3.3 Efficient Conditioning via Variational Inference

Here we discuss another approach to drive efficient approximate conditioning. The rough idea behind it is to approximate, in terms of KL-divergence, the conditional process by a simpler process. This idea is called *variational inference*. A key ingredient here is an approximating family of simpler processes, the most popular choices for which result in the so called *inducing variables* approximations [27, 14] which we will briefly describe further. One additional advantage of this method is that it allows to approximately condition Gaussian processes by non-Gaussian observations, thus making it possible to use Gaussian process based techniques, for example, for classification.

Let $f(\cdot) \sim GP(0, k(\cdot,\cdot))$ be a prior Gaussian process and $(\boldsymbol{y}, \boldsymbol{x})$ be a data set of size n. We still assume that $p(\boldsymbol{y}|f(\boldsymbol{x})) = \prod_{j=1}^n p(y_j|f(x_j))$, but no longer that $p(y_j|f(x_j))$ is Gaussian.

Let us introduce a set of *inducing points* $\boldsymbol{z} \in \mathbb{R}^m$, where $m \ll n$ and let $\boldsymbol{u} = f(\boldsymbol{z})$ be *inducing variables*. Then the hard-to-compute posterior distribution $p(f, \boldsymbol{u}|\boldsymbol{y})$ is approximated by a family of computationally simpler distributions $q(\boldsymbol{u}, f) = p(f|\boldsymbol{u})q(\boldsymbol{u})$, where $q(\boldsymbol{u})$ is $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as well as locations $\boldsymbol{z}$ being its parameters.

The parameters are found by minimizing Kullback–Leibler divergence

$$D_{KL}(q(\boldsymbol{u}, f)||p(u, f|\boldsymbol{y})) := \mathbb{E}_{q(\boldsymbol{u},f)} \ln \frac{q(\boldsymbol{u}, f)}{p(u, f|\boldsymbol{y})} = \tag{27}$$

$$= D_{KL}(q(\boldsymbol{u})||p(u)) - \sum_{i=1}^n \mathbb{E}_{q(\boldsymbol{u},f)} \ln p(y_i|f(x_i)) - p(y). \tag{28}$$

Note, that the term $p(y)$ does not depend on the parameters, so it is enough to minimize the sum of the first two terms. There are different ways to approach this optimization problem, but more or less all of the ones used in practice use some form of (stochastic) gradient descent.

One particularly simple and general way is called doubly stochastic variational inference algorithm [28]. This technique leads to a computational complexity $O(m^3)$ and when $m \ll n$ the complexity is sublinear in n.

1.4 Goals and Structure of the work

Now, after giving a brief account on Gaussian processes regression on Euclidean spaces we are ready to discuss the goals of this work in more detail. The main goal is to adapt the described notions and the corresponding computational techniques to the case of noncompact symmetric spaces such as the space of symmetric positive definite matrices or the hyperbolic space. We start with a description of stationary processes on them. This leads the statement of following questions.

Question 1. *How to define the notion of the stationarity in general situation?*

Question 2. *How to describe stationary kernels on symmetric spaces?*

Question 3. *How to generalize Matérn and heat kernels?*

When suitable answers to these questions are given we have to suggest computational routines for this kind of processes within the context of Gaussian process regression. This motivates the following additional questions.

Question 4. *How to approximately evaluate the kernels point-wise?*

Question 5. *How to efficiently build a finite dimensional approximation?*

Question 6. *How to adapt the method of a variational inference to this setting?*

Before answering these questions, in Section 2 we give a partial overview of symmetric spaces and harmonic analysis on them.

The Section 3 contains answers to the raised questions. The answer for the first question is given in the sense of invariance to a group action. This is relevant for symmetric spaces as they can be represented as homogeneous spaces of form G/H upon which the group G acts naturally. The answer for the second question can be found in Yaglom [31].

The main contribution of this work is answering the last four questions. To achieve this we adapt the spectral approach described previously. For this purpose we use the theory of spherical Fourier transform developed by Helgason [13]. We give an analogue of the method of random Fourier features and bound its error. Also we generalize Matérn and heat kernels and show how variational methods may be used on such spaces.

In the last section we apply the obtained results for the arguably the most interesting case, the space of symmetric positive definite matrices. For this specific case we additionally suggest an efficient way to sample from the spectral measures to drive better Monte Carlo approximations.

2 Symmetric Spaces

In this section we discuss the symmetric spaces, their properties, their classification and describe some of the key notions of harmonic analysis on these spaces. The section is based on the lecture notes [10] and the books [26, 13].

2.1 Definition and Classification

Symmetric spaces may be defined in geometric and algebraic terms. Let us start with a more intuitive geometric definition.

Definition 7. *Let M be a connected manifold with Riemannian metric g. For a point $p \in M$ let $T_p M$ be a tangent space at p. Let us denote by s_p the geodesic symmetry defined on the image U of a sufficiently small ball $B(0_p, \varepsilon) \subset T_p M$ under the exponential map at point p. It is defined by*

$$s_p \exp(v) = \exp(-v) \qquad \text{where } v \in B(0_p, \varepsilon). \tag{29}$$

Then the space M is called a locally symmetric space if for all points $p \in M$ the map s_p is an isometry. It is called a (globally) symmetric space *if this isometry may be extended to an isometry of the whole space M.*

Informally speaking, a Riemannian manifold (M, g) is a symmetric space if all point reflections are isometries of the manifold. This is true for such spaces as spheres, Grassmannians, Euclidian and hyperbolic spaces etc.

From the definition of a symmetric space it immediately follows that M is geodesically complete (since every geodesic can be continued by reflection). Also, for every $p, q \in M$ we can find a geodesic $\gamma : [0, 1] \mapsto M, \gamma(0) = p, \gamma(1) = q$ that connects p and q. Then point reflection on the middle of $\gamma(1/2)$ sends p to q and q to p. Thus the group of isometries acts transitively on M.

By the Myers-Steenrod theorem [20] the group of isometries of M, that we will denote by G, has natural structure of a Lie group. Let us fix an arbitrary point $p \in M$ on the manifold. It turns out that point p generates a Lie subgroup H of isometries that leaves p fixed (H is called isotropy group). This leads to the representation of M as a homogeneous space G/H.

This representation gives us a way to study symmetric spaces in a Lie theoretic (algebraic) way. The reflection s_p defines the involution σ on G, namely

$$\sigma(g) = s_p \circ g \circ s_p \qquad \text{where } g \in G. \tag{30}$$

Let $\mathfrak{g}$ and $\mathfrak{k}$ be the Lie algebras of G and H respectively. Since σ is an involution ($\sigma^2 = id$) the differential $\theta = \mathrm{D}_e \sigma$ of σ at the identity e is a linear map with $\{+1, -1\}$ eigenvalues. The eigenspace corresponding to $+1$ is exactly $\mathfrak{k}$ — the Lie algebra of H. The eigenspace that corresponds to -1 we denote by $\mathfrak{p}$. It can be identified with the tangent space $T_p M$. The pair $(\mathfrak{k}, \mathfrak{p})$ is called a *Cartan pair*.

Further we will assume that symmetric spaces we consider are simply connected since every symmetric space can be represented as quotient of a simply connected symmetric space and a discrete group. Every simply connected symmetric space M may be further decomposed into a direct product of manifolds

$$M = M_e \times M_c \times M_n, \tag{31}$$

where M_e is isomorphic to a Euclidean space; M_c is a symmetric space with positive sectional curvature (it is said to be of the compact type); M_n is a symmetric space with negative sectional curvature (it is said to be of the noncompact type).

At the same time M_c and M_n may in their turn be decomposed into a product of irreducible symmetric spaces i.e. the spaces that can not be represented as direct product of two symmetric spaces. Élie Cartan [5, 6] has given full classification of irreducible symmetric spaces. More precisely, there are four types of irreducible symmetric spaces, two of compact type (I, II) and two of noncompact type (III, IV):

I G/H where G is compact connected Lie group and H is the subgroup of points fixed by some involution σ of G ($\sigma^2 = Id$),

II compact connected simple Lie groups themselves,

III G/H where G is a connected noncompact simple Lie group and H is a maximal compact subgroup of G,

IV G/H where G is a noncompact connected Lie group and Lie algebra of G is a simple Lie algebra over $\mathbb{C}$ viewed as a real one, and H is a maximal compact subgroup of G.

In this work we will only consider groups of type III. Since, spaces of type I and II are compact, techniques described in [4] may be used. Spaces of type IV originate from complex Lie groups and therefore are less interesting to us. Most importantly, the main examples, e.g. the space of positive symmetric matrices and the hyperbolic space are of type III.

2.2 Lie Structure of Symmetric Spaces

In the preceding section we reviewed the general symmetric spaces. Further we focus on symmetric spaces which on the one the hand generalize spaces of type III and on the other hand possess well developed Harmonic analysis.

Recall that a Lie algebra is called *semisimple* if it can be decomposed into a direct sum of *simple*[2] Lie algebras. A Lie group is called semisimple if its Lie algebra is semisimple. Our spaces of interest have form G/H where G is a noncompact *semisimple* Lie Group with *finite center* and H is a maximal compact subgroup of H. The finite center assumption is made for convenience: in any case the center is always contained in H and is eliminated after factorization. See [12, Ch. VI, Th. 1.1] for details.

Again, let G be a noncompact semisimple Lie group with finite center and let H be its maximal compact subgroup. Also, let $\mathfrak{g}$ and $\mathfrak{k}$ be Lie algebras corresponding to G and to H. By e we denote identity element of the group G.

On a Lie algebra $\mathfrak{g}$ one can define the adjoint action of $X \in \mathfrak{g}$

$$ad_X : \mathfrak{g} \mapsto \mathfrak{g}, \qquad ad_X(Y) = [X, Y] \qquad \text{for } Y \in \mathfrak{g}, \tag{32}$$

where $[\cdot, \cdot]$ denotes the Lie bracket of $\mathfrak{g}$. The Killing form B of Lie algebra $\mathfrak{g}$ is a symmetric bilinear form on $\mathfrak{g} \times \mathfrak{g}$ defined by

$$B(X, Y) = Tr(ad_X \circ ad_Y). \tag{33}$$

Then the subalgebra $\mathfrak{p}$ (see previous section), can be viewed as the orthogonal complement of $\mathfrak{k}$ with respect to B and the differential at e of involution σ is of form

$$D_e\sigma(X + Y) = X - Y, \qquad \text{where } X \in \mathfrak{k}, Y \in \mathfrak{p}. \tag{34}$$

Let us select in the subalgebra $\mathfrak{p}$ the maximal abelian[3] subalgebra $\mathfrak{a}$ and let $\mathfrak{a}^*$ be the space of real linear functionals on $\mathfrak{a}$. For a linear functional $\mathfrak{a}^* \ni \alpha : \mathfrak{a} \mapsto \mathbb{R}$ we can define the root space

$$\mathfrak{g}_\alpha = \{X \in \mathfrak{g} \big| [X, Y] = \alpha(Y)X, \ \forall Y \in \mathfrak{a}\}. \tag{35}$$

and denote by m_α the dimension of g_α. If $m_\alpha > 0$ and $\alpha \neq 0$ then α is called a restricted root. By Σ we will denote the set of restricted roots. Note that the number of restricted roots is finite. One can choose a set of positive roots Σ^+ defined by the following conditions.

- For each restricted root α exactly one of α and $-\alpha$ is contained in Σ^+.
- For any $\alpha, \beta \in \Sigma^+$ either $\alpha + \beta \in \Sigma^+$ or $\alpha + \beta$ is not a restricted root.

Then we can define another Lie subalgebra $\mathfrak{n}$ of $\mathfrak{g}$ by

$$\mathfrak{n} = \oplus_{\alpha \in \Sigma^+} \mathfrak{g}_\alpha. \tag{36}$$

Finally, the *Iwasawa decomposition* of the Lie algebra $\mathfrak{g}$ is

$$\mathfrak{g} = \mathfrak{n} \oplus \mathfrak{a} \oplus \mathfrak{k}. \tag{37}$$

The following claim shows that this decomposition can be lifted to the group level.

[2] Non-abelian Lie algebra is *simple* if it does not contain any non-trivial ideals.

[3] Lie algebra $\mathfrak{a}$ is abelian if $[x, y] = 0$ for all x and y in $\mathfrak{a}$.

Theorem 8. *Let G and H be as above and let $\mathfrak{g} = \mathfrak{l} \oplus \mathfrak{a} \oplus \mathfrak{n}$ be an Iwasawa decomposition of the Lie algebra $\mathfrak{g}$ of G. Then there are subgroups A and N of G with Lie algebras $\mathfrak{a}$ and $\mathfrak{n}$ such that the multiplication map $N \times A \times H \mapsto G$ given by $(n, a, h) \mapsto nah$ is diffeomorphism. Moreover, the subgroups A and N are simply connected.*

For $g \in G$ let us denote by $n(g) \in N, a(g) \in A, h(g) \in H$ such elements that $g = n(g)a(g)h(g)$.

The abelian Lie algebra $\mathfrak{a}$ endowed with the inner product given by the restriction of the Killing form $B|_{\mathfrak{a}\times\mathfrak{a}}$ can be identified with the Euclidean space $\mathbb{R}^l$ for some l. The dimension $\dim \mathfrak{a} := l$ is called the *rank* of the symmetric space G/H.

A vector $X \in \mathfrak{a}$ is called regular if $\alpha(X) \neq 0$ for all $\alpha \in \Sigma$. Since Σ is a finite set, the connected components of the set of regular elements are convex cones. They are called Weyl chambers. By construction, the sign of $\alpha(X)$ is constant on each Weyl chamber. By a_+ we will denote the *positive Weyl chamber*, i.e. the unique Weyl chamber such that all elements of a_+ take positive values on Σ^+. The *Weyl group* W is the finite subgroup of the group of bijective linear transforms $GL(\mathfrak{a})$ generated by reflections relative to the hyperplanes $L_\alpha = \{\alpha(X) = 0 | X \in \mathfrak{a}\}$, where $\alpha \in \Sigma$. This group acts by permutations on the set of Weyl chambers.

Using the Killing form B we can identify $\mathfrak{a}$ and $\mathfrak{a}^*$: for $\lambda \in \mathfrak{a}^*$ we associate $X_\lambda \in \mathfrak{a}$ such that $B(X_\lambda, X) = \lambda(X)$ for all $X \in \mathfrak{a}$. Then, the space $\mathfrak{a}^*$ is naturally endowed with the inner product $\langle \lambda, \mu \rangle = B(X_\lambda, X_\mu)$. Finally, the Weyl chamber and the Weyl group can lifted to $\mathfrak{a}^*$:

$$\mathfrak{a}^*_+ = \{\lambda \in \mathfrak{a}^* | X_\lambda \in \mathfrak{a}^+\} \tag{38}$$

and the Weyl group W is generated by reflections relative to the hyperplanes $L^*_\alpha = \{\alpha(X_\lambda) = 0 | \lambda \in \mathfrak{a}^*\}$ where $\alpha \in \Sigma$.

2.3 Harmonic Analysis

In this section we give a brief introduction to the topic of harmonic analysis on symmetric spaces not going far beyond the definition of the spherical Fourier transform. Again, we consider symmetric spaces of form $M = G/H$ where G is a noncompact semisimple Lie group with finite center and H is its maximal compact subgroup.

Harmonic analysis on symmetric spaces is similar to the harmonic analysis on abelian groups. In abelian case the Fourier transform can be defined in terms of the *Gelfand transform*. The Banach algebra $L^1(A)$ of integrable functions on an abelian group A with multiplication given by convolution is commutative. Thanks to this the Gelfrand transform on algebra $L^1(A)$ is well-defined. For every function from $L^1(A)$ its Gelfand transform is a function on the space of multiplicative linear functionals from algebra $L^1(A)$ to $\mathbb{C}$. All such functionals in they turn may be represented as inner product with group characters i.e. continuous mappings from the group A to the unit circle $\mathbb{S}^1 \subseteq \mathbb{C}$.

The case of a symmetric space is quite similar. For a symmetric space $M = G/H$ let us consider the space $L^1_H(G)$ of H bi-invariant integrable functions on G. This space can be naturally identified with the space of H-invariant functions on G/H. It turns out that $L^1_H(G)$ with multiplication given by convolution is also a commutative Banach algebra. Hence it admits the Gelfand transform as well and the key question to characterize the continuous homomorphisms (multiplicative linear functionals). As shown by Helgason [13, Ch. III, Th. 12.5], these are the inner products with some of the *spherical functions*.

Definition 9. *A* spherical function *on G is a continuous non-zero function φ that satisfies*

$$\varphi(x)\varphi(y) = \int_H \varphi(xhy)\, \mathrm{d}\mu_H(h) \qquad \text{for all } x, y \in G, \tag{39}$$

where μ_H is the Haar measure on the compact group H normalized to be probabilistic.

Note the similarity to the definition of a character χ on an abelian group requiring $\chi(x)\chi(y) = \chi(xy)$.

All continuous homomorphisms of $L^1_H(G)$ onto $\mathbb{C}$ are given by

$$f \mapsto \int_G f(g)\varphi(g^{-1})\,\mathrm{d}\mu_G(g), \tag{40}$$

where μ_G is a Haar measure on G and φ is a *bounded* spherical function.

The following explicit formula for spherical functions was found by Harish-Chandra.

Theorem 10. *Spherical functions on the group G are in one-to-one correspondence with $\mathfrak{a}^*_{\mathbb{C}}/W$, where $\mathfrak{a}^*_{\mathbb{C}}$ is a complexification of the dual space $\mathfrak{a}^*$ and W is the Weyl group[4] (see definitions in Section 2.2). For $\lambda \in \mathfrak{a}^*_{\mathbb{C}}$ we assign the function φ_λ given by*

$$\varphi_\lambda(g) = \int_H e^{(i\cdot\lambda+\rho)A(hg)}\,\mathrm{d}\mu_H(h), \tag{41}$$

here $\rho = \frac{1}{2}\sum_{\alpha\in\Sigma^+} m_\alpha \alpha$ is the half-sum of all positive roots α weighted by the dimensions m_α of the corresponding root spaces $\mathfrak{g}_\alpha$ while $A(g) = \log a(g)$ is the logarithm of the a-part of the Iwasawa decomposition $g = n(g)a(g)k(g)$. Factorization by W means that for all $w \in W$ and $\lambda \in \mathfrak{a}^*_{\mathbb{C}}$ one has $\varphi_\lambda \equiv \varphi_{w\lambda}$.

Let us here describe some properties of spherical functions.

Result 11.

- *All spherical functions have the same value at the identity*

$$\varphi_\lambda(e) = 1. \tag{42}$$

- *Spherical functions are eigenfunctions of the Laplace–Beltrami operator Δ_G on G:*

$$\Delta_G\,\varphi_\lambda = -(|\lambda|^2 + |\rho|^2)\varphi_\lambda. \tag{43}$$

- *Spherical functions have the "symmetry" property*

$$\varphi_\lambda(g_1^{-1}g_2) = \int_H e^{(i\lambda+\rho)A(h^{-1}g_2)}e^{(-i\lambda+\rho)A(h^{-1}g_1)}\,\mathrm{d}\mu_H(h). \tag{44}$$

In contrast to the abelian case not all spherical functions are bounded and non-negative definite.

Theorem 12. *The subset Λ_b of $\mathfrak{a}^*_{\mathbb{C}}$ that corresponds to the bounded spherical functions is*

$$\Lambda_b = \mathfrak{a}^* + i\cdot C(\rho), \qquad \text{where } C(\rho) = \mathrm{Conv}\{w\rho \,|\, w \in W\}, \tag{45}$$

ρ was defined in Theorem 10 and Conv *is the notation for convex hull.*

The subset Λ_+ corresponding to non-negative definite functions is a subset of Λ_b. This subset includes $\mathfrak{a}^*$, since by Eq. (44) we can represent φ_λ as a scalar product of functions $p_{\lambda,g}(h) = e^{(i\lambda+\rho)A(h^{-1}g)}$ with respect to the Haar measure on H.

The spherical Fourier transform of a function f on G defined by

$$\hat{f}(\lambda) = \int_G f(g)\varphi_\lambda(g^{-1})\,\mathrm{d}\mu_G(g). \tag{46}$$

Our next goal is to describe Plancherel formula and the inversion formula. We start with the definition of Plancherel measure that is given in terms of a certain c-function.

[4]Elements of Weyl group are linear mappings which can be naturally extended to act on the complexification.

Definition 13. *The Haris-Chandra c-function is defined as*

$$c(\lambda) = I(i\lambda)/I(\rho), \qquad \text{where } I(\nu) = \prod_{\alpha\in\Sigma^+} B\left(\frac{1}{2}m_\alpha, \frac{1}{4}m_{\alpha/2} + \frac{(\nu,\alpha)}{(\alpha,\alpha)}\right) \tag{47}$$

and $B(n,z)$ is considered as the unique analytical continuation of Beta function on subset of the complex plane $\{z \neq 0, \mathfrak{Re}\, z \geq 0\}$, m_α are as in Theorem 10 and inner product is induced by the Killing form.

Theorem 14. *Let $f \in L^2_H(G) \cap L^1_H(G)$ and $\hat{f}$ be the Fourier transform as defined by Eq. (46). Then the inverse transform is given by*

$$f(x) = \text{const} \int_{\mathfrak{a}^*_+} \hat{f}(\lambda)\varphi_\lambda(x)|c(\lambda)|^{-2}\,\mathrm{d}\lambda, \tag{48}$$

where $d\lambda$ is the Lebesgue measure. Moreover, the analog of Plancherel theorem holds true:

$$\int_G |f(g)|^2\, d\mu_G(g) = \int_{\mathfrak{a}^*_+} |f(\lambda)|^2 |c(\lambda)|^{-2}\,\mathrm{d}\lambda. \tag{49}$$

Finally, the map $f \mapsto \hat{f}$ can be extended to an isometry between $L^2_H(G)$ and $L^2\big(\mathfrak{a}^*_+, |c(\lambda)|^{-2}\, d\lambda\big)$.

3 Gaussian Processes on Symmetric Spaces

3.1 Stationary Gaussian Processes

As it was mentioned in the introduction, we are interested in *stationary* Gaussian processes. Because of this, we start with presenting a general notion of stationarity which will be applicable, in particular, to symmetric spaces.

Let X be a set and let G be a group acting on X. Then we will call a kernel $k : X \times X \mapsto \mathbb{C}$ on X stationary with respect to G if

$$k(gx, gx') = k(x, x') \tag{50}$$

for all $g \in G$ and x, x' in X.

A zero-mean process $f = GP(0, k)$ will be called stationary if the kernel k is stationary. One can check that if f is stationary then the processes $f(g\,\cdot)$ and $f(\cdot)$ have same finite-dimensional distributions.

Let us additionally assume that the group G acts transitively on X. This means that for each pair $x, y \in G$ there is an element $g \in G$ such that $gx = y$. In this case X is called a homogeneous space. Further, if we fix some element $x \in G$ and let H be the stabilizer subgroup of G with respect to x i.e.

$$H = \{h \in G |\, hx = x\}, \tag{51}$$

then X will be isomorphic to the left coset space G/H.

Let us now point out an important property of stationary kernels on homogeneous spaces. For any $h \in H$

$$k(g_1 H, g_2 H) = k(g_2^{-1} g_1 H, eH) = k(g_2^{-1} g_1 H, hH) = \tag{52}$$

$$= k(h^{-1} g_2^{-1} g_1 H, H) = k(H g_1^{-1} g_2 H, H), \tag{53}$$

so the kernel k may be identified with a one argument function k' on the *double coset space* $H\backslash G/H$.

Remark 15. *When $X = \mathbb{R}^d$ we can consider two groups of isometries. The group $S(d)$ of shifts and the group $M(d)$ of motions. In the first case the stabilizer is trivial and $S(d) \equiv \mathbb{R}^d$, so the kernel is a one function on $\mathbb{R}^d$ as in Section 1.2. In the second case the stabilizer is $O(d)$ — the group of rotations of $\mathbb{R}^d$ and the kernel is the function of distance between points.*

3.2 Stationary Gaussian Processes on Symmetric Spaces

Since symmetric spaces are also homogeneous, the notion of stationarity from Section 3.1 is applicable for them. All stationary kernels on symmetric spaces of a rather general sort are described by the following result.

Theorem 16. *Consider a symmetric space $M = G/H$ where G is a separable locally compact group of type I.*[5] *A Gaussian process $f \sim GP(0, k)$ is stationary on M if and only if k is of form*

$$k(g_1 H, g_2 H) = \int_{\Lambda_+} \varphi_\lambda(g_2^{-1} g_1) \, \mathrm{d}\mu_k(\lambda), \tag{54}$$

where Λ_+ is the set of indices of non-negative definite spherical functions[6] *and μ_k is a nonnegative finite measure over Λ_+.*

Proof. See [13, Ch. III, Th. 12.9] and [31, Th. 6]. □

As before, we restrict ourselves to the case when G is a noncompact *semisimple* Lie group with *finite center* and H is the maximal compact subgroup of G.

We also impose the square integrability assumption on the function $k'(g) = k(H, gH)$ to be able to use the spherical Fourier transform. As we will see further, our main examples (heat and Matérn kernels) satisfy this assumption. With the assumption, Theorem 16 is modified as follows.

Theorem 17. *Consider a symmetric space $M = G/H$ where G is a* semisimple *Lie group with* finite center *and H is its maximal compact subgroup.*

The Gaussian process $f \sim \mathrm{GP}(0, k)$ is stationary on M with square integrable $k'(g) = k(gH, H)$ if and only if there exists a non-negative function

$$\hat{k} \in L^1(\mathfrak{a}_+^*, |c(\lambda)|^{-2} \, d\lambda) \cap L^2(\mathfrak{a}_+^*, |c(\lambda)|^{-2} \, d\lambda) \tag{55}$$

such that

$$k(g_1 H, g_2 H) = \int_{\mathfrak{a}_+^*} \varphi_\lambda(g_2^{-1} g_1) \hat{k}(\lambda) |c(\lambda)|^{-2} \, \mathrm{d}\lambda, \tag{56}$$

where $\varphi_\lambda(\cdot)$ are spherical functions, $c(\lambda)$ is the Harish-Chandra c-function and $\mathfrak{a}_+^*$ is the positive Weyl chamber.

Proof. Under the square integrability assumption due to Theorem 14 there is a function $\hat{k} \in L^2(\mathfrak{a}_+^*, |c(\lambda)|^{-2})$ on the $\mathfrak{a}_+^*$ such that Eq. (56) is true. The condition $\hat{k} \in L^1(\mathfrak{a}_+^*, |c(\lambda)|^{-2} \, d\lambda)$ is obtained by substituting $g_1 = g_2 = e$ into the Eq. (56). This proves the forward implication.

Substituting $\mu_k(\lambda) = \hat{k}(\lambda)|c(\lambda)|^{-2} \, d\lambda$ into Eq. (54) we get the backwards implication. □

As in the Euclidean case, we call the function $\hat{k}$ the *spectral density* of process f and the measure $\mu_k = \hat{k}(\lambda)|c(\lambda)|^{-2} d\lambda$ the *spectral measure* of process f.

3.3 Computational Algorithms

In this section we develop the counterparts of the computational techniques from Section 1.3 in the setting of symmetric spaces. Notably, in contrast to the Euclidean case, it is unclear how to evaluate the kernels of interest since there are no closed form expressions for them. We discuss how to solve this problem first.

Hereinafter we always assume that a stationary kernel k satisfies the conditions of Theorem 17 and is given in terms of its spectral measure μ_k.

[5] For precise definition see [15]. All symmetric spaces that we consider further satisfy this assumption.

[6] Note that in the paper [31] spherical functions are called *zonal spherical functions* and "spherical functions" are something different.

3.3.1 Pointwise Kernel Evaluation

We start with the evaluation problem. Applying a Monte Carlo approximation to Eq. (56) we can approximate the kernel k by

$$k(g_1H, g_2H) \approx \frac{\sigma^2}{L} \sum_{j=1}^{L} \varphi_{\lambda_j}(g_2^{-1}g_1) \qquad \text{where } \lambda_j \overset{i.i.d.}{\sim} \mu_k/\sigma^2, \tag{57}$$

where $\sigma^2 = k(H, H)$ is the variance of process f.

Eq. (57) still requires integration because $\varphi_\lambda(g) = \int_H e^{(i\lambda+\rho)A(hg)} \, \mathrm{d}\mu_H(h)$. To overcome this we suggest using Monte Carlo approximation again:

$$k(g_1H, g_2H) \approx \frac{\sigma^2}{L} \sum_{j=1}^{L} e^{(i\lambda_j+\rho)A(h_j g_2^{-1} g_1)}, \qquad \text{where } \lambda_j \overset{i.i.d.}{\sim} \mu_k/\sigma^2, h_j \overset{i.i.d.}{\sim} d\mu_H(h). \tag{58}$$

This gives rise to a computational algorithm provided there's a way to sample μ_k/σ^2 and the uniform distribution μ_H over the compact group H.

A natural question arising from this is how to estimate the convergence rate of this approximation.

Theorem 18. *The estimator on the right-hand side of Eq. (58) is unbiased and its standard deviation is bounded by $\sigma^2/\sqrt{L}$ uniformly with respect to g_1 and g_2.*

Proof. Because of the properties of Monte Carlo approximation, it is clear that the estimator is unbiased, hence we only need to estimate the variance. Since the estimator is a sum of independent random variables it is enough to bound the variance of the random variable $\sigma^2 e^{(i\lambda+\rho)A(hg_2^{-1}g_1)}$, where $\lambda \sim \mu_k/\sigma^2$ and $h \sim \mu_H$

$$\mathrm{Var}(\sigma^2 e^{(i\lambda+\rho)A(hg_2^{-1}g_1)}) \leq \int_{\mathfrak{a}_+^*} \int_H \sigma^2 |e^{(i\lambda+\rho)A(hg_2^{-1}g_1)}|^2 \, \mathrm{d}\mu_H(h) \, \mathrm{d}\mu_k(\lambda) \tag{59}$$

Using equations Eq. (44), Eq. (42) and noticing that $A(g), \rho, \lambda$ are real

$$\int_H |e^{(i\lambda+\rho)A(hg_2^{-1}g_1)}|^2 \, \mathrm{d}\mu_H(h) = \varphi_\lambda(e) = 1. \tag{60}$$

Finally,

$$\mathrm{Var}(\sigma^2 e^{(i\lambda+\rho)A(hg_2^{-1}g_1)}) \leq \sigma^2 \int_{\mathfrak{a}_+^*} 1 \, \mathrm{d}\mu_k(\lambda) = \sigma^4. \tag{61}$$

□

3.3.2 Finite-dimensional Feature Maps

As in Euclidean case, a finite-dimensional feature map can be found to drive an analogue of random Fourier features technique.

By virtue of Eq. (44) we may obtain from Eq. (57) the finite-dimensional approximation of the process $f \sim GP(0, k)$

$$f(g) \sim \frac{\sigma}{\sqrt{L}} \sum_{l=1}^{L} w_l e^{(i\lambda_l+\rho)A(h_l g)}, \qquad \text{where } w_l \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \lambda_l \overset{i.i.d.}{\sim} \mu_k/\sigma^2, h_l \overset{i.i.d.}{\sim} \mu_H. \tag{62}$$

As in Eq. (25) the vector-valued feature map $\mathbf{\Phi}(\cdot) = [\varphi_1(\cdot), \dots, \varphi_{2L}(\cdot)]^\top$ is given by

$$\begin{cases} \varphi_{2l-1} = \sqrt{\frac{2}{L}} e^{\rho A(h_l g)} \cdot \cos(\lambda_l A(h_1 g)); \\ \varphi_{2l} = \sqrt{\frac{2}{L}} e^{\rho A(h_l g)} \cdot \sin(\lambda_l A(h_1 g)). \end{cases} \tag{63}$$

Let us denote the process on the right-hand side of Eq. 62 by f_L. By construction the process f_L is Gaussian with covariance

$$k_L(g_1 H, g_2 H) = \frac{\sigma^2}{L} \sum e^{(i\lambda_j + \rho) A(h_j g_1)} e^{(-i\lambda_j + \rho) A(h_j g_2)}. \tag{64}$$

One of the ways to quantify the quality of approximation of f by f_L is to estimate the difference between k and k_L.

In contrast to the previous statement, the speed of convergence is non-uniform and depends on $g_1 H, g_2 H$, but for every compact subset of M the uniform convergence is still true.

On the bright side, unlike approximation 58, approximation 62 is guaranteed to be non-negative definite. This consideration is crucial for many practical applications.

Theorem 19. *Let $k(g_1 H, g_2 H)$ and $k_L(g_1 H, g_2 H)$ be as above, then $k_L(g_1 H, g_2 H)$ is an unbiased estimator of $k(g_1 H, g_2 H)$ and for every compact subset $U \subset M$ the standard deviation is bounded by $C_U \sigma^2 / \sqrt{L}$ uniformly with respect $g_1 H, g_2 H \in U$.*

Proof. As in previous statement, since all terms are independent and identically distributed, it is enough to estimate the variance of one term

$$\operatorname{Var}\left(\sigma^2 e^{(i\lambda+\rho)A(hg_2)} e^{(-i\lambda+\rho)A(hg_2)}\right) \le \int_{\mathfrak{a}_+^*} \int_H \sigma^2 |e^{(i\lambda+\rho)A(hg_1)} e^{(-i\lambda+\rho)A(hg_2)}|^2 \, \mathrm{d}\mu_H(h) \, \mathrm{d}\mu_k(\lambda) = \tag{65}$$

$$= \int_{\mathfrak{a}_+^*} \int_H \sigma^2 e^{2\rho A(hg_1)} e^{2\rho A(hg_2)} \, \mathrm{d}\mu_H(h) \, \mathrm{d}\mu_k(\lambda) = \sigma^4 \int_H e^{2\rho A(hg_1)} e^{2\rho A(hg_2)} \, \mathrm{d}\mu_H(h) \le \tag{66}$$

$$\le \sigma^4 \left(\int_H e^{4\rho A(hg_1)} \, \mathrm{d}\mu_H(h) \right)^{1/2} \left(\int_H e^{4\rho A(hg_2)} \, \mathrm{d}\mu_H(h) \right)^{1/2} = \sigma^4 \varphi_{-3i\rho}^{1/2}(g_1) \varphi_{-3i\rho}^{1/2}(g_2). \tag{67}$$

So the question is reduced to the bounding $\varphi_{-3i\rho}$. Since spherical functions are continuous, $\varphi_{-3i\rho}$ is continuous, so it is bounded on the compact set U and therefore we get the uniform estimation $\sigma^4 \max_{g \in U} \varphi_{-3i\cdot\rho}(g)$. □

Remark 20. *Recall that by Theorem 12 the space of bounded spherical functions Λ_+ is*

$$\Lambda_+ = \mathfrak{a}^* + i \cdot C(\rho), \qquad \text{where } C(\rho) = \operatorname{Conv}\{w\rho \,|\, w \in W\}, \tag{68}$$

ρ is the half-sum of positive roots and W is the Weyl group. Since the Weyl group is generated by reflections, we have that imaginary part for all $\lambda \in \Lambda_+$ is bounded by $\|\rho\|$ which implies that $-3i \cdot \rho \notin \Lambda_+$ and $\varphi_{-3i\cdot\rho}$ is not globally bounded.

Substituting $g_1 = g_2 = g$ into Eq. (65) the Cauchy inequality turns into an equality and since $\operatorname{Var}(X) = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2$ we get

$$\operatorname{Var}\left(\sigma^2 e^{(i\lambda+\rho)A(hg)} e^{(-i\lambda+\rho)A(hg)}\right) = \sigma^4 (\varphi_{-3i\cdot\rho}(g) - 1). \tag{69}$$

Since the spherical function $\varphi_{-3i\cdot\rho}$ is not globally bounded the unbounded variance is not a product of loose bounds, it is an inherent property of the estimator.

3.3.3 Variational Inference

The variational approach mentioned in Section 1.3 is applicable in our current setting as well. As we will see, the only difference is the optimization of locations of inducing points.

Let us recall the main points of variational approach. Let $\boldsymbol{x}, \boldsymbol{y}$ the observed data set of size n and $f \sim GP(0, k)$ be a prior distribution. The first step is to choose some initial set of inducing points $\boldsymbol{z} \in M^n$ and denote $\boldsymbol{u} = f(\boldsymbol{z})$. The second step is to approximate the posterior $p(f, \boldsymbol{u}|\boldsymbol{y})$ with the variational family $p(f|\boldsymbol{u})q(\boldsymbol{u})$, where $q(\boldsymbol{u}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The third step is to optimize $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{z}$ by minimizing the KL-divergence (Eq. 27) using a (stochastic) gradient optimization method.

The process of optimization of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is exactly the same as in Euclidean case. The optimization of locations $\boldsymbol{z}$ of inducing points calls for a gradient optimization technique on the symmetric space. For many examples this optimization is available [1] and can be performed e.g. by means of the PyManOpt library [29].

3.4 Matérn and Heat Kernels

As mentioned in the introduction, Matérn kernels are arguably the most popular family of kernels used in applications in the Euclidean setting. To generalize this family to the setting of symmetric spaces we will utilize their characterization in terms of heat kernel $k_{\infty,\kappa,\sigma^2}$ and heat kernel itself will be defined in terms of the appropriate version of the heat equation.

Recall that heat kernel on an arbitrary smooth Riemannian manifold defined in the same way as in Section 1.2.1

$$\frac{\partial \mathcal{P}}{\partial t}(t, x, y) = \Delta_x K(t, x, y), \tag{70}$$

with initial condition

$$\lim_{t \to 0} \mathcal{P}(t, x, y) = \delta_x(y), \tag{71}$$

where where $t > 0$ and Δ is the Laplace–Beltrami operator and the limit is taken in the sense of distributions.

For a symmetric space the heat kernel [26] can be found in terms of the spherical Fourier transform, in the same way as in Euclidean case, because $\Delta_G \varphi_\lambda(gH) = -(|\lambda|^2 + |\rho|^2)\varphi_\lambda(gH)$:

$$\mathcal{P}(t, xH, yH) = \int_{\mathfrak{a}_+^*} e^{-t(|\lambda|^2+|\rho|^2)} \cdot \varphi_\lambda(y^{-1}x) \cdot |c(\lambda)|^{-2} \, \mathrm{d}\lambda. \tag{72}$$

Then a heat kernel $k_{\infty,\kappa,\sigma^2}$ defined as

$$k_{\infty,\kappa,\sigma^2}(xH, yH) = \sigma^2 \mathcal{P}(\kappa^2/2, xH, yH) = \sigma^2 \int_{\mathfrak{a}_+^*} e^{-\frac{1}{2}\kappa^2(|\lambda|^2+|\rho|^2)} \cdot \varphi_\lambda(y^{-1}x) \cdot |c(\lambda)|^{-2} \, \mathrm{d}\lambda. \tag{73}$$

Note, that unlike the Euclidean case, we do not normalize the kernel so that $k_{\infty,\kappa,\sigma^2}(xH, xH) = \sigma^2$ since $C_{\kappa,\sigma} = k_{\infty,\kappa,\sigma}(xH, xH)$ cannot be computed analytically.

Finally the Matérn kernel k_{ν,κ,σ^2} is defined by Eq. (15) with different normalization:

$$k_{\nu,\kappa,\sigma^2}(xH, yH) = \frac{1}{\Gamma(\nu)} \int_0^\infty u^{\nu-1} e^{-\frac{2\nu}{\kappa^2}u} \, k_{\infty,\sqrt{2u},\sigma^2}(x, y) \, \mathrm{d}u. \tag{74}$$

This formula can be simplified

$$k_{\nu,\kappa,\sigma^2}(xH, yH) = \frac{1}{\Gamma(\nu)} \int_0^\infty u^{\nu-1} e^{-\frac{2\nu}{\kappa^2}u} k_{\infty,\sqrt{2u},\sigma^2}(xH, yH)\,\mathrm{d}u = \tag{75}$$

$$= \frac{1}{\Gamma(\nu)} \int_0^\infty u^{\nu-1} e^{-\frac{2\nu}{\kappa^2}u} \left(\sigma^2 \int_{\mathfrak{a}_+^*} e^{-u(|\lambda|^2+|\rho|^2)} \cdot \varphi_\lambda(y^{-1}x) \cdot |c(\lambda)|^{-2}\,\mathrm{d}\lambda \right) \mathrm{d}u = \tag{76}$$

$$\sigma^2 \frac{1}{\Gamma(\nu)} \int_{\mathfrak{a}_+^*} \left(\int_0^\infty u^{\nu-1} e^{-u(\frac{2\nu}{\kappa^2}+|\lambda|^2+|\rho|^2)}\,\mathrm{d}u \right) \varphi_\lambda(y^{-1}x) \cdot |c(\lambda)|^{-2}\,\mathrm{d}\lambda. \tag{77}$$

By Gradshteyn and Ryzhik [11], Section 3.326, Item 2 the following relation holds

$$\int_0^\infty u^n e^{-au}\,\mathrm{d}u = \Gamma(n+1)a^{-n-1}. \tag{78}$$

Hence, substituting $n = \nu - 1$ and $a = \frac{2\nu}{\kappa^2} + |\lambda|^2 + |\rho|^2$ we finally arrive at

$$k_{\nu,\kappa,\sigma^2}(xH, yH) = \sigma^2 \int_{\mathfrak{a}_+^*} \left(\frac{2\nu}{\kappa^2} + |\lambda|^2 + |\rho|^2 \right)^{-\nu} \varphi_\lambda(y^{-1}x) \cdot |c(\lambda)|^{-2}\,\mathrm{d}\lambda. \tag{79}$$

The derivation of Eq. (79) is somewhat informal but is enough to motivate why this equation should serve as the definition of Matérn kernels in this setting. As by Section 3.2, for Matérn kernels to be well defined we must have

$$\left(\frac{2\nu}{\kappa^2} + |\lambda|^2 + |\rho|^2 \right)^{-\nu} \in L^1(\mathfrak{a}_+^*, |c(\lambda)|^{-2}\,d\lambda) \cap L^2(\mathfrak{a}_+^*, |c(\lambda)|^{-2}\,d\lambda). \tag{80}$$

Theorem 21. *Matérn kernel k_{ν,κ,σ^2} is well defined in terms of Theorem 17 if and only if $\nu > d/2$, where d is the dimension of manifold G/H.*

Remark 22. *Note, that the constant $d/2$ is the same as in the Euclidean case.*

Proof. We need to study for which a, b the integral

$$\int_{\mathfrak{a}_+^*} (a + |\lambda|^2)^{-b} |c(\lambda)|^{-2}\,\mathrm{d}\lambda < \infty, \tag{81}$$

is finite. Since the integral is rotational invariant

$$\int_{\mathfrak{a}_+^*} (a + |\lambda|^2)^{-b} |c(\lambda)|^{-2}\,\mathrm{d}\lambda = \frac{1}{|W|} \int_{\mathfrak{a}^*} (a + |\lambda|^2)^{-b} |c(\lambda)|^{-2}\,\mathrm{d}\lambda \tag{82}$$

where W is the Weyl group. So it is enough to work with the integral on $\mathfrak{a}^*$. To do this we need to look closer at the c-function. Recall that this function is equal up to a constant (this is denoted by $\propto$) to

$$c(\lambda) \propto \prod_{\alpha \in \Sigma^+} B\left(\frac{1}{2} m_\alpha, \frac{1}{4} m_{\alpha/2} + i \frac{(\lambda, \alpha)}{(\alpha, \alpha)} \right). \tag{83}$$

Because $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ we can rewrite formula for $c(\lambda)^{-1}$

$$c(\lambda)^{-1} \propto \prod_{\alpha \in \Sigma^+} \frac{\Gamma\left(\frac{1}{2} m_\alpha + \frac{1}{4} m_{\alpha/2} + i \frac{(\lambda,\alpha)}{(\alpha,\alpha)} \right)}{\Gamma\left(\frac{1}{4} m_{\alpha/2} + i \frac{(\lambda,\alpha)}{(\alpha,\alpha)} \right)}. \tag{84}$$

It is well known [22, Ch. 4, Eq. (5.02)] that

$$\frac{\Gamma(z+s)}{\Gamma(z)} = z^s(1+O(1/z)), \tag{85}$$

as $z \to \infty$ in the sector $|\arg z| \le \pi - \delta$ for any $0 < \delta < \pi$.

On the other hand for a positive fixed $s > 0$ and $\mathfrak{Re} z \ge 0$ function $\Gamma(z+s)/\Gamma(z)$ is continuous, since $\Gamma(z+s)$ is analytic on $\mathfrak{Re} z > -s$ and $\frac{1}{\Gamma(z)}$ is an entire function. Then for any $R > 0$ the function $\frac{\Gamma(z+s)}{\Gamma(z)}$ is bounded in $\{z : |z| < R, \operatorname{Re} z \ge 0\}$. Thus for big enough C_α, C'_α we have

$$\left|\frac{\Gamma\left(\frac{1}{2}m_\alpha + \frac{1}{4}m_{\alpha/2} + i\frac{(\lambda,\alpha)}{(\alpha,\alpha)}\right)}{\Gamma\left(\frac{1}{4}m_{\alpha/2} + i\frac{(\lambda,\alpha)}{(\alpha,\alpha)}\right)}\right| \le C_\alpha + 2\left|\frac{1}{4}m_{\alpha/2} + i\frac{(\lambda,\alpha)}{(\alpha,\alpha)}\right|^{\frac{1}{2}m_\alpha} \le 2\left|C'_\alpha + i\frac{(\lambda,\alpha)}{(\alpha,\alpha)}\right|^{\frac{1}{2}m_\alpha}. \tag{86}$$

Then for $C > C'_\alpha$ we get

$$|c(\lambda)|^{-2} \le \prod_{\alpha\in\Sigma^+} 2\left|C + i\frac{(\lambda,\alpha)}{(\alpha,\alpha)}\right|^{m_\alpha}. \tag{87}$$

Applying the Cauchy inequality for a large positive constant $D > 0$ we can get the upper bound

$$|c(\lambda)|^{-2} \le (2C + D|\lambda|)^{\sum_{\alpha\in\Sigma^+} m_\alpha}. \tag{88}$$

On the other side, for a small enough $\delta > 0$ the set $\Lambda = \{\lambda \in \mathfrak{a}^* : \ (\lambda,\alpha) \ge \delta|\lambda||\alpha| \text{ for all } \alpha \in \Sigma_+\}$ has infinite Lebesgue measure. Therefore, by Eq. (85) there is $\varepsilon > 0$ such that

$$|c(\lambda)|^{-2} \ge (C + \varepsilon|\lambda|)^{\sum_{\alpha\in\Sigma^+} m_\alpha} \tag{89}$$

for all $\lambda \in \Lambda$ such that $|\lambda|$ is big enough.

Comparing Eq. (87) and Eq. (89) we conclude that integral in Eq. (82) converge if and only if the following integral converges

$$\int_{\mathfrak{a}^*} (1+|\lambda|)^{\sum_{\alpha\in\Sigma^+} m_\alpha - 2b}\,\mathrm{d}\lambda. \tag{90}$$

Finally, since $\mathfrak{a}^* \equiv \mathbb{R}^{\dim \mathfrak{a}}$ the conditions for this are quite clear: $2b > \dim \mathfrak{a} + \sum_{\alpha\in\Sigma^+} m_\alpha$. Computing the dimensions in Eq. (37) we conclude that $b > d/2$. Returning to the theorem we get that integrability condition implies $\nu > d/2$ and that square integrability condition implies $\nu > d/4$ which is even weaker. □

4 Application to the Space of Symmetric Positive Definite Matrices

The space of *symmetric positive-definite matrices* arises in many real-world applications [2, 9, 16, 17]. In this section we demonstrate how to apply techniques described in Section 3 to get efficient computational techniques for Gaussian process regression.

Further we fix $d \in \mathbb{N}$ and denote by $\mathcal{P}_d$ the space of symmetric positive definite matrices of size $d \times d$. We denote by $\mathrm{GL}_d(\mathbb{R})$ and $\mathrm{O}_d(\mathbb{R})$ the groups of invertible matrices and orthogonal matrices of size $d \times d$ correspondingly. As $\mathrm{SL}_d(\mathbb{R})$ and $\mathrm{SO}_d(\mathbb{R})$ we denote determinant one subspaces of $\mathrm{GL}_d(\mathbb{R})$ and $\mathrm{O}_d(\mathbb{R})$. And as $\mathbf{I}_d$ we denote the identity matrix of size d.

Let us consider the action of $\mathrm{GL}_d(\mathbb{R})$ on the space $\mathcal{P}_d$

$$\mathbf{Y}[\mathbf{M}] = \mathbf{M}\mathbf{Y}\mathbf{M}^T, \qquad \mathbf{Y} \in \mathcal{P}_d \text{ and } \mathbf{M} \in GL_d(\mathbb{R}). \tag{91}$$

Every symmetric positive-definite matrix $\mathbf{Y} \in \mathcal{P}_d$ admits Cholesky decomposition: $\mathbf{Y} = \mathbf{U}\mathbf{U}^T$ where $\mathbf{U}$ is upper–triangular matrix. Because of that it is clear that the map $\mathbf{M} \mapsto \mathbf{I}_d[\mathbf{M}]$ is surjective. Moreover, for $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathcal{P}_d$ having the Cholesky factors $\mathbf{U}_1, \mathbf{U}_2$ respectively we have that $\mathbf{Y}_1[\mathbf{U}_2\mathbf{U}_1^{-1}] = \mathbf{Y}_2$. Because $\mathbf{I}_d[\mathbf{H}] = \mathbf{H}\mathbf{H}^T = \mathbf{I}_d$ if and only if $\mathbf{H} \in \mathrm{O}_d(\mathbb{R})$ the stabilizer of the action (91) is $\mathrm{O}_d(\mathbb{R})$. Therefore we conclude that $\mathcal{P}_d$ may be identified with $\mathrm{GL}_d(\mathbb{R})/\mathrm{O}_d(\mathbb{R})$. Similarly, the subspace $\mathcal{SP}_d$ of the space $\mathcal{P}_d$ consisting of the matrices with determinant equal to 1 may be represented as $\mathrm{SL}_d(\mathbb{R})/\mathrm{SO}_d(\mathbb{R})$.

The group $\mathrm{GL}_d(\mathbb{R})$ is not semisimple so, strictly speaking, it does not satisfy the assumptions made in Section 3. However, pulling out the determinant, we may represent $\mathcal{P}_d = \mathbb{R}_+ \times \mathcal{SP}_d$, where $\mathbb{R}_+$ is the group of positive real numbers under multiplication. Then the Fourier transform on $\mathcal{P}_d$ is given [26, Th. 1.3.1] in terms of the *Mellin transform*[7] on $\mathbb{R}_+$ and the *spherical Fourier transform* on $\mathrm{SL}_d(\mathbb{R})/\mathrm{SO}_d(\mathbb{R})$. Further in this section we use the notation from [26] and denote by G and H the groups $\mathrm{GL}_d(\mathbb{R})$ and $\mathrm{O}_d(\mathbb{R})$.

All approximation techniques and definitions of heat and Matérn kernels from Section 3 are applicable to $\mathcal{P}_d$. Now we discuss the specifics needed for practical algorithms on the space $\mathcal{P}_d$.

The Iwasawa decomposition of $\mathrm{GL}_d(\mathbb{R})$ states that a matrix $\mathbf{M} \in \mathrm{GL}_d(\mathbb{R})$ can be represented as $\mathbf{M} = n(\mathbf{M})a(\mathbf{M})h(\mathbf{M})$ where $n(\mathbf{M})$ is an upper-triangular matrix with ones on the diagonal, $a(\mathbf{M})$ is a diagonal matrix with positive entries and $h(\mathbf{M})$ is an orthogonal matrix. Note that after grouping the first two factors this decomposition coincides with the RQ[8] decompositon with $n(\mathbf{M})a(\mathbf{M})$ being the upper-triangular and $h(\mathbf{M})$ being the orthogonal parts.

Denote the group of diagonal matrices of size d with positive entries by A_d (this is the A part of the Iwasawa decomposition). Denote its Lie algebra by $\mathfrak{a}$. It is isomorphic to $\mathbb{R}^d$. For $\boldsymbol{\lambda} \in \mathfrak{a}^*$ we define a *power function* $p_{\boldsymbol{\lambda}}$ by

$$p_{\boldsymbol{\lambda}}(\mathbf{Y}) = p_{\boldsymbol{\lambda}}(\mathbf{I}_d[\mathbf{U}]) = \prod_{j=1}^{d} e^{2(i\cdot\lambda_j + \frac{j}{2} - \frac{d+1}{2})\log U_{jj}} = \prod_{j=1}^{d} U_{jj}^{2i\cdot\lambda_j + j - (d+1)/2}, \tag{92}$$

where $\mathbf{U} \in \mathrm{GL}_d(\mathbb{R})$ is an upper triangular matrix such that $\mathbf{Y} = \mathbf{I}_d[\mathbf{U}]$ (the Cholesky factor of $\mathbf{Y}$). These functions play the same role as exponents $e^{(i\lambda+\rho)A(g)}$ in previous section.

The spherical functions are then defined by

$$\varphi_{\boldsymbol{\lambda}}(\mathbf{Y}) = \int_H p_{\boldsymbol{\lambda}}(\mathbf{Y}[\mathbf{H}])\,\mathrm{d}\mu_H(\mathbf{H}) \tag{93}$$

and they satisfy Result 11 with $\boldsymbol{\rho} = [\frac{1}{2} - \frac{d+1}{4}, \ldots, \frac{j}{2} - \frac{d+1}{4}, \ldots, \frac{d}{2} - \frac{d+1}{4}]$.

The Plancherel measure is given in terms of the Harish-Chandra $c-$function which simplifies to

$$|c(\boldsymbol{\lambda})|^{-2} = \prod_{1\le i<j\le d} \pi|\lambda_i - \lambda_j|\tanh(\pi|\lambda_i - \lambda_j|). \tag{94}$$

4.1 Efficient Evaluation of Heat and Matérn Kernels

In this section we focus on Matérn kernel and computational approaches associated with their usage. As we have shown in Section 1.3 this problem can be reduced to the problem of sampling from the corresponding spectral measure.

Let us fix $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathcal{P}_d$ and denote as $\mathbf{U}_1, \mathbf{U}_2$ corresponding Cholesky factors. Then the heat kernel [26, Ex. 1.3.8] is

$$k_{\infty,\kappa,\sigma^2}(\mathbf{Y}_1, \mathbf{Y}_2) = \sigma^2 \int_{\boldsymbol{\lambda}\in\mathfrak{a}^*} e^{-\frac{\kappa^2}{2}\left(|\boldsymbol{\lambda}|^2 + \frac{d^3-d}{48}\right)} \cdot \varphi_{\boldsymbol{\lambda}}(\mathbf{U}_1^{-1}\mathbf{U}_2)|c(\boldsymbol{\lambda})|^{-2}\,\mathrm{d}\boldsymbol{\lambda} \tag{95}$$

[7] Mellin transform $\mathcal{M}$ of function f is $(\mathcal{M}f)(s) = \int_0^\infty x^{s-1}f(x)\,\mathrm{d}x$.

[8] RQ decomposition of a matrix $\mathbf{M}$ is the representation $\mathbf{M} = \mathbf{R}\mathbf{Q}$ where $\mathbf{R}$ is an upper-triangular matrix and $\mathbf{Q}$ is an orthogonal matrix.

and as in Eq. (79) the Matérn kernel is

$$k_{\nu,\kappa,\sigma^2}(\mathbf{Y}_1,\mathbf{Y}_2)=\sigma^2\int_{\boldsymbol{\lambda}\in\mathfrak{a}^*}\left(\frac{2\nu}{\kappa^2}+|\boldsymbol{\lambda}|^2+\frac{d^3-d}{48}\right)^{-\nu}\cdot\varphi_{\boldsymbol{\lambda}}(t_1^{-1}t_2)|c(\boldsymbol{\lambda})|^{-2}\,\mathrm{d}\boldsymbol{\lambda} \tag{96}$$

where $\nu>d(d+1)/4$ by Theorem 21.

Substituting Eq. (94) we obtain that Matérn kernel $k_{\nu,\kappa,\sigma^2}(\mathbf{Y}_1,\mathbf{Y}_2)$ is given by

$$\sigma^2\int_{\boldsymbol{\lambda}\in\mathfrak{a}^*}\left(\frac{2\nu}{\kappa^2}+|\boldsymbol{\lambda}|^2+\frac{d^3-d}{48}\right)^{-\nu}\cdot\varphi_{\boldsymbol{\lambda}}(\mathbf{U}_1^{-1}\mathbf{U}_2)\prod_{1\le i<j\le d}\pi|\lambda_i-\lambda_j|\tanh(\pi|\lambda_i-\lambda_j|)\,\mathrm{d}\boldsymbol{\lambda}. \tag{97}$$

The spectral measure is

$$\left(\frac{2\nu}{\kappa^2}+|\boldsymbol{\lambda}|^2+\frac{d^3-d}{48}\right)^{-\nu}\prod_{1\le i<j\le d}\pi|\lambda_i-\lambda_j|\tanh(\pi|\lambda_i-\lambda_j|)d\boldsymbol{\lambda}. \tag{98}$$

Unfortunately, we are unaware of a direct way to sample from this measure. The naive approach, since we know the density, is to use the MCMC-based sampling. However, such methods are often difficult to use and slow to converge. It turns out though that we can invent a way to sample a similar density. Specifically, let us rearrange the terms

$$\sigma^2\int_{\boldsymbol{\lambda}\in\mathfrak{a}^*}\varphi_\lambda(\mathbf{U}_1^{-1}mU_2)\prod_{1\le i<j\le d}\tanh(\pi|\lambda_i-\lambda_j|)\underbrace{\left(\frac{2\nu}{\kappa^2}+|\boldsymbol{\lambda}|^2+\frac{d^3-d}{48}\right)^{-\nu}\cdot\prod_{1\le i<j\le d}\pi|\lambda_i-\lambda_j|\,\mathrm{d}\boldsymbol{\lambda}}_{p_{\nu,\kappa}(\boldsymbol{\lambda})d\boldsymbol{\lambda}} \tag{99}$$

and sample from the measure with density proportional to $p_{\nu,\kappa}(\lambda)$ as described further in the text.

The way of getting of a finite-dimensional approximation is similar to Eq. (63), it is given by $\boldsymbol{\Phi}(\cdot)=[\varphi_1(\cdot),\dots,\varphi_{2L}(\cdot)]^\top$ with

$$\begin{cases}\varphi_{2l-1}(\mathbf{Y}) &= \sqrt{\frac{2}{L}\prod\tanh(\pi|\lambda_i^l-\lambda_j^l|)}\cdot e^{2\boldsymbol{\rho}(a(\mathbf{H}^l\mathbf{U}))}\cdot\cos(2\boldsymbol{\lambda}^l(a(\mathbf{H}^l\mathbf{U})));\\ \varphi_{2l}(\mathbf{Y}) &= \sqrt{\frac{2}{L}\prod\tanh(\pi|\lambda_i^l-\lambda_j^l|)}\cdot e^{2\boldsymbol{\rho}(a(\mathbf{H}^l\mathbf{U}))}\cdot\sin(2\boldsymbol{\lambda}^l(a(\mathbf{H}^l\mathbf{U}))),\end{cases} \tag{100}$$

where $\mathbf{Y}=\mathbf{I}_n[\mathbf{U}]$ and $\boldsymbol{\lambda}^l,\mathbf{H}^l$ are sampled independently from $p_{\nu,\kappa}(\boldsymbol{\lambda})d\boldsymbol{\lambda},\mu_H(\mathbf{H})$.

Similarly, for the heat kernel we can define the measure

$$p_{\infty,\kappa}(\boldsymbol{\lambda})d\boldsymbol{\lambda}=e^{-\frac{\kappa^2}{2}(|\boldsymbol{\lambda}|^2+|\boldsymbol{\rho}|^2)}\prod_{1\le i<j\le d}\pi|\lambda_i-\lambda_j|d\lambda. \tag{101}$$

and use the same feature map as above.

We start with demonstrating a way of sampling from the measure corresponding to $p_{\infty,\kappa}$.

Theorem 23. *Consider a random matrix $\mathbf{X}=(X_{ij})_{1\le i,j\le d}$, where $X_{ij}\overset{i.i.d.}{\sim}\mathcal{N}(0,1)$ and put $\mathbf{M}=(\mathbf{X}+\mathbf{X}^T)/\sqrt{2}$. The distribution of $\mathbf{M}$ is called* Gaussian Orthogonal Ensemble (GOE).

Let $\lambda_1,\lambda_2,\dots,\lambda_d$ be the eigenvalues of the matrix $\mathbf{M}$ defined above. Let p_{GOE} be the joint density function of the random vector $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\dots,\lambda_n)$. Then

$$p_{GOE}(\lambda_1,\lambda_2,\dots,\lambda_n)\propto e^{-|\boldsymbol{\lambda}|^2/2}\prod_{i<j}|\lambda_i-\lambda_j|. \tag{102}$$

Proof. see [19, Th. 3.3.1]. □

As corollary, we have that the density of the random vector $\kappa^{-1}\boldsymbol{\lambda}$ is proportional to $p_{\infty,\kappa}$.

To sample from measure $p_{\nu,\kappa}(\boldsymbol{\lambda})\,d\boldsymbol{\lambda}$ we present the lemma that generalizes the representation of a Student's t-distribution as the quotient of the standard normal and χ distributions.

Lemma 24. *Let X be a random vector on $\mathbb{R}^n$ with density proportional to $e^{-|\boldsymbol{x}|^2/2}f(\boldsymbol{x})$, where $f(\boldsymbol{x})$ is a non-negative function such that for some $k \in \mathbb{R}$*

$$f(a\boldsymbol{x}) = a^k f(\boldsymbol{x}). \tag{103}$$

Let Y be a random variable independent of X with the χ_d distribution (the squared root of the χ^2_d distribution) for some $d > 0$. Then the random variable $Z = X/Y$ has density

$$p_Z(\boldsymbol{z}) \propto f(\boldsymbol{z})(1+|\boldsymbol{z}|^2)^{-\nu}, \qquad \text{where } \nu = (n+k+d)/2. \tag{104}$$

Proof. We prove this by a direct computation of the density of Z. Recall that $p_Y(y) \propto y^{d-1}e^{-y^2/2}$. Thus

$$p_{Z\times Y}(\boldsymbol{z}, y) = p_{X\times Y}(y\boldsymbol{z}, y)y^n \propto e^{-y^2(|\boldsymbol{z}|^2+1)/2}f(\boldsymbol{z})y^{n+k+d-1}. \tag{105}$$

The marginal distribution of Z is, up to a constant,

$$p_Z(\boldsymbol{z}) \propto f(\boldsymbol{z})\int_0^{+\infty} e^{-y^2(|\boldsymbol{z}|^2+1)/2}y^{n+k+d-1}\,\mathrm{d}y \propto f(\boldsymbol{z})(1+|\boldsymbol{z}|^2)^{-(n+k+d)/2}. \tag{106}$$

□

Substituting the distribution p_{GOE} and $\chi_{2\nu}$ into Lemma 24 we get the random variable Z with density

$$p_Z(\boldsymbol{\lambda}) \propto (1+|\boldsymbol{\lambda}|^2)^{-d(d+1)/4-\nu}\prod|\lambda_i-\lambda_j|, \tag{107}$$

which is up to rescaling the density we need for Matérn kernels.

Remark 25. *Similarly, for this purposes we can use measures with densities*

$$p'_{\nu,\kappa}(\boldsymbol{\lambda}) \propto \left(\frac{2\nu}{\kappa^2}+|\boldsymbol{\lambda}|^2+|\boldsymbol{\rho}|^2\right)^{-\nu}, \qquad p'_{\infty,\kappa}(\boldsymbol{\lambda}) \propto e^{-\frac{\kappa^2}{2}(|\boldsymbol{\lambda}|^2+|\boldsymbol{\rho}|^2)}. \tag{108}$$

As in Eq. (18) up to re-normalization they are the densities of normal and Student t-distributions. The advantage of such approach is that this method is generally applicable. However, as can be seen from Figure 2, for $\mathcal{P}_d$ the approach with $p'_{\nu,\kappa}$ and $p_{\infty,\kappa}$ behaves more favorably.

Let us visualize the kernels and processes defined above. For illustration purposes, we consider kernels on $\mathcal{P}_3$ restricted on the subspace $\mathcal{SP}_3$. We match numbers $\lambda_1, \lambda_2 \in \mathbb{R}$ with the matrix $\mathbf{Y}_{\lambda_1,\lambda_2} \in \mathcal{SP}_3$ given by

$$\mathbf{Y}_{\lambda_1\lambda_2} = \begin{pmatrix} e^{\lambda_1} & 0 & 0 \\ 0 & e^{\lambda_2} & 0 \\ 0 & 0 & e^{-\lambda_1-\lambda_2} \end{pmatrix}. \tag{109}$$

On Figure 1 for $\lambda_1, \lambda_2 \in [-5,5]^2$ and $\kappa \in [0.5, 1, 2, 4]$ the covariance function $k_{\infty,\kappa,1}(\mathbf{I}_3, \mathbf{Y}_{\lambda_1,\lambda_2})$ is plotted, where $\mathbf{I}_3 = \mathbf{Y}_{00}$ the identity matrix. Samples are obtained using the approximation (100) with $L = 1000$. It is clear from Figure 1 that when κ increases samples oscillate less.

On Figure 2 we compare different ways of approximate computations. For random matrices $\mathbf{Y} \in \mathcal{P}_5$ and $\mathbf{M} \in \mathrm{GL}_5(\mathbb{R})$ the kernel $k_{\infty,\kappa,1}(\mathbf{Y}, \mathbf{Y}[e^{c\mathbf{M}}])$ is computed for $c \in [-2, 2]$ and $\kappa \in [0.5, 1, 2, 4]$ using different methods. In the first case we use Eq. (57) for measures $p_{\infty,\kappa}$ (Eq. 101) and $p'_{\infty,\kappa}$ (Eq. 108). In the second case for finite-dimensional approximations with the same measures we compute the kernel using Eq. (64). In all cases the approximations have $L = 2000$. This illustrates the advantage of using $p_{\infty,\kappa}$.

Kernel $k_{\infty,\kappa,1}$: Sample from $GP(0, k_{\infty,\kappa,1})$:

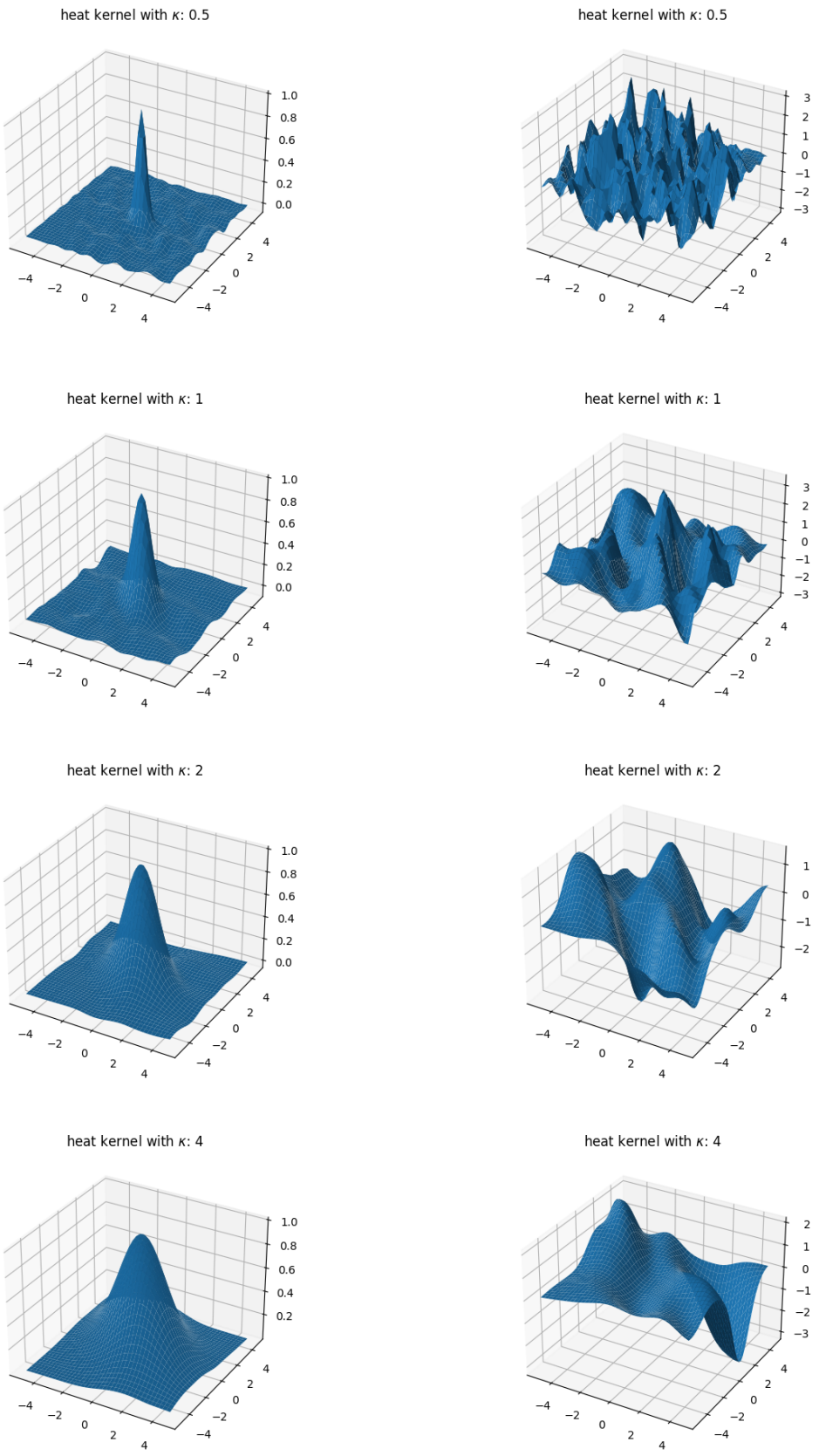

Figure 1: Kernels $k_{\infty,\kappa,1}$ on the left and samples from $GP(0, k_{\infty,\kappa,1})$ on the right.

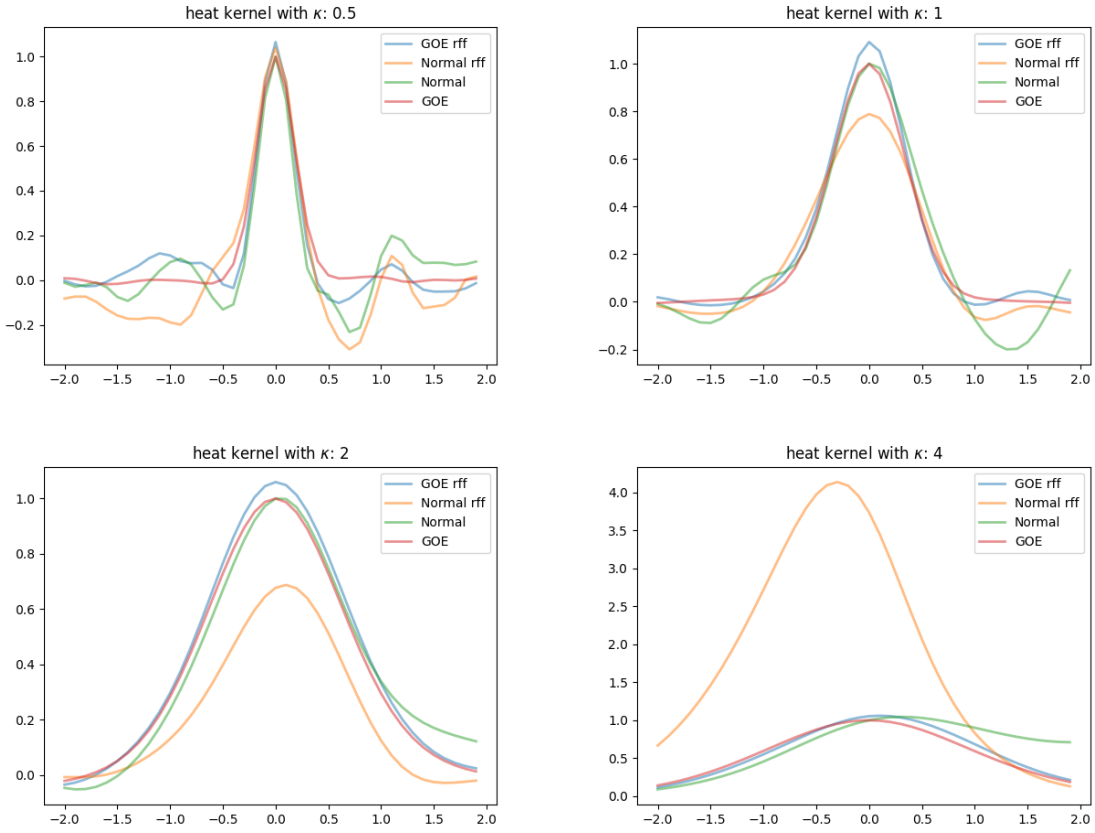

Figure 2: Comparing the approximation approaches for heat kernels with different values of κ. By "Normal" we denote the approximation corresponding to $p'_{\infty,\kappa}$ and by "GOE" we denote approximations corresponding to $p_{\infty,\kappa}$. The "rff" signifies that the approximation was made using the finite-dimensional feature transform.

Acknowledgements

I am profoundly grateful to Viacheslav Borovitskiy for his extensive help and mentorship during my studies. I am grateful to my scientific advisor Sergey Tikhomirov for his guidance. I am also grateful to my collaborators Andrei Smolensky and Alexander Terenin.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. Cited on page 16.

[2] M. Arnaudon, F. Barbaresco, and L. Yang. Riemannian Medians and Means With Applications to Radar Signal Processing. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):595–604, 2013. Cited on page 18.

[3] V. Borovitskiy, I. Azangulov, A. Terenin, P. Mostowsky, M. Deisenroth, and N. Durrande. Matérn Gaussian Processes on Graphs. 130:2593–2601. Cited on page 2.

[4] V. Borovitskiy, A. Terenin, P. Mostowsky, and M. Deisenroth. Matérn Gaussian Processes on Riemannian Manifolds. In *Advances in Neural Information Processing Systems*, pages 12426–12437, 2020. Cited on pages 2, 8.

[5] É. Cartan. Sur une classe remarquable d'espaces de Riemann. *Bulletin de la Société Mathématique de France*, 54:214–264, 1926. Cited on page 8.

[6] É. Cartan. Sur une classe remarquable d'espaces de Riemann. II. *Bulletin de la Société Mathématique de France*, 55:114–134, 1927. Cited on page 8.

[7] J.-P. Chiles and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, 2009. Cited on page 3.

[8] M. P. Deisenroth and C. E. Rasmussen. PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In ICML'11, pages 465–472. Omnipress, 2011. Cited on page 2.

[9] G. Dong and G. Kuang. Target Recognition in SAR Images via Classification on Riemannian Manifolds. *IEEE Geoscience and Remote Sensing Letters*, 12(1):199–203, 2015. Cited on page 18.

[10] J. Eschenburg. Lecture Notes on Symmetric Spaces, Jan. 1997. Cited on page 7.

[11] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 7th edition, 2014. Cited on page 17.

[12] S. Helgason. *Differential Geometry and Symmetric Spaces*. Pure and Applied Mathematics: Academic Press. Academic Press, 1962. ISBN: 9780123384508. Cited on page 9.

[13] S. Helgason. *Geometric Analysis on Symmetric Spaces*. Mathematical surveys and monographs. American Mathematical Soc., 1993. Cited on pages 7, 10, 13.

[14] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian Processes for Big Data. In UAI'13, pages 282–290. AUAI Press, 2013. Cited on page 6.

[15] E. Hewitt, M. Naimark, A. Stern, and E. Hewitt. *Theory of Group Representations*. Grundlehren der mathematischen Wissenschaften. Springer New York. Cited on page 13.

[16] N. Jaquier, V. Borovitskiy, A. Smolensky, A. Terenin, T. Asfour, and L. Rozo. Geometry-aware Bayesian Optimization in Robotics using Riemannian Matérn Kernels. In *Conference on Robot Learning*, pages 794–805. PMLR, 2022. Cited on page 18.

[17] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. 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, 2013. Cited on page 18.

[18] P. D. Jean-Paul Chilès. *Geostatistics: Modeling Spatial Uncertainty, Second Edition*. 2012. Cited on page 3.

[19] M. L. Mehta. *Random Matrices (Revised and Enlarged Second Edition)*. Academic Press, San Diego, revised and enlarged second edition edition, 1991. Cited on page 21.

[20] S. B. Myers and N. E. Steenrod. The Group of Isometries of a Riemannian Manifold. *Annals of Mathematics*, 40(2):400–416, 1939. ISSN: 0003486X. Cited on page 8.

[21] W. Neiswanger, K. A. Wang, and S. Ermon. Bayesian Algorithm Execution: Estimating Computable Properties of Black-box Functions Using Mutual Information, 2021. Cited on page 2.

[22] F. Olver and W. Rheinbolt. *Asymptotics and Special Functions*. Elsevier Science, 2014. Cited on page 18.

[23] A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. Cited on page 6.

[24] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. Cited on page 2.

[25] D. J. Sutherland and J. Schneider. On the Error of Random Fourier Features, 2015. Cited on page 6.

[26] A. Terras. *Harmonic Analysis on Symmetric Spaces—Higher Rank Spaces, Positive Definite Matrix Space and Generalizations*. 2016. Cited on pages 7, 16, 19.

[27] M. Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574. PMLR, 2009. Cited on page 6.

[28] M. Titsias and M. Lázaro-Gredilla. Doubly Stochastic Variational Bayes for non-Conjugate Inference. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of number 2, pages 1971–1979. PMLR, 2014. Cited on page 6.

[29] J. Townsend, N. Koep, and S. Weichwald. Pymanopt: A Python Toolbox for Optimization on Manifolds using Automatic Differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016. Cited on page 16.

[30] J. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. Deisenroth. Efficiently sampling functions from Gaussian process posteriors. In *International Conference on Machine Learning*, pages 10292–10302. PMLR, 2020. Cited on page 3.

[31] 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*, pages 593–622. University of California Press, 1961. Cited on pages 7, 13.