New random projections for isotropic kernels using stable spectral distributions

Nicolas Langrené*[1], Xavier Warin[2], and Pierre Gruet[2]

[1]Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College
[2]EDF Lab Paris-Saclay, FiME (Laboratoire de Finance des Marchés de l'Énergie)

November 7, 2024

Rahimi and Recht [31] introduced the idea of decomposing shift-invariant kernels by randomly sampling from their spectral distribution. This famous technique, known as Random Fourier Features (RFF), is in principle applicable to any shift-invariant kernel whose spectral distribution can be identified and simulated. In practice, however, it is usually applied to the Gaussian kernel because of its simplicity, since its spectral distribution is also Gaussian. Clearly, simple spectral sampling formulas would be desirable for broader classes of kernel functions. In this paper, we propose to decompose spectral kernel distributions as a scale mixture of α-stable random vectors. This provides a simple and ready-to-use spectral sampling formula for a very large class of multivariate shift-invariant kernels, including exponential power kernels, generalized Matérn kernels, generalized Cauchy kernels, as well as newly introduced kernels such as the Beta, Kummer, and Tricomi kernels. In particular, we show that the spectral densities of all these kernels are scale mixtures of the multivariate Gaussian distribution. This provides a very simple way to modify existing Random Fourier Features software based on Gaussian kernels to cover a much richer class of multivariate kernels. This result has broad applications for support vector machines, kernel ridge regression, Gaussian processes, and other kernel-based machine learning techniques for which the random Fourier features technique is applicable.

Keywords: random Fourier features, isotropic kernels, spectral distribution, stable distribution, generalized Matérn kernel, generalized Cauchy kernel, Tricomi kernel, spectral Monte Carlo, Gaussian processes.

*Corresponding author, nicolaslangrene@uic.edu.cn

1 Introduction

We start by fixing some notation and terminology regarding kernel functions. Let $\tilde{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel function. It is said to be *shift-invariant* (a.k.a. translation-invariant, radially-symmetric, or stationary) if for any $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{x}_j \in \mathbb{R}^d$, $\tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_i - \mathbf{x}_j)$ only depends on $\mathbf{x}_i$ and $\mathbf{x}_j$ through the difference $\mathbf{x}_i - \mathbf{x}_j$. Moreover, the shift-invariant kernel $K : \mathbb{R}^d \to \mathbb{R}$ is said to be *isotropic* [12] if it only depends on $\mathbf{x}_i$ and $\mathbf{x}_j$ through the Euclidean norm $\|\mathbf{x}_i - \mathbf{x}_j\|$ of the difference $\mathbf{x}_i - \mathbf{x}_j$.

Next, we say that the shift-invariant kernel $K : \mathbb{R}^d \to \mathbb{R}$ is *positive definite* if for any $N \geq 1$, $(\mathbf{x}_1, \dots, \mathbf{x}_N) \in \mathbb{R}^{d \times N}$ and $(z_1, \dots, z_N) \in \mathbb{R}^N$,

$$\sum_{i=1}^{N} \sum_{j=1}^{N} z_i z_j K(\mathbf{x}_i - \mathbf{x}_j) \geq 0. \tag{1}$$

According to Bochner's theorem [1, 2], a continuous, shift-invariant kernel $K : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if there exists a finite measure μ such that

$$K(\mathbf{u}) = \int_{\mathbb{R}^d} \exp(i\mathbf{x}^\top \mathbf{u}) d\mu(\mathbf{x}), \quad \mathbf{u} \in \mathbb{R}^d.$$

In other words, K is proportional to a characteristic function. If the kernel is scaled such that $K(\mathbf{0}) = 1$, then μ is a probability measure. Suppose that it admits a density f. Then

$$K(\mathbf{u}) = K(\mathbf{0}) \int_{\mathbb{R}^d} \exp(i\mathbf{x}^\top \mathbf{u}) f(\mathbf{x}) d\mathbf{x}, \quad \mathbf{u} \in \mathbb{R}^d. \tag{2}$$

This shows that K is the multivariate Fourier transform of f. According to Bochner's theorem, f is nonnegative if and only if K is positive definite. In other words, f is a probability density function, known as the *spectral density* of K, if and only if K is positive definite.

In general, equation (2) is complex-valued. Since a characteristic function is real if and only if f is symmetric around zero [9, Lemma 1 page 499], we can further explicit equation (2) by assuming that $f(\mathbf{x}) = f(-\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$, in which case $K(\mathbf{u}) = K(-\mathbf{u})$ for all $\mathbf{u} \in \mathbb{R}^d$ and

$$K(\mathbf{u}) = K(\mathbf{0}) \int_{\mathbb{R}^d} \cos(\mathbf{x}^\top \mathbf{u}) f(\mathbf{x}) d\mathbf{x}, \quad \mathbf{u} \in \mathbb{R}^d \tag{3}$$

Moreover, K and f are Fourier duals:

$$f(\mathbf{x}) = \frac{1}{K(\mathbf{0})(2\pi)^d} \int_{\mathbb{R}^d} \cos(\mathbf{x}^\top \mathbf{u}) K(\mathbf{u}) d\mathbf{u}, \quad \mathbf{x} \in \mathbb{R}^d \tag{4}$$

The interesting aspect of equation (3) is that if K is positive definite, then f is a probability density function, and the following probabilistic representation holds

$$K(\mathbf{u}) = K(\mathbf{0}) \mathbb{E}\left[\cos(\boldsymbol{\eta}^\top \mathbf{u})\right] \tag{5}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)$ is a continuous random vector with density f. The random vector $\boldsymbol{\eta}$ is known as *random projection*. The probabilistic representation (5) suggests that the kernel K can be approximated by Monte Carlo simulations:

$$K(\mathbf{u}) \simeq K_M(\mathbf{u}) := \frac{K(\mathbf{0})}{M} \sum_{m=1}^{M} \cos(\boldsymbol{\eta}_m^\top \mathbf{u}), \quad \mathbf{u} \in \mathbb{R}^d. \tag{6}$$

This approach is known as *random Fourier features* [31]. It provides an explicit feature map to approximate kernel functions in an efficient way, with broad applications in machine learning and statistical learning, for example for algorithms such as kernel ridge regression, Gaussian process inference, kernel principal component analysis, support vector machines, and other kernel-based methods [11].

In order to implement equation (6), the spectral distribution of K needs to be precomputed. Rahimi and Recht [31] provide three examples of shift-invariant multivariate kernels amenable to this method: the Gaussian kernel, the Laplace kernel and the Cauchy kernel:

- The Gaussian kernel admits a Gaussian spectral density

$$K(\mathbf{u}) = \exp\left(-\frac{1}{2}\|\mathbf{u}\|^2\right) \quad , \quad f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right) , \tag{7}$$

- The Laplace kernel admits a Cauchy spectral density

$$K(\mathbf{u}) = \exp\left(-\sum_{\ell=1}^{d} |u_\ell|\right) \quad , \quad f(\mathbf{x}) = \frac{1}{\pi^d} \prod_{\ell=1}^{d} \frac{1}{1 + x_\ell^2} , \tag{8}$$

- The Cauchy kernel admits a Laplace spectral density

$$K(\mathbf{u}) = \prod_{\ell=1}^{d} \frac{1}{1 + u_\ell^2} \quad , \quad f(\mathbf{x}) = \frac{1}{2^d} \exp\left(-\sum_{\ell=1}^{d} |x_\ell|\right) \tag{9}$$

where $\mathbf{u} = (u_1, \ldots, u_d) \in \mathbb{R}^d$ and $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Remark that the two kernels (8)-(9) are not isotropic kernels but tensor kernels, obtained as the product of univariate kernels; for the Gaussian kernel (7), the isotropic and tensor formulations coincide. Remark also how the roles of K and f can be swapped: if K is a nonnegative integrable kernel with spectral density f, then $f/f(\mathbf{0})$ is a kernel with spectral density $K/((2\pi)^d f(\mathbf{0}))$. This is a particular case of the duality theorem in Harrar et al. [14].

In principle, the random Fourier features formula (6) can be applied to any shift-invariant kernel for which the spectral density can be computed and simulated. In practice, the simplicity and convenience of the three examples (7)-(8)-(9), especially the Gaussian kernel (7), means that there is not much work available on the application of random Fourier features to other multivariate parametric kernels. One clear challenge is the ability to compute the multivariate inverse Fourier transform (4) analytically, and to find a suitable sampling algorithm for the corresponding distribution. This is the task we propose to address in this paper. We will show how the random projection technique can be explicitly extended to a large class of shift-invariant kernels of interest in machine learning, including Matérn kernels, exponential power kernels, and generalized Cauchy kernels. To do this, we make use of results from the theory of stable distributions [33, 6, 24]. We show how the spectral distributions of all these kernels of interest can be expressed as a scale mixture of a multivariate stable distribution, and provide explicit formulas to simulate them. We then use this approach to create new multivariate isotropic kernels to which the random Fourier features methodology is readily applicable. Examples of newly introduced kernels include the Beta, Kummer and Tricomi kernels. The availability of new kernel functions with additional parameters is of great practical interest, especially in situations where classical parametric kernels are too rigid to properly capture the information available in the data. We illustrate these kernels along with their random Fourier features approximations.

The rest of the paper is organized as follows. Section 2 describes our decomposition of random projections, and provides examples of the resulting multivariate isotropic kernels. Section 3 describes how to sample such random projections. Section 4 provides numerical examples, and finally Section 5 concludes the paper.

2 New random projections

In this section, we describe how to decompose random projections using symmetric stable vectors (Definition 1) so as to generate a large class of multivariate isotropic kernels (Theorem 1) and easily create new ones (Table 1).

To describe our representation of multivariate random projections, we need the following definition.

Definition 1. For any $\alpha \in (0, 2]$, let $\boldsymbol{S}_\alpha$ be a d-dimensional random vector with characteristic function ϕ_α given by

$$\phi_\alpha(\mathbf{u}) = \mathbb{E}\left[e^{i\boldsymbol{S}_\alpha^\top \mathbf{u}}\right] = e^{-\|\mathbf{u}\|^\alpha}\ , \ \ \mathbf{u} \in \mathbb{R}^d \tag{10}$$

where $\|\mathbf{u}\|$ is the Euclidean norm of $\mathbf{u} \in \mathbb{R}^d$. The vector $\boldsymbol{S}_\alpha$ is called *symmetric stable* (see for example [33]).

Theorem 1. *For any $\alpha \in (0, 2]$, let $\boldsymbol{S}_\alpha$ be a d-dimensional symmetric stable random vector (Definition 1), let R be a real-valued nonnegative random variable with characteristic function ϕ, and let $\lambda > 0$ be a positive constant. Then, the random projection vector defined by*

$$\boldsymbol{\eta} = (\lambda R)^{\frac{1}{\alpha}} \boldsymbol{S}_\alpha \tag{11}$$

spans the following isotropic kernel $K : \mathbb{R}^d \to \mathbb{R}$:

$$K(\mathbf{u}) = \mathbb{E}\left[e^{i\boldsymbol{\eta}^\top \mathbf{u}}\right] = \phi(i\lambda \|\mathbf{u}\|^\alpha)\ , \ \ \mathbf{u} \in \mathbb{R}^d \tag{12}$$

Proof. The characteristic function of $\boldsymbol{\eta}$ is given by

$$\begin{aligned}
\mathbb{E}\left[e^{i\boldsymbol{\eta}^\top \mathbf{u}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(i\boldsymbol{S}_\alpha^\top\left((\lambda R)^{\frac{1}{\alpha}}\mathbf{u}\right)\right)|R\right]\right] \\
&= \mathbb{E}\left[\exp\left(-\lambda R \|\mathbf{u}\|^\alpha\right)\right] \\
&= \mathbb{E}\left[\exp\left(iR\left(i\lambda \|\mathbf{u}\|^\alpha\right)\right)\right] \\
&= \phi(i\lambda \|\mathbf{u}\|^\alpha)
\end{aligned}$$

which proves equation (12). □

Remark 1. The nonnegative random variable R in Theorem 1 can equivalently be characterized by its Laplace-Stieltjes transform $\mathcal{L}$. The resulting kernel (12) becomes $K(\mathbf{u}) = \mathcal{L}(\lambda \|\mathbf{u}\|^{\alpha})$. The two characterizations are equivalent, since $\mathcal{L}(x) = \phi(ix)$ for a positive-valued random variable. The Laplace-Stieltjes version of Theorem 1 has been used explicitly in [28] in the univariate case. In this paper, we favor the characteristic function formulation of Theorem 1, since these functions are much more commonly precomputed and available for a vast range of random variables [25, 26], making it easy to construct new positive-definite multivariate isotropic kernels, as done in Table 1.

Theorem 1 provides a very convenient way to generate multivariate isotropic kernels. Table 1 provides several examples of interest. Table 1 makes use of the following special functions:

- Γ is the gamma function [7, 5.2],
- $\mathcal{B}(a, b)$ is the beta function [7, 5.12],
- $\mathcal{K}_{\beta}$ is the modified Bessel function [7, 10.25],
- $\mathcal{M}(a, b, z)$ is the Kummer confluent hypergeometric function [7, 13.2], also denoted as ${}_1F_1(a, b, z)$ in some references,
- $\mathcal{U}(a, b, z)$ is the Tricomi confluent hypergeometric function, a.k.a. Kummer's function of the second kind [7, 13.2].

Table 1 also uses the following random variables:

- $\boldsymbol{N}$ is a standard multivariate Gaussian random variable,
- G_{β}, $\beta > 0$, is a Gamma random variable, with density $f(x) = \frac{1}{\Gamma(\beta)} x^{\beta-1} e^{-x}$, $x > 0$,
- $B_{\beta,\gamma}$, $\beta > 0$, $\gamma > 0$, is a Beta random variable, with density $f(x) = \frac{x^{\beta-1}(1-x)^{\gamma-1}}{\mathcal{B}(\beta,\gamma)}$, $x \in (0, 1)$. It can be obtained from two independent Gamma random variables G_{β} and G_{γ} as $B_{\beta,\gamma} \stackrel{d}{=} \frac{G_{\beta}}{G_{\beta}+G_{\gamma}}$,
- $F_{2\beta,2\gamma}$, $\beta > 0$, $\gamma > 0$, is a Fisher-Snedecor random variable, with density $f(x) = \frac{1}{x\mathcal{B}(\beta,\gamma)} \frac{(\beta x)^{\beta}\gamma^{\gamma}}{(\beta x+\gamma)^{\beta+\gamma}}$, $x > 0$. It can be obtained from two independent Gamma random variables G_{β} and G_{γ} as $F_{2\beta,2\gamma} \stackrel{d}{=} \frac{\gamma G_{\beta}}{\beta G_{\gamma}}$.

	Name	Formula	
$R =$	Constant 1	$\phi(x) = e^{ix}$	
$K =$	Exponential power	$K(\mathbf{u}) = e^{-\|\mathbf{u}\|^\alpha}$	$\lambda = 1$
$R =$	Gamma G_β , $\beta > 0$	$\phi(x) = (1 - ix)^{-\beta}$	
$K =$	Generalized Cauchy	$K(\mathbf{u}) = \frac{1}{\left(1+\frac{\|\mathbf{u}\|^\alpha}{2\beta}\right)^\beta}$	$\lambda = \frac{1}{2\beta}$
$R =$	Inverse Gamma $1/G_\beta$, $\beta > 0$	$\phi(x) = \frac{2(-ix)^{\beta/2}}{\Gamma(\beta)} \mathcal{K}_\beta(\sqrt{-4ix})$	
$K =$	Generalized Matérn	$K(\mathbf{u}) = \frac{(\sqrt{2\beta}\|\mathbf{u}\|^{\frac{\alpha}{2}})^\beta}{\Gamma(\beta)2^{\beta-1}} \mathcal{K}_\beta(\sqrt{2\beta}\,\|\mathbf{u}\|^{\frac{\alpha}{2}})$	$\lambda = \frac{\beta}{2}$
$R =$	Beta $B_{\beta,\gamma}$, $\beta > 0$, $\gamma > 0$	$\phi(x) = \mathcal{M}(\beta, \beta + \gamma, ix)$	
$K =$	Kummer	$K(\mathbf{u}) = \mathcal{M}(\beta, \beta + \gamma, -\|\mathbf{u}\|^\alpha)$	$\lambda = 1$
$R =$	Beta-exponential $-\log(B_{\beta,\gamma})$, $\beta > 0$, $\gamma > 0$	$\phi(x) = \frac{\mathcal{B}(\beta-ix,\gamma)}{\mathcal{B}(\beta,\gamma)}$	
$K =$	Beta	$K(\mathbf{u}) = \frac{\mathcal{B}(\beta+\|\mathbf{u}\|^\alpha,\gamma)}{\mathcal{B}(\beta,\gamma)}$	$\lambda = 1$
$R =$	F-distribution $F_{2\beta,2\gamma}$, $\beta > 0, \gamma > 0$	$\phi(x) = \frac{\Gamma(\beta+\gamma)}{\Gamma(\gamma)} \mathcal{U}\left(\beta, 1-\gamma, -i\frac{\gamma}{\beta}x\right)$	
$K =$	Tricomi	$K(\mathbf{u}) = \frac{\Gamma(\beta+\gamma)}{\Gamma(\gamma)} \mathcal{U}\left(\beta, 1-\gamma, \frac{\gamma}{\beta}\|\mathbf{u}\|^\alpha\right)$	$\lambda = 1$

Table 1: Characteristic functions of random radii $R \geq 0$ and resulting covariance kernels K. The notations R, K and λ are defined in Theorem 1.

Particular cases of covariance kernels of interest from Table 1 include:

- The Laplace kernel $K(\mathbf{u}) = e^{-\|\mathbf{u}\|}$ ($R = 1$, $\lambda = 1$, $\alpha = 1$), also known as exponential kernel [32] (see for example [11, equation (4)]). The random projection $\boldsymbol{\eta} = \boldsymbol{S}_1$ follows a standard multivariate Cauchy distribution [6].
- The Gaussian kernel $K(\mathbf{u}) = e^{-\frac{\|\mathbf{u}\|^2}{2}}$ ($R = 1$, $\lambda = 1/2$, $\alpha = 2$), also known as squared exponential kernel [32]. The random projection $\boldsymbol{\eta} = \boldsymbol{S}_2(1/2)^{\frac{1}{2}} = \boldsymbol{N}$ follows a standard multivariate Gaussian distribution.
- The exponential power kernel $K(\mathbf{u}) = e^{-\|\mathbf{u}\|^\alpha}$ ($R = 1$, $\lambda = 1$, $\alpha \in (0, 2]$) [3, (21.4)],

also known as generalized Gaussian [8], generalized normal [29], γ-exponential [32] or Subbotin [35] kernel. The random projection $\boldsymbol{\eta} = \boldsymbol{S}_\alpha$ follows a symmetric stable distribution ([5], [6]).

- The Matérn-ν kernel $K(\mathbf{u}) = \frac{(\sqrt{2\nu}\|\mathbf{u}\|)^\nu}{\Gamma(\nu)2^{\nu-1}}\mathcal{K}_\nu(\sqrt{2\nu}\,\|\mathbf{u}\|)$, $\nu > 0$ ($R = 1/G_\nu$, $\lambda = \frac{\nu}{2}$, $\alpha = 2$) [30], also known as multivariate symmetric Laplace kernel [17, equation (5.2.2)]. The random projection $\boldsymbol{\eta} = \boldsymbol{S}_2(\nu/(2G_\nu))^{\frac{1}{2}} = \boldsymbol{N}/\sqrt{G_\nu/\nu}$ follows a standard multivariate Student $t_{2\nu}$ distribution with 2ν degrees of freedom [18, equation (1.2)].

- The power kernel $K(\mathbf{u}) = \frac{1}{1+\|\mathbf{u}\|^\alpha}$ ($R = G_1$, $\lambda = 1$, $\alpha \in (0, 2]$). The random projection $\boldsymbol{\eta} = \boldsymbol{S}_\alpha E^{\frac{1}{\alpha}}$, where E is a standard exponential random variable, follows a Linnik distribution [22], [20], [4], also known as Linnik-Laha distribution.

- The Student t kernel $K(\mathbf{u}) = \left(1 + \frac{\|\mathbf{u}\|^2}{2\beta}\right)^{-\beta}$, $\beta > 0$, with $2\beta - 1$ degrees of freedom ($R = G_\beta$, $\lambda = \frac{1}{2\beta}$, $\alpha = 2$), also known as rational quadratic kernel [32] or generalized inverse multiquadric kernel [15]. The random projection $\boldsymbol{\eta} = \boldsymbol{S}_2(G_\beta/(2\beta))^{\frac{1}{2}} = \boldsymbol{N}\sqrt{\frac{G_\beta}{\beta}}$ follows a Matérn distribution, also known as generalized Laplace distribution [17, Definition 4.1.1] (see [17, Proposition 4.1.2]), which is a particular case of the variance-gamma distribution [10].

- The generalized Cauchy kernel $K(\mathbf{u}) = \frac{1}{(1+\|\mathbf{u}\|^\alpha)^\beta}$, $\beta > 0$ ($R = G_\beta$, $\lambda = 1$, $\alpha \in (0, 2]$) [13], also known as generalized Pearson VII kernel [3, page 144]. The random projection $\boldsymbol{\eta} = \boldsymbol{S}_\alpha G_\beta^{\frac{1}{\alpha}}$ follows a generalized Linnik distribution [5, 6]. The scaling $\lambda = \frac{1}{2\beta}$ proposed in Table 1 (by analogy with Matérn kernels) is such that $K(\mathbf{u}) = \frac{1}{\left(1+\frac{\|\mathbf{u}\|^\alpha}{2\beta}\right)^\beta} \underset{\beta\to\infty}{\longrightarrow} e^{-\frac{\|\mathbf{u}\|^\alpha}{2}}$, which is an exponential power kernel (with scaling $\lambda = 1/2$) which contains the Gaussian kernel as the particular case $\alpha = 2$.

- Since $F_{2\beta,2\gamma} \overset{d}{=} (G_\beta/\beta)/(G_\gamma/\gamma)$ and $\lim_{a\to\infty} \frac{G_a}{a} = 1$, the proposed Tricomi kernel $K(\mathbf{u}) = \frac{\Gamma(\beta+\gamma)}{\Gamma(\gamma)}\mathcal{U}\left(\beta, 1-\gamma, \frac{\gamma}{\beta}\,\|\mathbf{u}\|^\alpha\right)$ contains generalized Matérn kernels (when $\beta \to \infty$) and generalized Cauchy kernels (when $\gamma \to \infty$) as limit cases, and therefore contains all the classical stationary kernels (Laplace, Gaussian, Matérn, Student, Power, Exponential Power, etc.) as particular limit cases.

Remark 2. Symmetric stable random vectors (Definition 1) only exist when $\alpha \in (0, 2]$. As a result, for every kernel in Table 1, the parameter α in restricted to the interval $(0, 2]$ to enforce positive definiteness. When setting α to a value larger than 2, the inverse Fourier transform (4) of the kernel is not a density anymore as it can take negative values. Explicit examples from the literature include the exponential power kernel with $\alpha = 4$ and $\alpha = 6$ [16], and $\alpha = 3$ [8], as well as the generalized Cauchy kernel with $\alpha = 4$ (Laha distribution [19], its inverse Fourier transform is the Silverman kernel [36]) and $\alpha = 6$ [16].

Remark 3. This section focuses on isotropic kernels, which can be written as $K(\mathbf{u}) = k(\|\mathbf{u}\|)$ where k is a univariate kernel and $\|\mathbf{u}\|$ is the Euclidean norm of the vector

$\mathbf{u} \in \mathbb{R}^d$. Another classical way to construct multivariate kernels is the tensor approach, where $K(\mathbf{u}) = \prod_{\ell=1}^d k(u_\ell)$ is the product of univariate kernels, as shown with the examples (8) and (9) from Rahimi and Recht [31]. The random projection $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)$ of a tensor kernel is a vector with i.i.d. components with distribution equal to the spectral distribution of the univariate kernel k. In view of this, the results from this section, in particular Table 1, can also be used to generate new tensor kernels, by setting $d = 1$ in equation (11) and simulating the resulting univariate random projection $\eta = S_\alpha(\lambda R)^{\frac{1}{\alpha}}$ d times independently. However, the results of this paper suggest that there is little reason to favour tensor kernels over isotropic kernels when resorting to the random Fourier feature approximation, if only because simulating the spectral distribution of isotropic kernels is faster than for tensor kernels, as shown in Remark 5 in the next section.

Remark 4. Let $\Sigma \in \mathbb{R}^{d\times d}$ be a symmetric, positive definite matrix. By noting that $\|\mathbf{u}\|^\alpha = (\mathbf{u}^\top\mathbf{u})^{\frac{\alpha}{2}}$, one can replace $\|\mathbf{u}\|^\alpha$ by $(\mathbf{u}^\top\Sigma\mathbf{u})^{\frac{\alpha}{2}}$ in the definition of all the kernels from Table 1, by changing the random projection formula (11) to $\boldsymbol{\eta} = (\lambda R)^{\frac{1}{\alpha}}\boldsymbol{S}_\alpha\Sigma^{\frac{1}{2}}$. This is another possible approach to introduce additional parameters into the formula of a multivariate parametric kernel.

3 Sampling random projections

In order to simulate the random projections from Table 1, we need to be able to simulate multivariate symmetric stable random vectors $\boldsymbol{S}_\alpha$, $\alpha \in (0, 2]$, as well as the nonnegative random variables R.

Gamma random variables All the nonnegative random variables R in Table 1 can be obtained from independent simulations of Gamma random variables, which are elementary distributions for which simulation routines are widely available. Popular approaches to simulate Gamma random variables include acceptance-rejection and numerical inversion, see Luengo [23].

Symmetric stable random vectors The following Proposition describes how to obtain simulations of symmetric stable random vectors, in the form of scale mixtures of multivariate Gaussian random vectors.

Proposition 1. *Let $\boldsymbol{N}$ be a d-dimensional standard Gaussian vector, let U_1 and U_2 be two independent standard uniform random variables, independent of $\boldsymbol{N}$, let $W = -\log(U_1)$ be a standard exponential random variable, and let $\Theta = \pi\left(U_2 - \frac{1}{2}\right)$ be a uniform random variable in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Then, for any $\alpha \in (0, 2]$, the multivariate symmetric stable distribution $\boldsymbol{S}_\alpha$ (Definition 1) admits the following decomposition*

$$\boldsymbol{S}_\alpha \overset{d}{=} \sqrt{2A_\alpha}\boldsymbol{N} \tag{13}$$

where

$$A_\alpha := \frac{\sin\left(\frac{\alpha\pi}{4} + \frac{\alpha}{2}\Theta\right)}{(\cos(\Theta))^{2/\alpha}}\left(\frac{\cos\left(\frac{\alpha\pi}{4} + \left(\frac{\alpha}{2} - 1\right)\Theta\right)}{W}\right)^{\frac{2}{\alpha}-1} \tag{14}$$

Proof. To obtain the Gaussian mixture representation (13)-(14) of symmetric stable vectors, apply [33, Proposition 2.5.2], use the fact that $\boldsymbol{S}_2 \sim \mathcal{N}(\mathbf{0}, 2\mathbf{I}_d)$ is a multivariate Gaussian random vector with independent components with mean zero and variance 2 (where $\mathbf{I}_d$ is the d-dimensional identity matrix), and use the simulation formula for nonsymmetric stable random variable from [24, Theorem 1.3] with maximum skew $\beta = 1$. □

Corollary 1. *A consequence of Theorem 1 ($\boldsymbol{\eta} = (\lambda R)^{\frac{1}{\alpha}} \boldsymbol{S}_\alpha$) and Proposition 1 ($\boldsymbol{S}_\alpha = \sqrt{2A_\alpha}\boldsymbol{N}$) is that all the kernels discussed in this Section (Table 1) admit a representation of their random projections as a scale mixture of Gaussians, explicitly given by*

$$\boldsymbol{\eta} = \left((\lambda R)^{\frac{1}{\alpha}} \sqrt{2A_\alpha} \right) \boldsymbol{N}. \tag{15}$$

This corollary has important practical applications. For example, it suggests that a task such as kernel learning via learning spectral distributions can be brought down to learning the distribution of the univariate nonnegative random radius R. It can also suggest variance reduction techniques by splitting the effort between the Gaussian vector $\boldsymbol{N}$ [27] and the random scaling factor $(\lambda R)^{\frac{1}{\alpha}} \sqrt{2A_\alpha}$.

Remark 5. In view of equation (15), simulating the random projection vector of an isotropic kernel requires to simulate $d + 2$ independent random variables (the vector $\boldsymbol{N}$ of size d, the random variable A_α and the random variable R). By contrast, simulating the random projection vector of a tensor kernel requires to simulate $3d$ independent random variables (d independent simulations of the scalar random variable $\eta = \left((\lambda R)^{\frac{1}{\alpha}} \sqrt{2A_\alpha} \right) N$ which involves the three random variables N, A_α and R). This shows that the isotropic formulation of multivariate kernels is almost three times more efficient than the tensor formulation when using the random Fourier features approach with the decomposition formula (15).

4 Numerical examples

This section gives several examples of kernels from Table (1) along with their random Fourier features approximation (6), using the random projection decomposition (11) and the simulation algorithms described above.

Remark that all the known theoretical results about the convergence of the random Fourier features approximation with respect to the number of random projections, such as [34], [21] or [37], still apply to the kernels discussed in this paper.

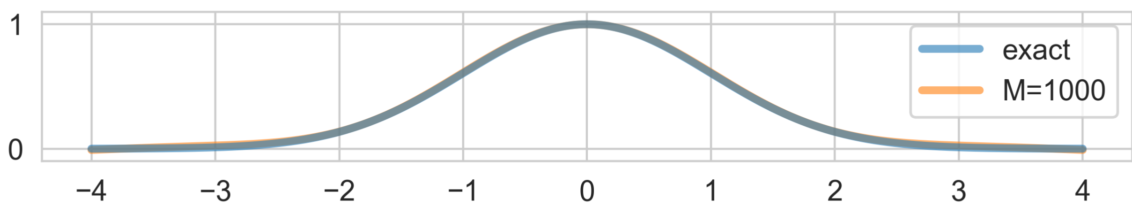

Figure 1: Univariate Gaussian kernel and its random Fourier features approximation (6) using $M = 1000$ random projections with Gaussian distribution.

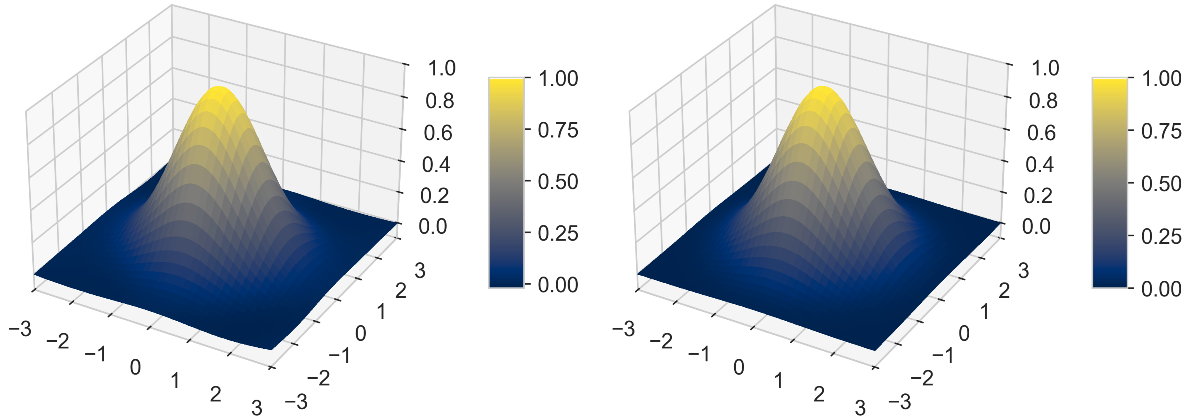

Figure 2: Bivariate Gaussian kernel (right) and its random Fourier features approximation (6) (left) using $M = 4000$ random projections.

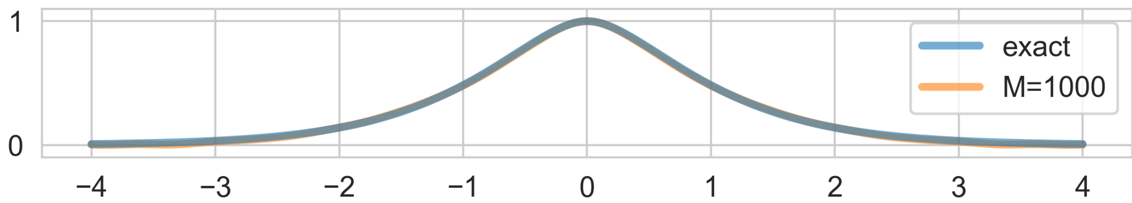

Figure 3: Univariate Matérn-3/2 kernel and its random Fourier features approximation (6) using $M = 1000$ random projections.

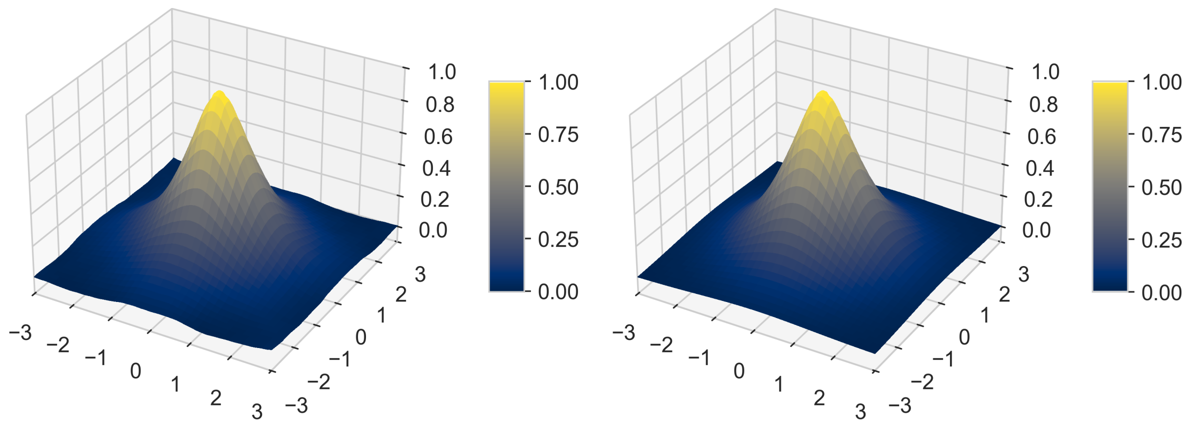

Figure 4: Bivariate Matérn-3/2 kernel (right) and its random Fourier features approximation (6) (left) using $M = 4000$ random projections.

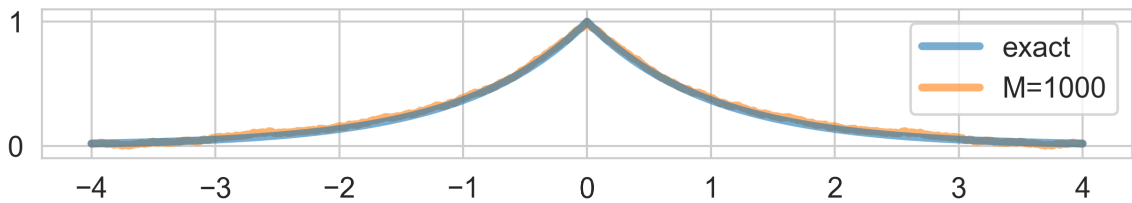

Figure 5: Univariate Laplace (a.k.a. Matérn-1/2) kernel and its random Fourier features approximation (6) using $M = 1000$ random projections.

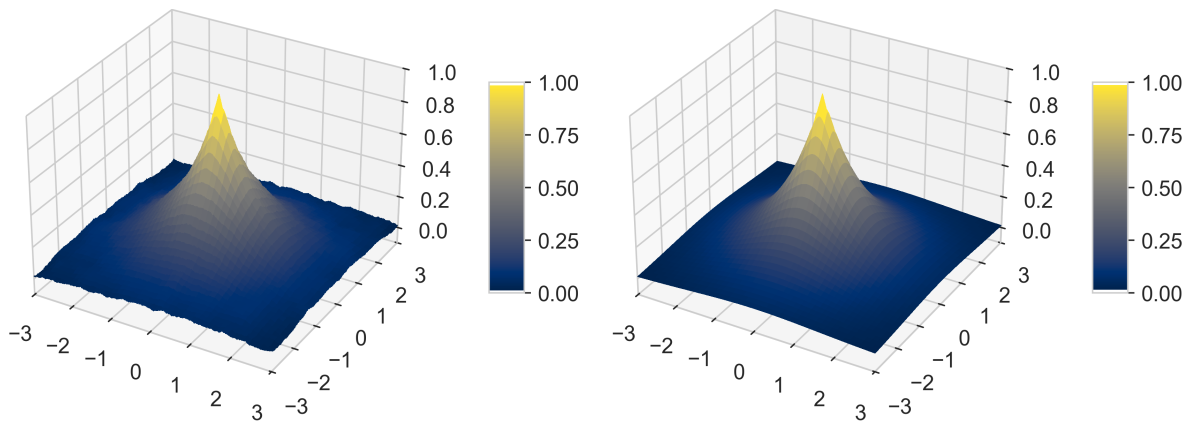

Figure 6: Bivariate Laplace (a.k.a. Matérn-1/2) kernel (right) and its random Fourier features approximation (6) (left) using $M = 4000$ random projections.

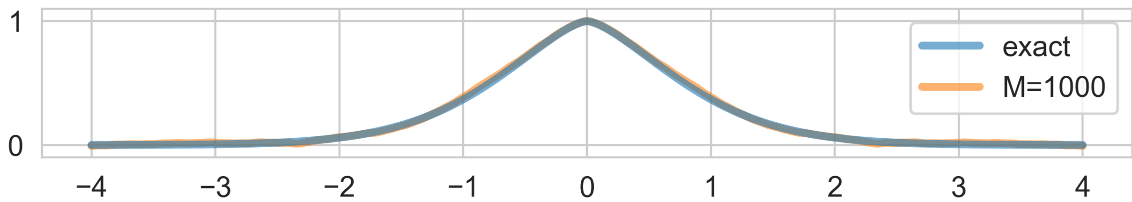

Figure 7: Univariate exponential power kernel with $\alpha = 1.5$ and its random Fourier features approximation (6) using $M = 1000$ random projections.

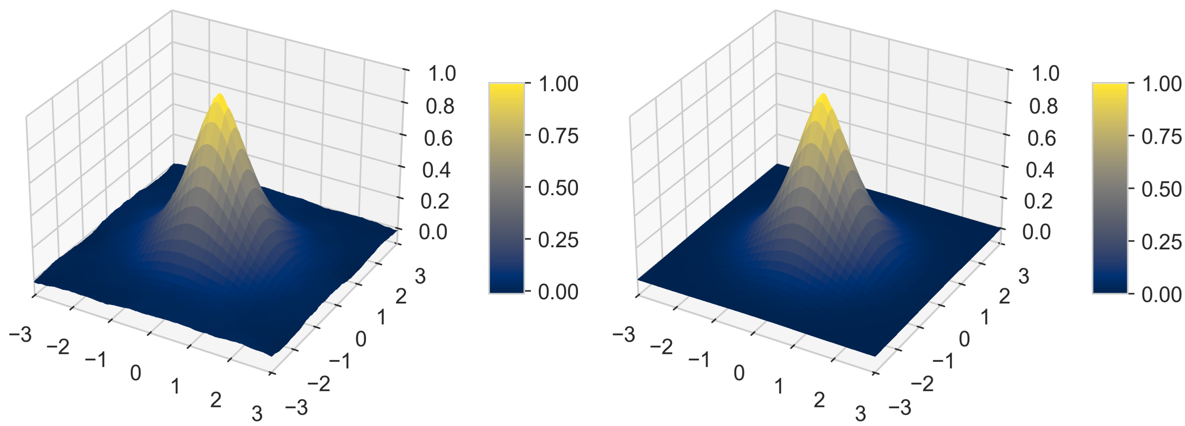

Figure 8: Bivariate exponential power kernel with $\alpha = 1.5$ (right) and its random Fourier features approximation (6) (left) using $M = 4000$ random projections.

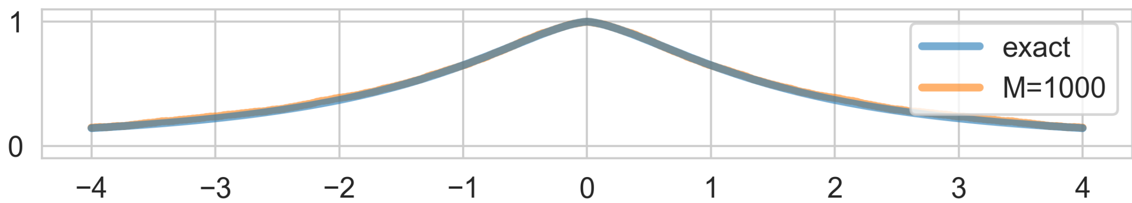

Figure 9: Univariate generalized Cauchy kernel with $(\alpha, \beta) = (1.5, 1.5)$ and its random Fourier features approximation (6) using $M = 1000$ random projections.

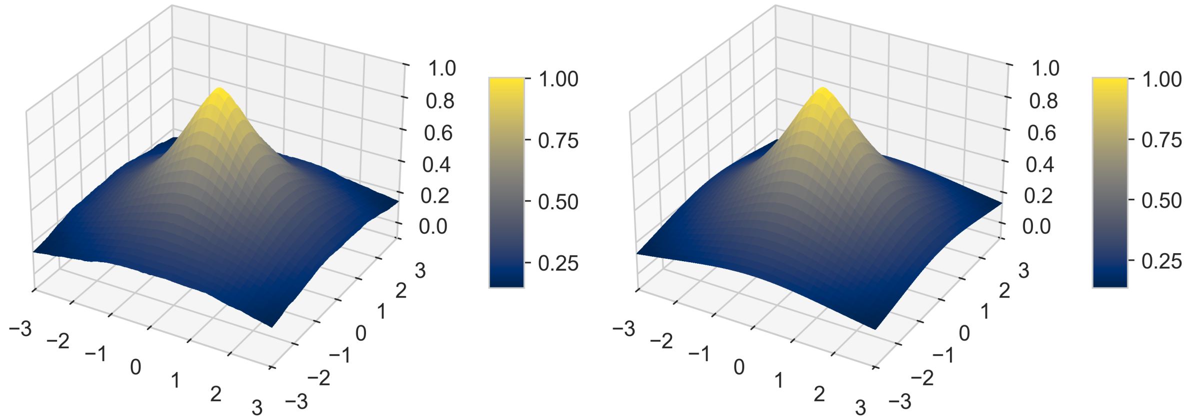

Figure 10: Bivariate generalized Cauchy power kernel with $(\alpha, \beta) = (1.5, 1.5)$ (right) and its random Fourier features approximation (6) (left) using $M = 4000$ random projections.

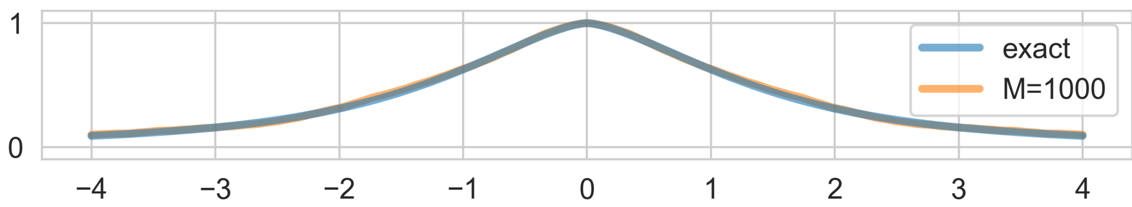

Figure 11: Univariate Kummer kernel with $(\alpha, \beta, \gamma) = (1.5, 1.5, 1.5)$ and its random Fourier features approximation (6) using $M = 1000$ random projections.

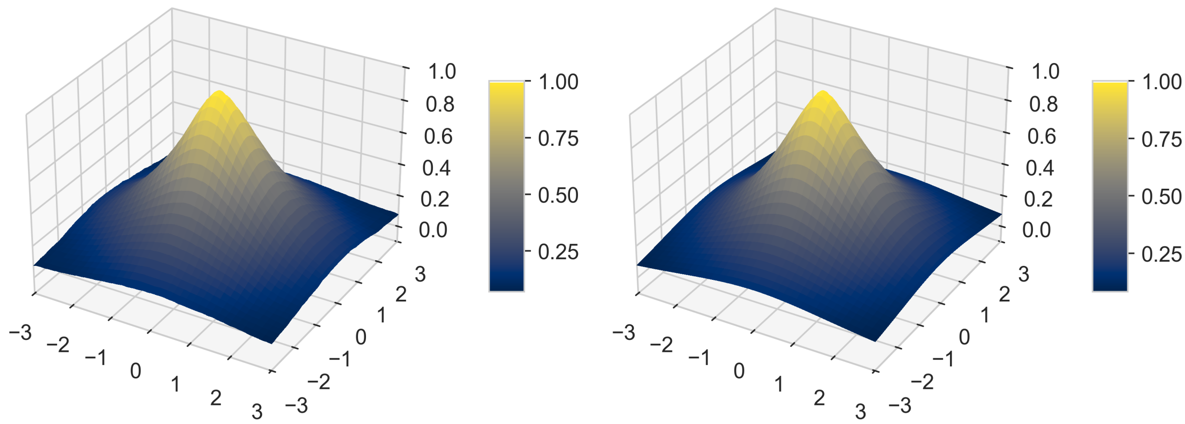

Figure 12: Bivariate Kummer kernel with $(\alpha, \beta, \gamma) = (1.5, 1.5, 1.5)$ (right) and its random Fourier features approximation (6) (left) using $M = 4000$ random projections.

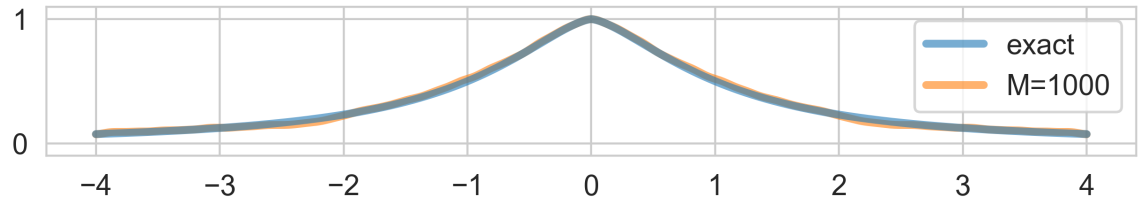

Figure 13: Univariate beta kernel with $(\alpha, \beta, \gamma) = (1.5, 1.5, 1.5)$ and its random Fourier features approximation (6) using $M = 1000$ random projections.

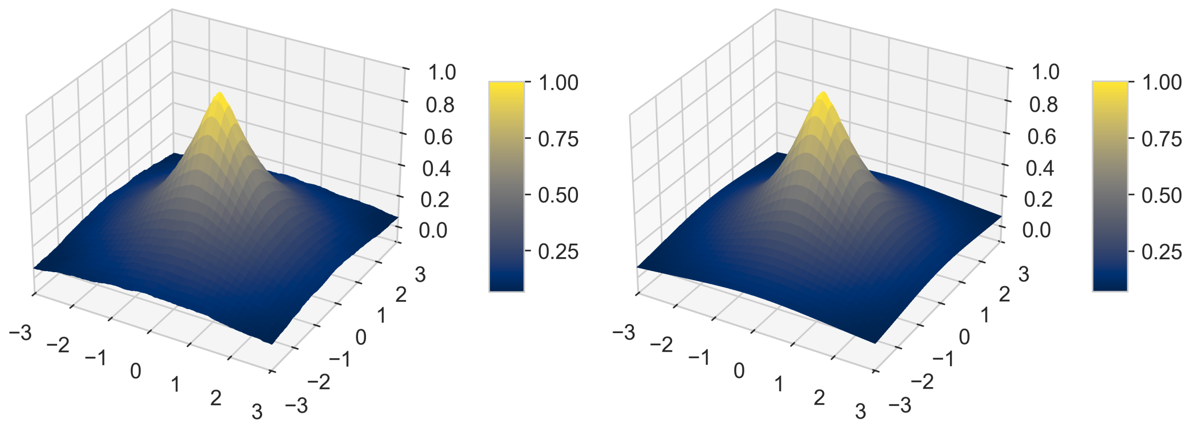

Figure 14: Bivariate beta kernel with $(\alpha, \beta, \gamma) = (1.5, 1.5, 1.5)$ (right) and its random Fourier features approximation (6) (left) using $M = 4000$ random projections.

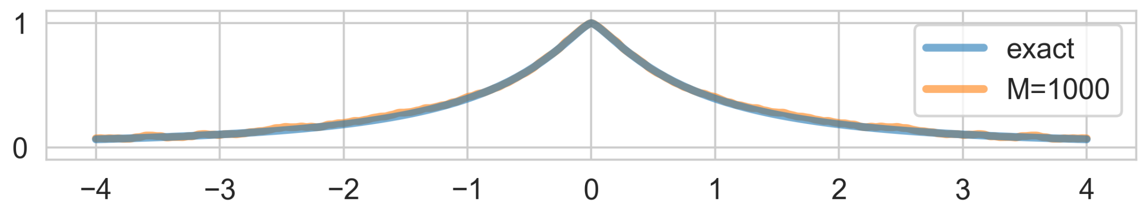

Figure 15: Univariate Tricomi kernel with $(\alpha, \beta, \gamma) = (1.5, 1.5, 1.5)$ and its random Fourier features approximation (6) using $M = 1000$ random projections.

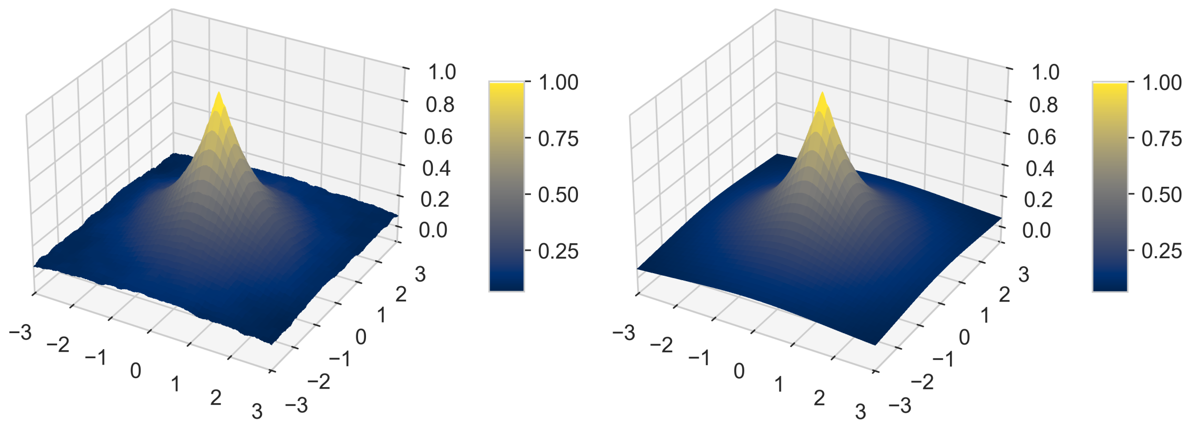

Figure 16: Bivariate Tricomi kernel with $(\alpha, \beta, \gamma) = (1.5, 1.5, 1.5)$ (right) and its random Fourier features approximation (6) (left) using $M = 4000$ random projections.

5 Conclusion

In this paper, we showed that the random projection $\boldsymbol{\eta} = (\lambda R)^{\frac{1}{\alpha}} \boldsymbol{S}_\alpha$, where $\boldsymbol{S}_\alpha$, $\alpha \in (0,2]$ is a symmetric stable random vector, $\lambda > 0$ is a constant, and $R \geq 0$ is a nonnegative random variable with characteristic function ϕ, spans the multivariate isotropic kernel $K(\mathbf{u}) = \mathbb{E}\left[e^{i\boldsymbol{\eta}^\top \mathbf{u}}\right] = \phi(i\lambda \|\mathbf{u}\|^\alpha)$ (Theorem 1). This decomposition applies to the Gaussian, Laplace, Matérn, Generalized Cauchy, and many other kernels of interest in machine learning and statistical learning. We provide a simple sampling algorithm for the vector $\boldsymbol{S}_\alpha$ (Proposition 1). In particular, we show how to decompose it as a scale mixture of a Gaussian vector (equation (13)), which shows that the random projection $\boldsymbol{\eta}$ is also a scale mixture of a Gaussian vector (Corollary 1).

These results provide a very simple way to apply the random Fourier features technique to a very large class of multivariate kernels, not only isotropic kernels but also tensor kernels (Remark 3). We show how the kernel of a Gaussian random Fourier features implementation can be easily modified by simply adding a random scaling to the random projections, and we show how to easily generate new isotropic kernels by changing the distribution of R.

Table 1 provides a number of important examples, including newly proposed kernels such as the Kummer, Beta, and Tricomi kernels. We implemented these random projections for each kernel in Table 1, and illustrated the resulting kernels in dimension 1 and 2. The approximation error and convergence are in line with the known theoretical results about random Fourier features.

Our work provides substantial benefits to the random Fourier features methodology: it simplifies its practical implementation, readily adapts it to a large class of kernels, simplifies the creation of new multivariate positive definite kernels, and simplifies kernel learning to the identification of the distribution of the random radius R. We believe that further developments ought to be achievable at this intersection between the theories of random projections and stable distributions. Future work could include the extension of our proposed random projection decomposition to more general classes of kernels, including asymmetric, non positive definite, and perhaps non-stationary kernels, with a broad impact for kernel-based machine learning techniques.

Acknowledgements

The authors would like to thank Christian Walder (Google Brain Montréal, Australian National University) for fruitful discussions. Nicolas Langrené acknowledges the partial support of the Guangdong Provincial Key Laboratory IRADS (2022B1212010006, R0400001-22) and the UIC Start-up Research Fund UICR0700041-22. Pierre Gruet and Xavier Warin acknowledge support from the FiME Lab.

References

[1] Bochner, S. (1933). Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, 108(1):378–410. (Cited on page 2)

[2] Bochner, S. (1959). *Lectures on Fourier integrals*, volume 42. Princeton University Press. (Cited on page 2)

[3] Crooks, G. E. (2019). *Field guide to continuous probability distributions*. Berkeley Institute for Theoretical Science. (Cited on pages 6 and 7)

[4] Devroye, L. (1990). A note on Linnik's distribution. *Statistics & Probability Letters*, 9(4):305–306. (Cited on page 7)

[5] Devroye, L. (2006). Nonuniform random variate generation. In *Handbooks in Operations Research and Management Science*, volume 13, chapter 4, pages 83–121. Elsevier. (Cited on page 7)

[6] Devroye, L. and James, L. (2014). On simulation and properties of the stable law. *Statistical methods & applications*, 23(3):307–343. (Cited on pages 4, 6, and 7)

[7] DLMF (2024). NIST Digital Library of Mathematical Functions. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. (Cited on page 5)

[8] Dytso, A., Bustin, R., Poor, H. V., and Shamai, S. (2018). Analytical properties of generalized Gaussian distributions. *Journal of Statistical Distributions and Applications*, 5(1):1–40. (Cited on page 7)

[9] Feller, W. (1971). *An introduction to probability theory and its applications*, volume II. John Wiley & Sons. (Cited on page 2)

[10] Fischer, A., Gaunt, R., and Sarantsev, A. (2024). The Variance-Gamma distribution: a review. *Statistical Science*. (Cited on page 7)

[11] Francis, D. P. and Raimond, K. (2021). Major advancements in kernel function approximation. *Artificial Intelligence Review*, 54:843–876. (Cited on pages 3 and 6)

[12] Genton, M. (2001). Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2:299–312. (Cited on page 2)

[13] Gneiting, T. and Schlather, M. (2004). Stochastic models that separate fractal dimension and the Hurst effect. *SIAM Review*, 46(2):269–282. (Cited on page 7)

[14] Harrar, S. W., Seneta, E., and Gupta, A. K. (2006). Duality between matrix variate t and matrix variate VG distributions. *Journal of Multivariate Analysis*, 97(6):1467–1475. (Cited on page 3)

[15] Hu, X.-G., Ho, T.-S., and Rabitz, H. (1998). The collocation method based on a generalized inverse multiquadric basis for bound-state problems. *Computer physics communications*, 113(2-3):168–179. (Cited on page 7)

[16] Jones, M. and Wand, M. (1992). Asymptotic effectiveness of some higher order kernels. *Journal of statistical planning and inference*, 31(1):15–21. (Cited on page 7)

[17] Kotz, S., Kozubowski, T., and Krzysztof, P. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer. (Cited on page 7)

[18] Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press. (Cited on page 7)

[19] Laha, R. (1958). An example of a nonnormal distribution where the quotient follows the Cauchy law. *Proceedings of the National Academy of Sciences*, 44(2):222–223. (Cited on page 7)

[20] Laha, R. G. (1961). On a class of unimodal distributions. *Proceedings of the American Mathematical Society*, 12(2):181–184. (Cited on page 7)

[21] Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. (2021). Towards a unified analysis of random Fourier features. *Journal of Machine Learning Research*, 22(108):1–51. (Cited on page 9)

[22] Linnik, Y. (1953). Linear forms and statistical criteria. *Ukrainian Mathematical Journal*, 5(2):207–243. (Cited on page 7)

[23] Luengo, E. A. (2022). Gamma pseudo random number generators. *ACM Computing Surveys*, 55(4):1–33. (Cited on page 8)

[24] Nolan, J. (2020). *Univariate stable distributions: models for heavy tailed data*, volume 10 of *Springer Series in Operations Research and Financial Engineering*. Springer. (Cited on pages 4 and 9)

[25] Oberhettinger, F. (1973). *Fourier transforms of distributions and their inverses: a collection of tables*, volume 16 of *Probability and Mathematical Statistics*. Academic Press. (Cited on page 5)

[26] Oberhettinger, F. (1990). *Tables of Fourier transforms and Fourier transforms of distributions*. Springer-Verlag. (Cited on page 5)

[27] Pagès, G. and Printems, J. (2003). Optimal quadratic quantization for numerics: the Gaussian case. *Monte Carlo Methods and Applications*, 9(2):135–165. (Cited on page 9)

[28] Pakes, A. (1998). Mixture representations for symmetric generalized Linnik laws. *Statistics and Probability Letters*, 37(3):213–221. (Cited on page 5)

[29] Pogány, T. K. and Nadarajah, S. (2010). On the characteristic function of the generalized normal distribution. *Comptes Rendus. Mathématique*, 348(3-4):203–206. (Cited on page 7)

[30] Porcu, E., Bevilacqua, M., Schaback, R., and Oates, C. (2024). The Matérn model: a journey through statistics, numerical analysis and machine learning. *Statistical Science*, 39(3):469–492. (Cited on page 7)

[31] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184. (Cited on pages 1, 3, and 8)

[32] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press. (Cited on pages 6 and 7)

[33] Samorodnitsky, G. and Taqqu, M. (1994). *Stable non-Gaussian processes: stochastic models with infinite variance*. Stochastic modeling. Chapman & Hall. (Cited on pages 4 and 9)

[34] Sriperumbudur, B. and Szabó, Z. (2015). Optimal rates for random Fourier features. *Advances in Neural Information Processing Systems*, 28:1144–1152. (Cited on page 9)

[35] Subbotin, M. F. (1923). On the law of frequency of error. *Matematicheskiĭ Sbornik*, 31(2):296–301. (Cited on page 7)

[36] Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer. (Cited on page 7)

[37] Yao, J., Erichson, N. B., and Lopes, M. E. (2023). Error estimation for random Fourier features. In *International Conference on Artificial Intelligence and Statistics*, pages 2348–2364. PMLR. (Cited on page 9)