RC25633 (WAT1610-090) October 27, 2016
Mathematics

IBM Research Report

On Bochner's and Polya's Characterizations of Positive-Definite Kernels and the Respective Random Feature Maps

Jie Chen
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 USA

Dehua Cheng, Yan Liu
University of Southern California

Research Division
Almaden – Austin – Beijing – Brazil – Cambridge – Dublin – Haifa – India – Kenya – Melbourne – T.J. Watson – Tokyo – Zurich

On Bochner's and Polya's Characterizations of Positive-Definite Kernels and the Respective Random Feature Maps

Jie Chen* Dehua Cheng† Yan Liu†

October 24, 2016

Abstract

Positive-definite kernel functions are fundamental elements of kernel methods and Gaussian processes. A well-known construction of such functions comes from Bochner's characterization, which connects a positive-definite function with a probability distribution. Another construction, which appears to have attracted less attention, is Polya's criterion that characterizes a subset of these functions. In this paper, we study the latter characterization and derive a number of novel kernels little known previously.

In the context of large-scale kernel machines, Rahimi and Recht (2007) proposed a random feature map (random Fourier) that approximates a kernel function, through independent sampling of the probability distribution in Bochner's characterization. The authors also suggested another feature map (random binning), which, although not explicitly stated, comes from Polya's characterization. We show that with the same number of random samples, the random binning map results in an Euclidean inner product closer to the kernel than does the random Fourier map. The superiority of the random binning map is confirmed empirically through regressions and classifications in the reproducing kernel Hilbert space.

1 Introduction

A positive-definite function (also coined *kernel* in this paper) is a complex-valued function $k : \mathbb{R} \to \mathbb{C}$ such that for any n real numbers $x_1, x_2, \ldots, x_n$, the matrix K with elements $K_{ij} = k(x_i - x_j)$ is positive semi-definite. A well-known relationship between positive-definite functions and probability distributions is given by the celebrated Bochner's theorem [19], which states that a continuous function k with $k(0) = 1$ is positive-definite if and only if it is the characteristic function (cf) of some random variable X. Let $F(x)$ be the cumulative distribution function (cdf) of X; then, the if-and-only-if condition is written in the form of a Stieltjes integral as

$$k(r) = E[e^{\mathbf{i}\, rX}] = \int_{\mathbb{R}} e^{\mathbf{i}\, rx}\, dF(x). \tag{1}$$

A practical significance of this characterization is that one may construct a positive-definite function from a probability distribution. For example, the squared exponential kernel[1] is constructed from

*IBM Thomas J. Watson Research Center. Email: chenjie@us.ibm.com

†University of Southern California. Email: (dehua.cheng, yanliu.cs)@usc.edu

[1]Also called the Gaussian kernel.

the normal distribution, the exponential kernel from the Cauchy distribution[2], and the Cauchy kernel from the Laplace distribution. Positive-definite functions are of vital importance in kernel methods and Gaussian processes. The kernel k in kernel methods defines a reproducing kernel Hilbert space (RKHS) [1], from which an optimal prediction function is sought with respect to some risk functional [18, 9]. In Gaussian processes, k serves as a covariance function and its Fourier transform, coined *spectral density*, dictates the smoothness and other behavior of the process [19, 21, 16, 6].

Another approach for constructing kernels from probability distributions, which appears to have attracted less attention, comes from Polya's criterion [8], which states that for any real continuous and even function k convex on $[0, \infty)$ with $k(0) = 1$ and $k(\infty) = 0$, there exists a random variable X with a positive support and a cdf $F(x)$ such that

$$k(r) = \int_{\mathbb{R}} \max\left\{0, 1 - \frac{|r|}{|x|}\right\} dF(x). \tag{2}$$

An informal argument why k is positive-definite, is that the integrand in (2) is the triangular function, whose Fourier transform is the squared sinc function that is nonnegative. With slightly extra work, in Section 2, we show that the converse of Polya's criterion is also true; that is, given any cdf F with a positive support, the function k defined in (2) possesses the said properties. Hence, (2) in fact characterizes a subset of positive-definite functions, the most salient property being convexly decreasing on $[0, \infty)$; and the respective probability distributions are those positively supported. We study in depth Polya's criterion and its consequences, particularly in the context of kernel constructions, in Section 2. Then, in Section 3, we consider a number of example distributions and derive explicit expressions for k and the associated Fourier transform. Such distributions include Poisson, gamma, Nakagami, Weibull, and other distributions that are special cases of the last three (e.g., exponential, chi-square, chi, half-normal, and Rayleigh).

One may recall that (2) resembles an equality established by Rahimi and Recht [15]

$$k(r) = \int_0^\infty \max\left\{0, 1 - \frac{r}{x}\right\} xk''(x)\, dx, \qquad r \geq 0, \tag{3}$$

for any twice differentiable function k on $(0, \infty)$ that vanishes at the infinity. Indeed, if $xk''(x)$ integrates to unity, it could be considered the probability density function (pdf) associated to F. Two important distinctions, however, should be noted. First, the expression (3) implicitly assumes the existence of a pdf, which occurs only for (well behaved) continuous distributions. To the contrary, (2), in the form of a Stieltjes integral, is more general, defined for distributions including notably discrete ones. Such does not contradict with (3), because a kernel function constructed from a discrete distribution may not be twice differentiable on $(0, \infty)$; in fact, it is at most once differentiable.

Second, (2) and (3) result in methods that utilize the relationship between a kernel function and a probability distribution in a completely opposite direction. The work [15], based on (3), starts from a known kernel k and seeks a valid pdf for constructing random feature maps. On the other hand, our work, based on (2), focuses on constructing new kernels. The theoretical appeal

[2]It turns out that the exponential kernel can be generalized to high dimensions by taking either the 1-norm or the 2-norm of the inputs. In the 1-norm case, the kernel is also called the Laplace kernel and the distribution is a tensor product of one-dimensional Cauchy distributions. In the 2-norm case, the distribution is multivariate Cauchy. More discussions appear in Section 4.3.

of (2) guarantees that the so defined function k is always a valid kernel; and the prosperous results in probability distributions provide a practical opportunity to derive explicit formulas for novel kernels k little known previously.

Whereas the mathematical properties of the proposed kernels are interesting in their own right, the work here stems from a practical purpose: we are interested in comparing the quality of the random feature maps resulting from (1) and (2), if possible, for the same kernel function k. A computational bottleneck in many kernel and Gaussian process applications is the factorization of the $n \times n$ shifted kernel matrix $K + \lambda I$, whose memory cost and time cost scale as $O(n^2)$ and $O(n^3)$, respectively, if no particular structures of K are known other than symmetry. A class of methods aiming at reducing the computational costs is the random feature approaches [15, 11, 20, 13, 23], which map a data point x to a random vector $\boldsymbol{z}(x)$ such that $\boldsymbol{z}(x)^T \boldsymbol{z}(x')$ approximates $k(x - x')$ for any pair of points x and x'. In the matrix form, let $\boldsymbol{z}$ be a column vector and let Z be the matrix $[\boldsymbol{z}(x_1), \boldsymbol{z}(x_2), \ldots, \boldsymbol{z}(x_n)]$; then, $Z^T Z$ approximates K.

Probably the most well-known and used random feature map is random Fourier (see, e.g., the original publication [15], a few extensions [25, 7, 2], analysis [24, 22], and applications [10, 5]); whereas a less popular, but more effective one as we argue in this paper, is random binning (see the same publication [15]). The random Fourier approach uses (1) to construct a dense Z of size $D \times n$, where D denotes the number of random Fourier samples. The random binning approach, as we extend in this work for an arbitrary distribution positively supported, uses (2) to construct a sparse Z where each column has D' nonzeros, with D' denoting the number of random binning samples. We analyze in Section 4 that $Z^T Z$ better approximates K by using the latter approach, if D is taken to be the same as D'. In other words, for a matching approximation quality, D' may be (much) smaller than D. Such an observation supports the use of the proposed formula (2) for kernel construction and approximation.

Note that analysis of the two random feature approaches exists in other works. Rahimi and Recht [15] give probabilistic bounds for the uniform error $\sup_{x,x'} |\boldsymbol{z}(x)^T \boldsymbol{z}(x') - k(x - x')|$. These results, however, do not directly compare the two approaches as we do. Wu et. al [22] consider the setting of risk functional minimization in the RKHS and bound the bias of the computed function from the optimal one, when the minimization is done through coordinate descent by taking one sample at a time. They argue that the optimization converges faster for the random binning approach, in the sense that if the same number of samples/iterations are used, the bias has a smaller upper bound. On the other hand, our analysis focuses on the matrix approximation error and gives exact values rather than bounds. As a result, the analysis also favors the random binning approach. Experimental results that gauge regression and classification performance further confirm the superiority of this approach; see Section 5.

We summarize the contributions of this work and conclude in Section 6.

2 Polya's Characterization

We start with the formal statement of Polya.

Theorem 1 (Polya's criterion)**.** *If $k : \mathbb{R} \to [0, 1]$ is a real, continuous and even function with $k(0) = 1$, $\lim_{r \to \infty} k(r) = 0$, and k is convex on $[0, \infty)$, then there exists a cumulative distribution function $F(x)$ on $(0, \infty)$ such that*

$$k(r) = \int_0^\infty \max\left\{0, 1 - \frac{|r|}{|x|}\right\} dF(x). \tag{4}$$

Hence, k is a characteristic function.

Proof. See, e.g., proof of Theorem 3.3.10 in [8]. □

Polya's criterion sheds deep insights between a kernel k and a cdf F connected by the relation (4). Let us stress a few.

First, being a characteristic function is equivalent to being positive-definite with $k(0) = 1$, a consequence of Bochner's theorem. The positive definiteness comes from the fact that the integrand in (4) is a triangular function with scaled width $|x|$. Based on the well-known relation

$$\int_{\mathbb{R}} e^{\mathbf{i}\, rt} \max\{0, 1 - |r|\}\, dr = \frac{4}{t^2} \sin\left(\frac{t}{2}\right)^2,$$

if k is absolutely integrable, then k admits an inverse Fourier transform

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} k(r) e^{-\mathbf{i}\, rt}\, dr = \frac{1}{2\pi} \int_0^{\infty} x \operatorname{sinc}^2\left(\frac{xt}{2}\right) dF(x) =: h(t), \tag{5}$$

where $\operatorname{sinc}(x) = \sin(x)/x$. Clearly, h is nonnegative for all t. Then, the Bochner's characterization (1) is satisfied with a stronger condition for the cdf, one that admits a density:

$$k(r) = \int_{\mathbb{R}} e^{\mathbf{i}\, rx} h(x)\, dx.$$

If k is not absolutely integrable, one invokes Lévy's continuity theorem and shows that a sequence of absolutely integrable and positive-definite functions converge to k. Both cases conclude that k is positive-definite.

Second, the cdf F in (4) may be constructed as

$$F(x) = 1 - k(x) + xg(x) \qquad \text{with} \qquad g(x) = \lim_{\delta \to 0^+} \frac{k(x+\delta) - k(x)}{\delta}.$$

Here, g is the right derivative of k. By the continuity and convexity of k, g is well defined. Clearly, if k is differentiable, then $F(x) = 1 - k(x) + xk'(x)$. Further, if k is twice differentiable, we have that F is differentiable with

$$F'(x) = xk''(x),$$

which recovers (3) established in [15].

Third, it is important to note that F is supported on $(0, \infty)$, not $[0, \infty)$. If one considers the domain of a cdf to be the whole real line, then the requirement for F in the theorem may be equivalently stated as $F(x) = 0$ for all $x \le 0$. Apart from an obvious practical constraint seen later, that the random variable X will be used as the width of a bin, which must be positive, we particularly note that $F(0)$ cannot be nonzero. Such a constraint is naturally satisfied by continuous distributions, because F must be continuous at 0. However, a discrete distribution may assign a nonzero mass for $X = 0$, which makes $F(0) \neq 0$, a case we must rule out in the theorem. The reason is that if $\Pr(X = 0)$ is nonzero, then $dF(x)$ makes a nontrivial contribution to the Stieltjes integral (4) when x approaches 0 from the right. In such a case, $k(r)$ does not converge to 1 when $r \to 0$. In other words, we have to sacrifice either the equality $k(0) = 1$ or the continuity of k in the theorem, if we want to relax the support of F to $[0, \infty)$ with particularly allowing $F(0) \neq 0$.

This is not a sacrifice we make in this paper. Later in Section 3 when we construct kernels from discrete distributions, we will reiterate the requirement that $\Pr(X = 0) = 0$.

It is not hard to show that the converse of Theorem 1 is also true. Then, we strengthen the theorem into the following result and call it *Polya's characterization*. The significance is that any distribution on $(0, \infty)$ defines a positive-definite function, an additional characterization besides that of Bochner's.

Corollary 2. *A real function k is continuous and even with $k(0) = 1$, $\lim_{r\to\infty} k(r) = 0$, and convex on $[0, \infty)$, if and only if there exists a cdf $F(x)$ with positive support such that*

$$k(r) = \int_0^\infty \max\left\{0, 1 - \frac{|r|}{|x|}\right\} dF(x).$$

Moreover, all such functions k are positive-definite.

Proof. Theorem 1 corresponds to the "only if" part. Hence, it suffices to show the "if "part. Clearly, k is continuous, even, and satisfies $k(0) = 1$ and $\lim_{r\to\infty} k(r) = 0$. We therefore focus on only the convexity.

Let $r_1 \geq 0$, $r_2 \geq 0$, $r_1 \neq r_2$, and $t \in [0, 1]$. Define

$$L = \max\left\{0,\ t\left[1 - \frac{r_1}{|x|}\right] + (1-t)\left[1 - \frac{r_2}{|x|}\right]\right\}$$

and

$$R = \max\left\{0,\ t\left[1 - \frac{r_1}{|x|}\right]\right\} + \max\left\{0,\ (1-t)\left[1 - \frac{r_2}{|x|}\right]\right\}.$$

When r_1 and r_2 are on the same side of $|x|$, we have $L = R$. When $r_1 \leq |x| \leq r_2$, we have

$$L \leq \max\left\{0,\ t\left[1 - \frac{r_1}{|x|}\right]\right\} \leq R.$$

Similarly, when $r_2 \leq |x| \leq r_1$, we have

$$L \leq \max\left\{0,\ (1-t)\left[1 - \frac{r_2}{|x|}\right]\right\} \leq R.$$

Hence, all cases point to that $L \leq R$. Therefore,

$$k(tr_1 + (1-t)r_2) \leq tk(r_1) + (1-t)k(r_2),$$

concluding the convexity of k on $[0, \infty)$. □

Because the central subject of this paper, the function k in Corollary 2, is even, its Fourier transform and inverse transform differ by only a factor of 2π. In what follows, we do not distinguish the two transforms and consider only the forward one, with formal notation

$$\mathcal{F}[k](t) \equiv \int_{-\infty}^{\infty} k(r) e^{\mathbf{i}rt}\, dr.$$

We will also use Fourier transforms in the more general setting—one that is defined for generalized functions—which does not require k to be absolutely integrable.

2.1 Special Case

Based on the foregoing, because $F(x) = 0$ for all $x \le 0$, we may get rid of the max operator and write equivalently,

$$k(r) = \int_r^\infty \left(1 - \frac{r}{x}\right) dF(x) = \int_r^\infty dF(x) - r \int_r^\infty \frac{dF(x)}{x}, \qquad r \ge 0, \tag{6}$$

omitting the obvious symmetric part $r < 0$. The second term on the right-hand side of (6), $r \int_r^\infty (1/x) dF(x)$, is finite when $r \to 0^+$, but not necessarily when the front factor r is dropped. In this subsection, we consider the special, but not-so-infrequent case, when

$$\int_r^\infty \frac{dF(x)}{x}$$

indeed converges to a finite number as $r \to 0^+$. A benefit of considering this case is that we may introduce another random variable to simplify the expressions for k and its Fourier transform sometimes. Later in Section 3 we show quite a few such examples. For convenience, the integration limit starts from $-\infty$ rather than 0.

Formally, let X be a random variable with cdf $F(x)$, where $F(x) = 0$ for all $x \le 0$. If

$$C := \int_{-\infty}^\infty \frac{dF(x)}{x} \tag{7}$$

is finite, define

$$\widetilde{F}(x) := \int_{-\infty}^x \frac{dF(t)}{Ct}. \tag{8}$$

Because $\widetilde{F}(-\infty) = 0$, $\widetilde{F}(\infty) = 1$, and $\widetilde{F}$ is nondecreasing and right continuous, it is the cdf of some random variable $\widetilde{X}$. The following theorem gives the expressions of k and its Fourier transform by using some quantities with respect to X and $\widetilde{X}$. For notational consistency, we will use $F_{\widetilde{X}}$ to replace $\widetilde{F}$ when appropriate.

Theorem 3. *Denote by F_Z and φ_Z the cdf and the cf of a random variable Z, respectively. If C defined in* (7) *is finite and $\widetilde{X}$ is the respective random variable of $\widetilde{F}$ defined in* (8)*, then,*

$$k(r) = [1 - F_X(r)] - Cr[1 - F_{\widetilde{X}}(r)], \qquad r \ge 0,$$

and

$$\mathcal{F}[k](t) = \frac{C}{t^2}[2 - \varphi_{\widetilde{X}}(t) - \varphi_{\widetilde{X}}(-t)].$$

Proof. The expression of k is straightforward in light of (6). To show the Fourier transform, we apply (5) and write

$$\int_{-\infty}^\infty k(r) e^{\mathbf{i}\, rt}\, dr = \frac{1}{t^2} \int_{-\infty}^\infty \frac{2 - 2\cos(xt)}{x}\, dF(x) = \frac{C}{t^2} \int_{-\infty}^\infty [2 - 2\cos(xt)]\, d\widetilde{F}(x).$$

Then, we have

$$\frac{C}{t^2} \int_{-\infty}^\infty [2 - 2\cos(xt)]\, d\widetilde{F}(x) = \frac{C}{t^2} \left[\int_{-\infty}^\infty 2\, d\widetilde{F}(x) - \int_{-\infty}^\infty e^{\mathbf{i}\, xt}\, d\widetilde{F}(x) - \int_{-\infty}^\infty e^{-\mathbf{i}\, xt}\, d\widetilde{F}(x) \right],$$

which simplifies to the second equality in the theorem. □

This theorem is extensively applied in Section 3. Let us note two cases. For the case of discrete distributions, let S be the support and denote by f the probability mass function (pmf). Then, (7) and (8) read

$$C = \sum_{x \in S} \frac{f(x)}{x} \quad \text{and} \quad \widetilde{f}(x) = \frac{f(x)}{Cx}, \tag{9}$$

where $\widetilde{f}$ is the pmf of the new random variable $\widetilde{X}$ stated in the theorem. In particular, if the elements of S are all ≥ 1, or if the number of elements < 1 is finite, or if the number of elements < 1 is infinite but all are bounded away from 0, then C must be finite.

For the case of continuous distributions, if F is differentiable on $(0, \infty)$ and f is the corresponding pdf (i.e., $f = F'$), then (7) and (8) become

$$C = \int_0^\infty \frac{f(x)}{x}\, dx \quad \text{and} \quad \widetilde{f}(x) = \frac{f(x)}{Cx}, \tag{10}$$

where $\widetilde{f}$ is the pdf of the new random variable $\widetilde{X}$ stated in the theorem.

2.2 Scaling

Substantial experiences in kernel methods suggest that the spread of a kernel is one of the most important factors that affect the performance of a regression/classification. A well-known (though improper) example is the scale parameter σ in a squared exponential kernel $k(r) = \exp[-r^2/(2\sigma^2)]$. This example is improper because the kernel does not correspond to any cdf F in (2); nevertheless, the spirit of the example is that one needs to properly scale a kernel in order to achieve optimal results.

Hence, we introduce a scaling factor $\rho > 0$ and turn $k(r)$ to $k(\rho r)$. Because of the vast difference in spreads among kernels constructed from different cdf's, a principled approach is to define $\rho = A/\tau$, where A is used to standardize all kernels and τ is a tuning parameter that adjust the spread of the standardized kernel. One approach of standardization is to let A be the area under curve, because then the area under $k(Ar)$ is 1. The following result gives A.

Theorem 4. *If the random variable X has a finite mean, then for k defined in Corollary 2 we have*

$$\int_{-\infty}^{\infty} k(r)\, dr = E[X]. \tag{11}$$

Proof. Because k is even, a direct calculation gives

$$\begin{aligned} \int_{-\infty}^{\infty} k(r)\, dr &= 2 \int_0^\infty \left[\int_0^\infty \max\left\{0, 1 - \frac{r}{x}\right\} dF(x) \right] dr \\ &= 2 \int_0^\infty \left[\int_0^\infty \max\left\{0, 1 - \frac{r}{x}\right\} dr \right] dF(x) = \int_0^\infty x\, dF(x), \end{aligned}$$

where the interchange of integration order is permissible under the assumption that $E[X] < \infty$. □

Remark 1. As a straightforward corollary, $\mathcal{F}[k](0) = E[X]$.

The scaling $\rho = A/\tau = E[X]/\tau$ is a key ingredient in parameter tuning when we compare the empirical performance of kernels. Note that with $k(r)$ scaled to $k(\rho r)$, the following facts occur simultaneously for continuous random variables:

1. The cdf that constructs $k(\rho r)$ is $F(\rho x)$;
2. The corresponding random variable is X/ρ;
3. The Fourier transform of $k(\rho r)$ evaluates to $\frac{1}{\rho}\mathcal{F}[k](\frac{t}{\rho})$.

For discrete variables, the same facts hold, too; but be minded that the support is possibly changed (e.g., from integers to real numbers).

3 Example Kernels

An application of Polya's characterization is to construct positive-definite functions from known probability distributions. In this section, we consider a number of applicable distributions, either discrete or continuous, and derive explicit formulas for the corresponding kernel k and its Fourier transform. There incur a number of special functions, whose definitions are given in Appendix A. The definitions generally conform to convention.

3.1 Constructed from (Shifted) Poisson Distribution

If Y is a random variable of the Poisson distribution $\mathrm{Pois}(\mu)$ with rate $\mu > 0$, we have the following known facts:

1. pmf $f_Y(x) = \dfrac{\mu^x e^{-\mu}}{x!}$, $x = 0, 1, 2, \ldots$
2. cdf $F_Y(x) = \dfrac{\Gamma(\lfloor x+1 \rfloor, \mu)}{\Gamma(\lfloor x+1 \rfloor)}$,
3. mean $E[Y] = \mu$,
4. cf $\varphi_Y(t) = \exp[\mu(e^{\mathbf{i}t} - 1)]$,

where $\Gamma(s)$ is the gamma function and $\Gamma(s,t)$ is the upper incomplete gamma function, with t being the lower integration limit (see Appendix A for the formal definition).

Because the support of Y includes zero, we shift the distribution and define $X = Y + 1$, such that the value of the random variable starts from 1. Then, we have

$$f_X(x) = f_Y(x-1), \quad F_X(x) = F_Y(x-1), \quad E[X] = E[Y] + 1, \quad x = 1, 2, \ldots$$

To derive the kernel and its Fourier transform, consider the random variable $\widetilde{X}$ stated in Theorem 3 and subsequently revealed by (9). We write

$$\frac{f_X(x)}{x} = \frac{f_Y(x-1)}{x} = \frac{1}{\mu} \cdot \frac{\mu^x e^{-\mu}}{x!} = \frac{1}{\mu} f_Y(x).$$

Then, clearly,

$$C = \frac{1}{\mu} \quad \text{and} \quad \widetilde{X} = Y.$$

Thus, applying Theorem 3, we immediately obtain the kernel and the Fourier transform explicitly:

$$k(r) = \begin{cases} 1 - \dfrac{r}{\mu}\left[1 - \dfrac{\Gamma(\lfloor r+1 \rfloor, \mu)}{\Gamma(\lfloor r+1 \rfloor)}\right], & 0 \le r < 1, \\ \left[1 - \dfrac{\Gamma(\lfloor r \rfloor, \mu)}{\Gamma(\lfloor r \rfloor)}\right] - \dfrac{r}{\mu}\left[1 - \dfrac{\Gamma(\lfloor r+1 \rfloor, \mu)}{\Gamma(\lfloor r+1 \rfloor)}\right], & r \ge 1, \end{cases} \tag{12}$$

and

$$\mathcal{F}[k](t) = \frac{2 - \exp[\mu(e^{\mathbf{i}\,t} - 1)] - \exp[\mu(e^{-\mathbf{i}\,t} - 1)]}{\mu t^2}. \tag{13}$$

Note that the constructed kernel k is piecewise linear.

3.2 Constructed from Gamma Distribution

If X is a random variable of the gamma distribution $\mathrm{Gamma}(s, \theta)$ with shape $s > 0$ and scale $\theta > 0$, we have the following known facts:

1. pdf $f(x) = \dfrac{x^{s-1}e^{-x/\theta}}{\Gamma(s)\theta^s}$,
2. cdf $F_X(x) = 1 - \dfrac{\Gamma(s, x/\theta)}{\Gamma(s)}$,
3. mean $E[X] = \theta s$,
4. cf $\varphi_X(t) = (1 - \mathbf{i}\,\theta t)^{-s}$.

We discuss three cases of the shape s. When $s > 1$, we write

$$\frac{f(x)}{x} = \frac{1}{(s-1)\theta} \cdot \frac{x^{s-2}e^{-x/\theta}}{\Gamma(s-1)\theta^{s-1}}.$$

Then, clearly, with respect to (10),

$$C = \frac{1}{(s-1)\theta} \quad \text{and} \quad \widetilde{X} \sim \mathrm{Gamma}(s-1, \theta).$$

Applying Theorem 3, we immediately obtain the kernel and the Fourier transform explicitly:

$$k(r) = \frac{\Gamma(s, r/\theta) - r/\theta \cdot \Gamma(s-1, r/\theta)}{\Gamma(s)}, \quad r \ge 0, \tag{14}$$

$$\mathcal{F}[k](t) = \frac{2\left[1 - \cos^{s-1}(\omega)\cos((s-1)\omega)\right]}{(s-1)\theta t^2}, \quad \text{with} \quad \cos(\omega) = (1 + \theta^2 t^2)^{-\frac{1}{2}}. \tag{15}$$

When $s = 1$, the distribution $\mathrm{Gamma}(s-1, \theta)$ is undefined. However, we may derive the kernel function directly from (6):

$$k(r) = e^{-r/\theta} - r/\theta \cdot E_1(r/\theta), \quad r \ge 0, \tag{16}$$

where E_1 is the exponential integral. The Fourier transform admits a closed form due to the known sine transform of E_1.

Theorem 5. *For k defined in* (16)*,*

$$\mathcal{F}[k](t) = \frac{\log(1+\theta^2 t^2)}{\theta t^2}. \tag{17}$$

Proof. Based on (16), we perform a reparameterization $\lambda = 1/\theta$ and write

$$\int_{-\infty}^{\infty} k(r) e^{\mathbf{i}\, rt}\, dr = 2\int_0^{\infty} e^{-\lambda r}\cos(rt)\, dr - 2\int_0^{\infty} \lambda r E_1(\lambda r)\cos(rt)\, dr.$$

The first term is a commonly used integral and it is evaluated to

$$\int_0^{\infty} e^{-\lambda r}\cos(rt)\, dr = \frac{\lambda}{\lambda^2 + t^2}.$$

Then, we perform integration by parts on the second term. Noting that $E_1'(r) = -e^{-r}/r$, we obtain

$$\begin{aligned}\int_0^{\infty} rE_1(\lambda r)\cos(rt)\, dr &= rE_1(\lambda r)\frac{\sin(rt)}{t}\bigg|_0^{\infty} - \int_0^{\infty} -\lambda r E_1'(\lambda r)\frac{\sin(rt)}{t}\, dr - \int_0^{\infty} E_1(\lambda r)\frac{\sin(rt)}{t}\, dr \\ &= 0 + \frac{1}{t}\int_0^{\infty} e^{-\lambda r}\sin(rt)\, dr - \frac{1}{t}\int_0^{\infty} E_1(\lambda r)\sin(rt)\, dr.\end{aligned}$$

The middle term is a commonly used integral and it is evaluated to

$$\int_0^{\infty} e^{-\lambda r}\sin(rt)\, dr = \frac{t}{\lambda^2 + t^2}.$$

According to Section 2.11, Equation (18) of [3, p.98], we have for the third term

$$\int_0^{\infty} E_1(\lambda r)\sin(rt)\, dr = \frac{1}{2t}\log\left(1 + \frac{t^2}{\lambda^2}\right).$$

Combining all these results, we obtain

$$\int_{-\infty}^{\infty} k(r) e^{\mathbf{i}\, rt}\, dr = \frac{\lambda}{t^2}\log\left(1 + \frac{t^2}{\lambda^2}\right),$$

which concludes the theorem. □

When $s < 1$, the expression of k in (6) incurs incomplete gamma functions with negative arguments. Such functions are not standard. We therefore do not consider this case. Note, however, that although we do not have an explicit expression for k, the results in the preceding section still guarantee that k is a valid kernel.

3.3 Constructed from Exponential Distribution

If X is a random variable of the exponential distribution $\mathrm{Exp}(\theta)$ with scale $\theta > 0$, that is,

$$f(x) = \frac{1}{\theta} e^{-x/\theta},$$

then it also belongs to the gamma distribution with shape $s = 1$ and scale θ. Hence, the corresponding kernel k and its Fourier transform are given in (16) and (17) of Section 3.2, respectively.

3.4 Constructed from Chi-Square Distribution

If X is a random variable of the chi-square distribution χ^2_ν with degree of freedom ν, that is,

$$f(x) = \frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)},$$

then it also belongs to the gamma distribution with shape $s = \nu/2$ and scale $\theta = 2$. In particular, when $\nu = 2$, the corresponding kernel k and its Fourier transform are given in (16) and (17) of Section 3.2, respectively. When $\nu > 2$, the respective formulas are give in (14) and (15).

3.5 Constructed from Chi Distribution

If X is a random variable of the chi distribution χ_ν with degree of freedom ν, we have the following known facts:

1. pdf $f(x) = \dfrac{2^{1-\nu/2}}{\Gamma(\nu/2)} x^{\nu-1}e^{-x^2/2}$,
2. cdf $F_X(x) = 1 - \dfrac{\Gamma(\nu/2, x^2/2)}{\Gamma(\nu/2)}$,
3. mean $E[X] = \sqrt{2}\dfrac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}$,
4. cf $\varphi_X(t) = M\left(\dfrac{\nu}{2}, \dfrac{1}{2}, \dfrac{-t^2}{2}\right) + \mathbf{i}\, t\sqrt{2}\dfrac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} M\left(\dfrac{\nu+1}{2}, \dfrac{3}{2}, \dfrac{-t^2}{2}\right)$,

where $M(a, b, z)$ is Kummer's confluent hypergeometric function.

We discuss two cases of ν. When $\nu > 1$, we write

$$\frac{f(x)}{x} = \frac{\Gamma((\nu-1)/2)}{\sqrt{2}\Gamma(\nu/2)} \frac{2^{1-(\nu-1)/2}}{\Gamma((\nu-1)/2)} x^{\nu-2}e^{-x^2/2}.$$

Then, clearly, with respect to (10),

$$C = \frac{\Gamma((\nu-1)/2)}{\sqrt{2}\Gamma(\nu/2)} \quad \text{and} \quad \widetilde{X} \sim \chi_{\nu-1}.$$

Applying Theorem 3, we immediately obtain the kernel and the Fourier transform explicitly:

$$k(r) = \frac{\Gamma(\nu/2, r^2/2) - r/\sqrt{2}\cdot\Gamma((\nu-1)/2, r^2/2)}{\Gamma(\nu/2)}, \quad r \geq 0, \tag{18}$$

$$\mathcal{F}[k](t) = \frac{\sqrt{2}\Gamma((\nu-1)/2)}{t^2\Gamma(\nu/2)} \left[1 - M\left(\frac{\nu-1}{2}, \frac{1}{2}, \frac{-t^2}{2}\right)\right]. \tag{19}$$

Note that as a special case, when $\nu = 2$, the distribution χ_ν is the same as the Rayleigh distribution with scale $\sigma = 1$; see Section 3.7. The explicit expressions for the kernel and the Fourier transform will be presented therein for a general scale parameter σ.

When $\nu = 1$, the distribution χ_ν is the same as the half-normal distribution with scale $\sigma = 1$; see Section 3.6. The explicit expressions will be presented therein for a general σ.

3.6 Constructed from Half-Normal Distribution

If X is a random variable of the half-normal distribution $\mathrm{HN}(\sigma)$ with scale $\sigma > 0$, we have the following known facts:

1. pdf $f(x) = \dfrac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\dfrac{x^2}{2\sigma^2}\right)$,
2. cdf $F_X(x) = \operatorname{erf}\left(\dfrac{x}{\sigma\sqrt{2}}\right)$,
3. mean $E[X] = \dfrac{\sigma\sqrt{2}}{\sqrt{\pi}}$,
4. cf $\varphi_X(t) = e^{-\sigma^2 t^2/2}[1 - \mathbf{i}\,\operatorname{erfi}(\sigma t/\sqrt{2})]$,

where erf is the error function and erfi is the imaginary error function (which, in fact, is a real-valued function when the argument is real).

We may not apply Theorem 3 to derive the explicit formula for k, because C is infinite. However, with a change of variable $y = x^2/(2\sigma^2)$, we see that a part of (6) is evaluated to

$$\int_r^\infty \frac{f(x)}{x}\,dx = \int_r^\infty \frac{\sqrt{2}}{x\sigma\sqrt{\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{r^2/(2\sigma^2)}^\infty \frac{e^{-y}}{y}\,dy = \frac{E_1(r^2/(2\sigma^2))}{\sigma\sqrt{2\pi}}.$$

Therefore, an explicit expression for the kernel is

$$k(r) = \operatorname{erfc}\left(\frac{r}{\sigma\sqrt{2}}\right) - \frac{1}{\sqrt{\pi}}\left(\frac{r}{\sigma\sqrt{2}}\right) E_1\left(\frac{r^2}{2\sigma^2}\right), \quad r \geq 0, \tag{20}$$

where erfc is the complementary error function. The following theorem gives the Fourier transform of k in the form of a sine transform, which unfortunately is hard to be further simplified.

Theorem 6. *For k defined in* (20)*,*

$$\mathcal{F}[k](t) = \frac{2}{t\sqrt{\pi}} \int_0^\infty E_1(r^2) \sin(\sigma\sqrt{2}tr)\,dr. \tag{21}$$

Proof. We first turn $k(r)$ to $k(\sigma\sqrt{2}r)$ in order to simply the math:

$$\begin{aligned}\int_{-\infty}^\infty k(r)e^{\mathbf{i}\,rt}\,dr &= 2\int_0^\infty k(r)\cos(rt)\,dr = 2\sqrt{2}\sigma \int_0^\infty k(\sigma\sqrt{2}r)\cos(\sigma\sqrt{2}rt)\,dr \\ &= 2\sqrt{2}\sigma\left[\int_0^\infty \operatorname{erfc}(r)\cos(rT)\,dr - \frac{1}{\sqrt{\pi}}\int_0^\infty rE_1(r^2)\cos(rT)\,dr\right],\end{aligned} \tag{22}$$

where $T = \sigma\sqrt{2}t$. For the first term, we rearrange the order of integration:

$$\begin{aligned}\int_0^\infty \operatorname{erfc}(r)\cos(rT)\,dr &= \int_0^\infty \frac{2}{\sqrt{\pi}}\int_r^\infty e^{-x^2}dx\cos(rT)\,dr \\ &= \frac{2}{\sqrt{\pi}}\int_0^\infty e^{-x^2}\int_0^x \cos(rT)\,drdx = \frac{2}{T\sqrt{\pi}}\int_0^\infty e^{-x^2}\sin(xT)\,dx.\end{aligned} \tag{23}$$

For the second term, we perform integration by parts:

$$\begin{aligned}\int_0^\infty rE_1(r^2)\cos(rT)\,dr &= \frac{rE_1(r^2)\sin(rT)}{T}\bigg|_0^\infty - \int_0^\infty \frac{\sin(rT)}{T}\left[E_1(r^2) - 2e^{-r^2}\right]\,dr \\ &= -\frac{1}{T}\int_0^\infty \sin(rT)E_1(r^2)\,dr + \frac{2}{T}\int_0^\infty \sin(rT)e^{-r^2}\,dr. \end{aligned} \tag{24}$$

Substituting (23) and (24) into (22), we obtain the result of the theorem. □

3.7 Constructed from Rayleigh Distribution

If X is a random variable of the Rayleigh distribution Rayleigh(σ) with scale $\sigma > 0$, we have the following known facts:

1. pdf $f(x) = \dfrac{x}{\sigma^2}e^{-x^2/(2\sigma^2)}$,
2. cdf $F_X(x) = 1 - e^{-x^2/(2\sigma^2)}$,
3. mean $E[X] = \sigma\sqrt{\dfrac{\pi}{2}}$.

To derive explicit expressions, we note that

$$\frac{f(x)}{x} = \frac{\sqrt{\pi}}{\sigma\sqrt{2}}\frac{\sqrt{2}}{\sigma\sqrt{\pi}}e^{-x^2/(2\sigma^2)}.$$

Then, clearly, with respect to (10),

$$C = \frac{1}{\sigma}\sqrt{\frac{\pi}{2}} \quad \text{and} \quad \widetilde{X} \sim \mathrm{HN}(\sigma).$$

Applying Theorem 3 with the known facts for the half-normal distribution listed in Section 3.6, we immediately obtain the kernel and the Fourier transform explicitly:

$$k(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right) - \sqrt{\pi}\left(\frac{r}{\sigma\sqrt{2}}\right)\operatorname{erfc}\left(\frac{r}{\sigma\sqrt{2}}\right), \quad r \geq 0, \tag{25}$$

$$\mathcal{F}[k](t) = \frac{\sqrt{2\pi}}{\sigma t^2}\left[1 - \exp\left(-\frac{\sigma^2 t^2}{2}\right)\right]. \tag{26}$$

3.8 Constructed from Nakagami Distribution

If X is a random variable of the Nakagami distribution Nakagami(m, Ω) with shape $m \geq 1/2$ and spread $\Omega > 0$, that is,

$$f(x) = \frac{2m^m x^{2m-1}e^{-mx^2/\Omega}}{\Gamma(m)\Omega^m},$$

we may perform a reparameterization

$$m = \nu/2, \quad \Omega = \nu\theta^2,$$

and obtain

$$f(x) = \frac{1}{\theta} \cdot \frac{2^{1-\nu/2}}{\Gamma(\nu/2)} (x/\theta)^{\nu-1} e^{-(x/\theta)^2/2}.$$

Clearly, f is a rescaling of the pdf of the chi distribution χ_ν, with the integer ν (degree of freedom) relaxed to a real number.

We discuss two cases of m. When $m > 1/2$ (i.e., $\nu > 1$), we will reuse the formulas (18) and (19) derived for χ_ν. The reason why (18) and (19) are valid for non-integers ν is that they are derived from the cdf and the cf of χ_ν, wherein the integration results are valid for any real numbers $\nu > 1$. Then, with a proper scaling, we have for the Nakagami distribution:

$$k(r) = \frac{\Gamma(m, mr^2/\Omega) - \sqrt{m}r/\sqrt{\Omega} \cdot \Gamma(m - 1/2, mr^2/\Omega)}{\Gamma(m)}, \quad r \geq 0, \tag{27}$$

$$\mathcal{F}[k](t) = 2\sqrt{\frac{m}{\Omega}} \frac{\Gamma(m - 1/2)}{t^2 \Gamma(m)} \left[1 - M\left(m - \frac{1}{2}, \frac{1}{2}, \frac{-\Omega t^2}{4m}\right)\right]. \tag{28}$$

When $m = 1/2$, the distribution is the same as the half-normal distribution with scale $\sigma = \sqrt{\Omega}$. Then, substituting $\sigma = \sqrt{\Omega}$ into (20) and (21), we have

$$k(r) = \operatorname{erfc}\left(\frac{r}{\sqrt{2\Omega}}\right) - \frac{1}{\sqrt{\pi}} \left(\frac{r}{\sqrt{2\Omega}}\right) E_1\left(\frac{r^2}{2\Omega}\right), \quad r \geq 0, \tag{29}$$

$$\mathcal{F}[k](t) = \frac{2}{t\sqrt{\pi}} \int_0^\infty E_1(r^2) \sin(\sqrt{2\Omega} t r)\, dr. \tag{30}$$

3.9 Constructed from Weibull Distribution

If X is a random variable of the Weibull distribution Weibull(θ, α) with scale $\theta > 0$ and shape $\alpha > 0$, we have the following known facts:

1. pdf $f(x) = \dfrac{\alpha}{\theta} \left(\dfrac{x}{\theta}\right)^{\alpha-1} e^{-(x/\theta)^\alpha}$,
2. cdf $F_X(x) = 1 - e^{-(x/\theta)^\alpha}$,
3. mean $E[X] = \theta\Gamma(1 + 1/\alpha)$.

We discuss two cases of α. When $\alpha > 1$, with a change of variable $y = (x/\theta)^\alpha$, we see that a part of (6) is evaluated to

$$\int_r^\infty \frac{f(x)}{x}\, dx = \int_r^\infty \frac{\alpha}{\theta} \left(\frac{x}{\theta}\right)^{\alpha-1} e^{-(x/\theta)^\alpha} \frac{1}{x}\, dx = \frac{1}{\theta} \int_{(r/\theta)^\alpha}^\infty y^{-1/\alpha} e^{-y}\, dy = \frac{1}{\theta}\Gamma(1 - 1/\alpha, (r/\theta)^\alpha).$$

Therefore, an explicit expression for the kernel is

$$k(r) = e^{-(r/\theta)^\alpha} - (r/\theta)\Gamma(1 - 1/\alpha, (r/\theta)^\alpha). \tag{31}$$

We do not have an explicit expression for the Fourier transform, unfortunately.

When $\alpha = 1$, the distribution is the same as the exponential distribution with scale θ; it is also the same as the gamma distribution with shape $s = 1$ and scale θ. Hence, the corresponding kernel k and its Fourier transform are given in (16) and (17) of Section 3.2, respectively.

3.10 Summary

We summarize the results obtained so far in Table 3 (located after the bibliography). This table lists many applicable distributions and the correspondingly constructed kernels. Accompanied with the distributions are the pmf/pdf's and the mean's. The pmf/pdf's are used to uniquely identify the distributions, because different authors may call the parameters differently. Moreover, for the Poisson distribution, it has been shifted to avoid a nonzero mass at the origin. Hence, one is suggested to fully digest the notations before usage. The mean's are used to standardize a kernel so that the area under curve is 1 (see Section 2.2). Accompanied with the kernels are the explicit expressions for k and the Fourier transform $\mathcal{F}[k]$. These expressions could be used, for example, for further deriving analytic properties.

The table consists of three parts. The top part contains a discrete distribution, whereas the other two parts contain continuous ones. The distributions in the bottom part are special cases of those in the middle. The equivalence is indicated in the last column. Therefore, we consider that practical use of the distributions focuses mainly on the top and middle parts of the table.

A practical aspect for the use of the distributions is the choice of parameters, which is reflected in the last column. All continuous distributions therein contain a "scale" parameter that affects the spread in one way or another. Because we use the distribution mean $A = E[X]$ to perform standardization, we may fix the scale parameter at an arbitrary value (particularly, 1) and let the actual spread be determined by a scaling factor $\rho = A/\tau$ where τ is tuned (see Section 2.2). Apart from the scale parameter, some distributions come additionally with a "shape" parameter, which appears as an exponent for x in the pdf. When tuning such a parameter, one may search for an optimal one from a grid (e.g., integers and half-integers). The same practice applies to the "rate" parameter of Poisson.

Figures 4 and 5 (located after the bibliography) plot the kernels listed on the top and the middle parts of Table 3, with several choices of a parameter (μ in Poisson, s in gamma, m in Nakagami, and α in Weibull). As expected, the kernels are all convex and monotonically decreasing from 1 to 0. The right column of the figure shows the kernels scaled by the distribution mean; therefore, the area under curve is 1. These curves smoothly vary with the parameter.

4 Random Feature Maps

Mercer's theorem [14] guarantees that there exists a feature map $z(x)$ such that $k(x - x')$ is equal to the inner product $\langle z(x), z(x')\rangle$, where z is a finite-dimensional or countably infinite-dimensional vector. The random feature approaches construct such maps so that $z(x)$ is random and that the expectation of $\langle z(x), z(x')\rangle$ is equal to $k(x - x')$. Naturally, one may define D independent copies of z, namely, $z^{(l)}$ for $l = 1, \ldots, D$, and use the Monte Carlo sample average $\frac{1}{D}\sum_{l=1}^{D}\langle z^{(l)}(x), z^{(l)}(x')\rangle$ to reduce the randomness of the inner product as an unbiased approximation to the kernel k. In this section, we compare the randomness of different approaches.

On notation: The data x in the general case is a vector; however in some cases (e.g., random binning), the kernel acts on a scalar input x. The feature map z may be a scalar-valued or a vector-valued map, depending on context. We define a random function $\tilde{k}$ as a shorthand notation of the inner product:

$$\tilde{k}(x, x') \equiv \langle z(x), z(x')\rangle$$

and write $\widetilde{K}$ as the corresponding kernel matrix. Note that although k is stationary, $\tilde{k}$ may not

(hence the notations are $k(x-x')$ and $\tilde{k}(x,x')$, respectively). Then, with D independent copies of the feature maps, the corresponding kernel matrix becomes $\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}^{(l)}$. We are interested in the probabilistic properties of $\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}^{(l)} - K$. Because the dimension of the data matters only in the Fourier transform of the kernel, the mathematical derivation here focuses on one-dimensional kernel functions. Generalizations to the multidimensional case are straightforward. The theorems in this section are presented to be applicable to the multidimensional case, too.

4.1 Random Fourier Map

The random Fourier approach defines the feature map $z(x) = e^{\mathbf{i}\,wx}$, where w is drawn from the cdf F in (1). Then, the same equation immediately verifies that the inner product $\langle z(x), z(x')\rangle = e^{\mathbf{i}\,w(x-x')}$ has an expectation $k(x-x')$. Additionally, we easily obtain that the variance of the inner product is

$$\mathrm{Var}[\langle z(x), z(x')\rangle] = \left[\int |e^{\mathbf{i}\,w(x-x')}|^2\, dF(w)\right] - k(x-x')^2 = 1 - k(x-x')^2. \tag{32}$$

In practice, it is often more desirable to use a feature map that is real-valued. Hence, the real version of the map is $z(x) = \sqrt{2}\cos(wx+b)$, where b is drawn from $\mathcal{U}(0, 2\pi)$. This map still yields expectation $k(x-x')$ for the inner product:

$$\begin{aligned} E[\langle z(x), z(x')\rangle] &= \int_{-\infty}^{\infty}\int_{0}^{2\pi}\Big(2\cos(wx+b)\cos(wx'+b)\Big)\frac{1}{2\pi}\,db\,dF(w) \\ &= \int_{-\infty}^{\infty}\cos(w(x-x'))\,dF(w) = k(x-x'), \end{aligned}$$

but gives a larger variance:

$$\begin{aligned} \mathrm{Var}[\langle z(x), z(x')\rangle] &= \left[\int_{-\infty}^{\infty}\int_{0}^{2\pi}\Big(2\cos(wx+b)\cos(wx'+b)\Big)^2\frac{1}{2\pi}\,db\,dF(w)\right] - k(x-x')^2 \\ &= \left[\int_{-\infty}^{\infty}\Big(1+\frac{1}{2}\cos(2w(x-x'))\Big)\,dF(w)\right] - k(x-x')^2 \\ &= 1 + \frac{1}{2}k(2(x-x')) - k(x-x')^2. \end{aligned} \tag{33}$$

The feature map is straightforwardly generalized to the multidimensional case through multidimensional Fourier transform, the details of which are omitted here. With one further generalization—using a Monte Carlo sample average of D independent copies to replace z—we arrive at the following result. It states that the random Fourier approach gives an unbiased approximation. It also gives the squared Frobenius norm error of the approximation.

Theorem 7. *Let K be the kernel matrix of a kernel k on data points x_i, $i = 1,\ldots,n$. Let $\widetilde{K}^{(l)}$, $l = 1,\ldots,D$ be the kernel matrices resulting from D independent random Fourier feature maps for k. We have*

$$E\left[\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}^{(l)}\right] = K.$$

Moreover, for the complex feature map,

$$E\left[\left\|\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}^{(l)} - K\right\|_F^2\right] = \frac{1}{D}(n^2 - \|K\|_F^2),$$

and for the real feature map,

$$E\left[\left\|\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}^{(l)}-K\right\|_F^2\right]=\frac{1}{D}\left(n^2+\frac{1}{2}\sum_{i,j=1}^{n}k(2(x_i-x_j))-\|K\|_F^2\right).$$

Proof. The first expectation is obvious and the second one is analogous to the third one. Thus, we prove only the second one. By the linearity of expectation, we have

$$E\left[\left\|\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}^{(l)}-K\right\|_F^2\right]=\sum_{i,j=1}^{n}E\left[\left(\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}_{ij}^{(l)}-K_{ij}\right)^2\right].$$

Note that inside the summation, each expectation is nothing but the variance of $\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}_{ij}^{(l)}$. Then, with independence,

$$E\left[\left(\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}_{ij}^{(l)}-K_{ij}\right)^2\right]=\mathrm{Var}\left[\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}_{ij}^{(l)}\right]=\frac{1}{D}\mathrm{Var}[\widetilde{K}_{ij}^{(1)}].$$

By (32), we see that $\mathrm{Var}[\widetilde{K}_{ij}^{(1)}]=1-k(x_i-x_j)^2$, which proves the second expectation in the theorem.

For the third expectation, follow the same argument and apply (33) at the end. □

4.2 Random Binning Map

The random binning approach applies to multidimensional kernel functions k that are a tensor product of one-dimensional kernels. The approach was originally proposed for only the exponential kernel, because based on (3), the term $wk''(w)$ happens to be a known pdf (gamma distribution of a certain shape). One easily generalizes the approach based on, instead, (2), through a reverse thinking: any cdf corresponds to a valid kernel. Hence, in the general setting, we consider the following construction, which defines a marginal distribution for the inner product $\tilde{k}=\langle z(x),z(x')\rangle$:

1. Let $F(w)$ be a cdf with positive support.
2. Let a random one-dimensional grid have spacing w and offset b, where $w\sim F(w)$ and $b\sim\mathcal{U}(0,w)$. In other words, we have the conditional probability density $f(b|w)=w^{-1}$.
3. Define the feature vector $z(x)$, one element for each grid bin, that takes 1 when x falls in the bin and 0 otherwise. For two points x and x', because the probability that they fall in the same bin is $\max\{0,1-r/w\}$ with $r=|x-x'|$, we have the conditional probability

$$\Pr(\tilde{k}=1\mid w,b)=\max\left\{0,1-\frac{r}{w}\right\},\qquad\Pr(\tilde{k}=0\mid w,b)=1-\Pr(\tilde{k}=1\mid w,b).$$

Therefore, this procedure defines a marginal distribution for $\tilde{k}$ whose pmf is

$$\Pr(\tilde{k}=1)=\int_0^\infty\int_0^w\Pr(\tilde{k}=1\mid w,b)f(b|w)\,db\,dF(w)=k(r),\qquad\text{(cf. (2))}$$

and $\Pr(\tilde{k} = 1) = 1 - \Pr(\tilde{k} = 0)$. In other words, $\tilde{k}$ is a Bernoulli variable with success probability k; hence, obviously,

$$E[\tilde{k}] = k \quad \text{and} \quad \mathrm{Var}[\tilde{k}] = k - k^2. \tag{34}$$

This feature map is straightforwardly generalized to the multidimensional case through using a multidimensional grid. With one further generalization—using a Monte Carlo sample average of D independent copies to replace z—we arrive at the following result, parallel to Theorem 7.

Theorem 8. *Let K be the kernel matrix of a kernel k and let $\widetilde{K}^{(l)}$, $l = 1, \ldots, D$ be the kernel matrices resulting from D independent random binning feature maps for k. We have*

$$E\left[\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}^{(l)}\right] = K,$$

and

$$E\left[\left\|\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}^{(l)} - K\right\|_F^2\right] = \frac{1}{D}\left(\sum_{i,j=1}^{n} K_{ij} - \|K\|_F^2\right).$$

Proof. The proof is analogous to that of Theorem 7, except that at the end we apply $\mathrm{Var}[\widetilde{K}_{ij}^{(1)}] = k(x_i - x_j) - k(x_i - x_j)^2$. □

4.3 Discussions

Theorems 7 and 8 indicate that for a kernel k that admits both random Fourier and random binning feature maps, the latter map results in an approximate kernel matrix closer to K than does the former map, if the same sample size D is used, because $0 \leq K_{ij} \leq 1$. Moreover, (32), (33), and (34) reveal that such a better approximation is elementwise. Of course, a better quality in matrix approximation does not necessarily imply a superior performance in a machine learning task, where the performance metric might not be directly connected with matrix approximation. In practice, however, our experience shows that random binning indeed performs better almost always, in the sense that it requires a (much) smaller D for a matching regression error/classification accuracy, compared with random Fourier. See experimental results in the next section.

One advantage of random Fourier, though, is that it generalizes more broadly to multidimensional inputs, through multidimensional Fourier transforms. As long as the respective multivariate probability distribution can be easily sampled from, the random Fourier map is efficient to compute. Such is the case, for example, for the squared exponential kernel (also called the Gaussian kernel), because the corresponding distribution is multivariate normal. As another example, the exponential kernel (note the vector norm)

$$\exp\left(-\frac{\|x - x'\|_2}{\sigma}\right)$$

is corresponded by multivariate Cauchy. In fact, both the squared exponential and the exponential kernels are special cases of the Matérn family of kernels [19, 16], whose corresponding distributions are the multivariate t-distributions, when the Matérn smoothness parameter is an integer or a half-integer.

On the other hand, random binning is applicable to only tensor-product kernels; e.g., the Laplace kernel

$$\exp\left(-\frac{\|x-x'\|_1}{\sigma}\right) = \exp\left(-\frac{|(x)_1-(x')_1|}{\sigma}\right)\exp\left(-\frac{|(x)_2-(x')_2|}{\sigma}\right)\cdots\exp\left(-\frac{|(x)_d-(x')_d|}{\sigma}\right),$$

where $(\cdot)_i$ is used to index the dimensions, not the data. Such is not the limitation of Polya's criterion, because one may easily generalize (2) to the multidimensional case by using a multidimensional positive-definite function to replace the triangular function in the integrand. However, the challenge is that if the integrand is not a tensor product, it is difficult to define a "bin" such that two points fall in the same bin with a probability equal to the integrand.

In the next section, we will perform an experiment that compares also the Gaussian kernel as an example kernel for random Fourier, which, despite the aforementioned advantage, performs less well than random binning.

5 Numerical Experiments

In this section, we demonstrate the empirical performance of random Fourier (denoted by "RF") and random binning (denoted by "RB"), as kernel approximation approaches for regression and classification, in the reproducing kernel Hilbert space (RKHS). We perform the experiments with eight benchmark data sets downloaded from `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`. The primary reason of using these data sets is their varying sizes n and dimensions d. Some of the data sets come with a train/test split; for those not, we performed a 4:1 split. Attributes were normalized to $[-1, 1]$. Table 1 gives the detailed information.

Table 1: Data sets.

Name	Type	d	n Train	n' Test
cadata	regression	8	16,512	4,128
YearPredictionMSD	regression	90	463,518	51,630
ijcnn1	binary classification	22	35,000	91,701
covtype.binary	binary classification	54	464,809	116,203
SUSY	binary classification	18	4,000,000	1,000,000
mnist	10 classes	780	60,000	10,000
acoustic	3 classes	50	78,823	19,705
covtype	7 classes	54	464,809	116,203

5.1 Matrix Approximation

The purpose of the following experiment is to empirically verify Theorems 7 and 8 regarding the kernel matrix approximation error, and show how large the gap could be between the two feature maps. For this, we use the Laplace kernel (tensor product of one-dimensional exponential kernels) as an example, because the two corresponding distributions, Cauchy for the random Fourier map and gamma for the random binning map, can be easily sampled from.

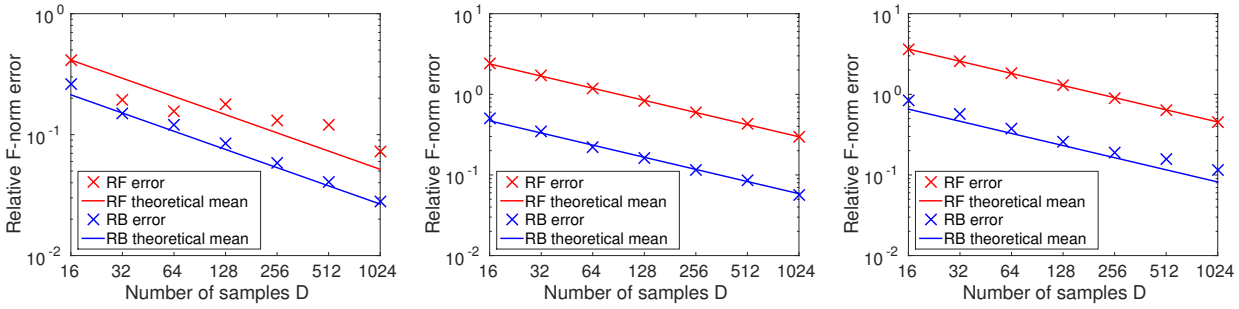
The examples are run on the three small data sets listed in Table 1—cadata, ijcnn1, and acoustic—whose full kernel matrices (sizes on the order $\sim 10^4$ to 10^5) are affordable to compute.

Before running a machine learning task, we do not know the optimal scale parameter σ in the kernel $k(r) = e^{-r/\sigma}$. Hence, we fix $\sigma = 1$ as a reasonable choice. For a related experiment that uses the tuned σ, see Section 5.3.

In Figure 1, we plot the relative Frobenius norm error, defined as

$$E\left[\left\|\frac{1}{D}\sum_{l=1}^{D}\widetilde{K}^{(l)} - K\right\|_F^2 \Big/ \|K\|_F^2\right]^{1/2}, \tag{35}$$

in straight lines. This quantity is similar to the so-called "standard deviation to mean ratio" in standard statistics. The lines are computed according to the results given by Theorems 7 and 8. Then, we plot the actual error (with the the expectation sign in (35) removed) as scattering crosses, overlaid with the lines.



(a) cadata (b) ijcnn1 (c) acoustic

Figure 1: Matrix approximation error as a function of the sample size D.

One sees that the actual error is well aligned with the theoretical mean. Furthermore, there is a clear gap between the two feature maps; random binning always yields a smaller error. The largest gap corresponds to almost a one-digit difference. Clearly, for different data sets, the gap may be different; and even for the same data set, the gap may also vary when the scale parameter σ varies. The spirit of this experiment, after all, is that the theoretical analysis gives a clear preference to random binning, empirically verified.

5.2 Regression/Classification

In the next experiment, we apply the random feature maps for regression and classification in the RKHS. The unified setting is that given data $\{x_i\}_{i=1}^n$ with targets $\{y_i\}_{i=1}^n$, we minimize the risk functional

$$\mathcal{L}(f) = \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \langle f, f \rangle_{\mathcal{H}_k}$$

within the RKHS $\mathcal{H}_k$ defined by a kernel function k, where $V(\cdot,\cdot)$ is a loss function, $\langle \cdot,\cdot \rangle$ is the inner product associated to $\mathcal{H}_k$, and λ is a regularization parameter. We choose to use the squared loss $V(t,y) = (t-y)^2$ as in [15], because due to the Representer Theorem [12, 17], the optimal function admits a well-known closed-form expression

$$f(x) = k_x(K + \lambda I)^{-1} y,$$

where k_x is the row vector of $k(x, x_i)$ for all i and y is the column vector of all targets. We will use the approximate kernel $\tilde{k}$, defined as the Euclidean inner product of the random feature maps, to replace a kernel k.

We perform the experiment with all data sets listed in Table 1. The performance metric is mean squared error (MSE) for regression and accuracy for classification. Parameters are tuned through cross validation. In particular, for random binning, the scaling factor $\rho = E[X]/\tau$ for the kernel is obtained through actually tuning the assisting parameter τ, as discussed in depth in Section 2.2.

We compare random Fourier with random binning, by using two kernels for the former map (Laplace and Gaussian) and four kernels for the latter (those constructed from shifted Poisson, gamma, Nakagami, and Weilbull distributions). Note that the Laplace kernel is equivalent to the one constructed from gamma distribution according to (2), with a particular shape $s = 2$. However, for the random binning map, the shape is considered a tuning parameter, which is not the case for the random Fourier map.

Figure 2 plots the regression/classification performance when the sample size D increases. One sees that the performance curves for the two random feature maps are separately clustered in general. The curves of random binning clearly indicate a better performance than do those of random Fourier. Table 2 lists the tuned parameters that generate the results of Figure 2. We display only those for random binning, because the parameters for random Fourier vary significantly when the number D of samples changes. One observation from the table is that the optimal shape s of the gamma distribution is not always achieved by 2. In other words, a better performance is obtained by treating s as a tuning parameter.

Table 2: Tuned parameters for the random binning maps.

cadata

Distri.	Param.	τ	λ
Poisson	$\mu = 2.0$	0.46	0.1
Gamma	$s = 0.5$	3.16	0.1
Nakagami	$m = 0.5$	1.66	0.1
Weilbull	$\alpha = 1.0$	0.87	0.1

YearPredictionMSD

Distri.	Param.	τ	λ
Poisson	$\mu = 4.0$	0.87	1
Gamma	$s = 2.5$	0.87	1
Nakagami	$m = 2.0$	0.87	1
Weilbull	$\alpha = 3.0$	1.66	1

ijcnn1

Distri.	Param.	τ	λ
Poisson	$\mu = 1.0$	0.87	1
Gamma	$s = 2.0$	0.87	1
Nakagami	$m = 2.0$	0.46	1
Weilbull	$\alpha = 3.0$	0.87	1

covtype.binary

Distri.	Param.	τ	λ
Poisson	$\mu = 4.0$	0.12	0.1
Gamma	$s = 2.0$	0.12	0.1
Nakagami	$m = 1.0$	0.24	0.1
Weilbull	$\alpha = 2.0$	0.24	0.01

SUSY

Distri.	Param.	τ	λ
Poisson	$\mu = 1.0$	1.33	1
Gamma	$s = 1.5$	1.77	1
Nakagami	$m = 1.5$	1.00	1
Weilbull	$\alpha = 1.0$	5.62	1

mnist

Distri.	Param.	τ	λ
Poisson	$\mu = 4.0$	21.5	0.1
Gamma	$s = 2.0$	21.5	0.01
Nakagami	$m = 1.5$	21.5	0.01
Weilbull	$\alpha = 2.0$	40.8	0.01

acoustic

Distri.	Param.	τ	λ
Poisson	$\mu = 0.5$	1.66	1
Gamma	$s = 1.5$	1.66	1
Nakagami	$m = 1.5$	0.87	1
Weilbull	$\alpha = 1.0$	5.99	1

covtype

Distri.	Param.	τ	λ
Poisson	$\mu = 4.0$	0.12	0.1
Gamma	$s = 2.0$	0.12	0.1
Nakagami	$m = 1.0$	0.24	0.01
Weilbull	$\alpha = 2.0$	0.24	0.01

Note that we particularly include the Gaussian kernel for comparison. Unlike other kernels, this kernel does not fall within Polya's characterization, because it is not convex on $[0, \infty)$. However, ones sees that its performance is often similar to that of Laplace. In the context of random feature

maps, they are not as good as the kernels approximated by random binning.

5.3 Matrix Approximation Error v.s. Prediction Performance

A link is missing between the kernel matrix approximation error and a machine learning task performance. We have shown that random binning yields a better approximation from the matrix angle, and we have also demonstrated that it yields a better prediction performance from the regression/classification angle. The purpose of the final experiment is to show that these two metrics appear to be closely related.

For demonstration, we use the same data sets as in Section 5.1, but perform the comparison with tuned parameters obtained from the preceding subsection. In Figure 3 we plot the approximation error versus prediction performance. These curves are obtained for the same Laplace kernel approximated by different approaches. One sees a clear trend that a better kernel approximation implies a better prediction. Moreover, the curves of the two random feature maps are generally well aligned, indicating that the approximation method does not play a significant role in the relation between approximation error and prediction performance.

Despite the appealing empirical evidence that approximation and prediction performance are positively correlated, we, however, hesitate the conclude firmly the relation. The reader may notice in Figure 2 that occasionally the prediction performance degrades when D becomes too large. These scenarios occur at a large n, or a large D, that prevents us from extending the plots in Figure 3 for a more complete account. Incidentally, other work also shows that using the approximate kernel $\tilde{k}$ from random Fourier maps, it could happen that the prediction results are better compared with those of the nonapproximate kernel [4]. Such phenomena appear to be beyond explanations of existing theory on the convergence of random feature maps or on the bounds of generalization error. Further theory is yet to be developed.

6 Summary of Contributions and Conclusion

This work aims at deepening the understanding of positive-definite functions, as well as the random feature maps proposed by Rahimi and Recht [15] for training large-scale kernel machines. We highlight a few contributions in the following.

First, we reveal that the random binning feature map is closely tied to Polya's criterion, a less used characterization of kernels compared to that of Bochner's. We derive a number of novel kernel functions (12), (14), (16), (18), (20), (25), (27), (29), and (31) based on Polya's characterization, which substantially enrich the catalog of kernels applicable to kernel methods and Gaussian processes. The work [15] focuses on the generation of random feature maps given a kernel; hence, the sampling distributions are restricted to those tied to known kernels. On the other hand, we exploit the relationship between kernels and distributions on the opposite direction; and show that any distribution with a positive support corresponds to a valid kernel (Corollary 2), which allows for the construction of new kernels through applying numerous known probability distributions. Additionally, we study a few properties of the kernels constructed from Polya's characterization (Theorems 3 and 4) and derive the Fourier transforms of the constructed kernels mentioned earlier.

Second, we compare the two approaches for generating random feature maps—random Fourier and random binning—through an analysis of the Frobenius norm error of the approximate kernel matrix (Theorems 7 and 8). The analysis points to a conclusion that random binning yields a

smaller error in expectation. The difference in errors is demonstrated in Figure 1 for a few data sets. This analysis favors the random binning approach from the kernel approximation angle. Meanwhile, empirical evidences in Section 5.2 on regression/classification performance also lead to the same preference.

Third, the revealed fact that the sampling distribution of random binning is not limited to the gamma distribution of a particular shape, allows us to treat the shape as a tuning parameter for obtaining better regression/classification performance. Moreover, it also allows us to use other distributions for chasing the performance. Figure 2 and Table 2 confirm this argument.

Acknowledgment

We would like to thank Michael Stein and Haim Avron for helpful discussions. J. Chen is supported in part by the XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323. D. Cheng and Y. Liu are supported in part by the NSF Research Grant IIS-1254206 and IIS-1134990. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agency, or the U.S. Government. Part of the work was done while D. Cheng was a summer intern at IBM Research.

A Special Functions Seen in Section 3

Gamma function

$$\Gamma(s) = \int_0^\infty x^{s-1} e^{-x}\, dx, \qquad \Re(s) > 0.$$

Upper incomplete gamma function

$$\Gamma(s,t) = \int_t^\infty x^{s-1} e^{-x}\, dx, \qquad \Re(s) \geq 0.$$

Exponential integral

$$E_1(z) = \int_z^\infty \frac{e^{-t}}{t}\, dt, \qquad |\arg(z)| < \pi.$$

Kummer's confluent hypergeometric function

$$M(a,b,z) = \sum_{n=0}^{\infty} \frac{a^{(n)} z^n}{b^{(n)} n!}, \qquad \text{where } a^{(0)} = 1,\ a^{(n)} = a(a+1)(a+2)\cdots(a+n-1).$$

Error function

$$\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}\, dt.$$

Imaginary error function

$$\operatorname{erfi}(x) = -\mathbf{i} \operatorname{erf}(\mathbf{i}\, x).$$

Complementary error function

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x).$$

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] Haim Avron and Vikas Sindhwani. High-performance kernel machines with implicit distributed optimization and randomization. *Technometrics*, 58(3):341–349, 2016.

[3] Harry Bateman. *Tables of Integral Transforms*, volume I. McGraw-Hill Book Company, 1954.

[4] Jie Chen, Haim Avron, and Vikas Sindhwani. Hierarchically compositional kernels for scalable nonparametric learning. arXiv:1608.00860, 2016.

[5] Jie Chen, Lingfei Wu, Kartik Audhkhasia, Brian Kingsbury, and Bhuvana Ramabhadran. Efficient one-vs-one kernel ridge regression for speech recognition. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[6] Jean-Paul Chilès and Pierre Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, 2012.

[7] Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina Balcan, and Le Song. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems 27*, 2014.

[8] Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 4th edition, 2010.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

[10] P. Huang, H. Avron, T. N. Sainath, V. Sindhwani, and B. Ramabhadran. Kernel methods match deep neural networks on TIMIT. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[11] Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *The 15th International Conference on Artificial Intelligence and Statistics*, 2012.

[12] George S. Kimeldorf and Grace Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[13] Quoc Le, Tamas Sarlos, and Alexander Smola. Fastfood — computing hilbert space expansions in loglinear time. In *Proc. of the 30th International Conference on Machine Learning (ICML)*, 2013.

[14] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209(441–458):415–446, 1909.

[15] Ali Rahimi and Ben Recht. Random features for large-scale kernel machines. In *Neural Infomration Processing Systems*, 2007.

[16] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[17] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. *Lecture Notes in Computer Science*, 2111:416426, 2001.

[18] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2001.

[19] Michael L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

[20] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):480–492, 2012.

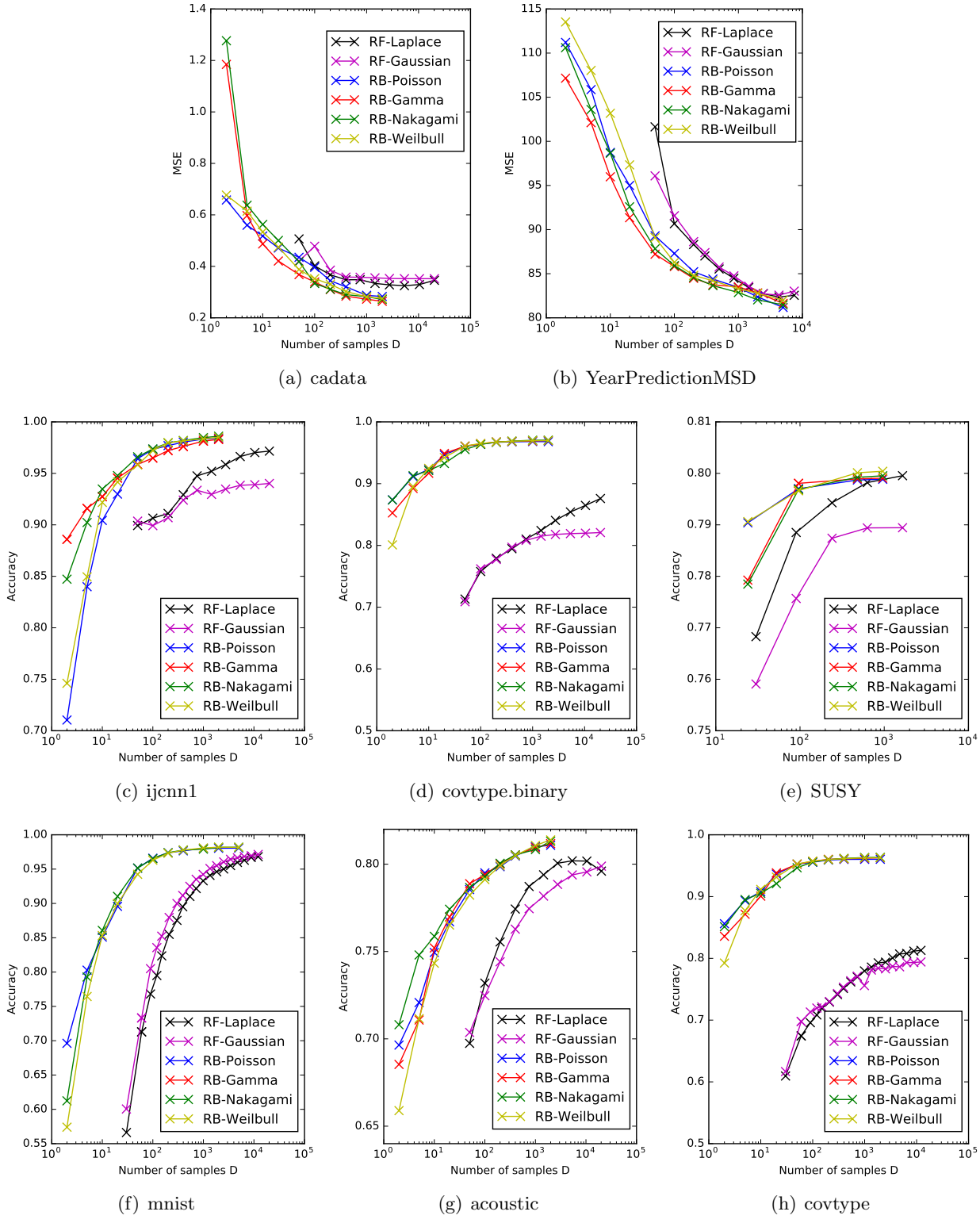
[21] Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004.

[22] Lingfei Wu, Ian E.H. Yen, Jie Chen, and Rui Yan. Revisiting random binning features: Fast convergence and strong parallelizability. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2016.

[23] J. Yang, V. Sindhwani, Q. Fan, H. Avron, and M. Mahoney. Random Laplace feature maps for semigroup kernels on histograms. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[24] Tianbao Yang, Yu feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems 25*, 2012.

[25] I. Yen, T. Lin, S. Lin, P. Ravikumar, and I. Dhillon. Sparse random feature algorithm as coordinate descent in Hilbert space. In *Neural Information Processing Systems*, pages 2456–2464, 2014.

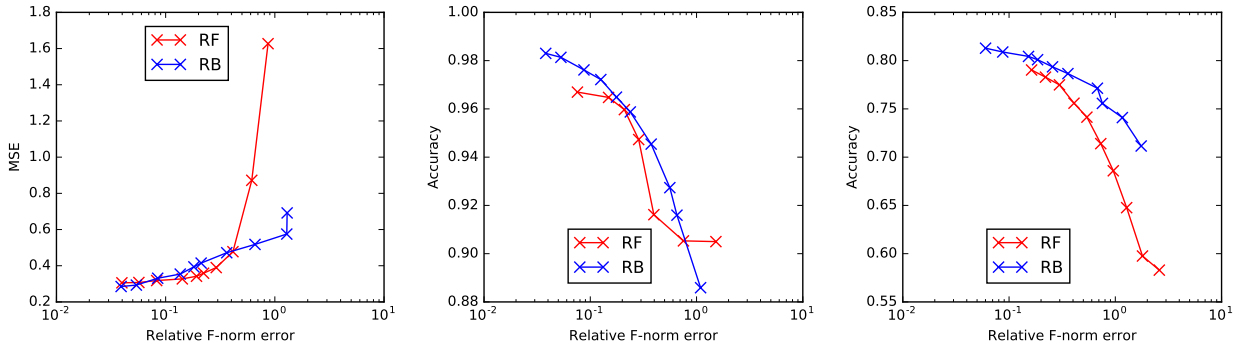


(a) cadata (b) YearPredictionMSD

(c) ijcnn1 (d) covtype.binary (e) SUSY

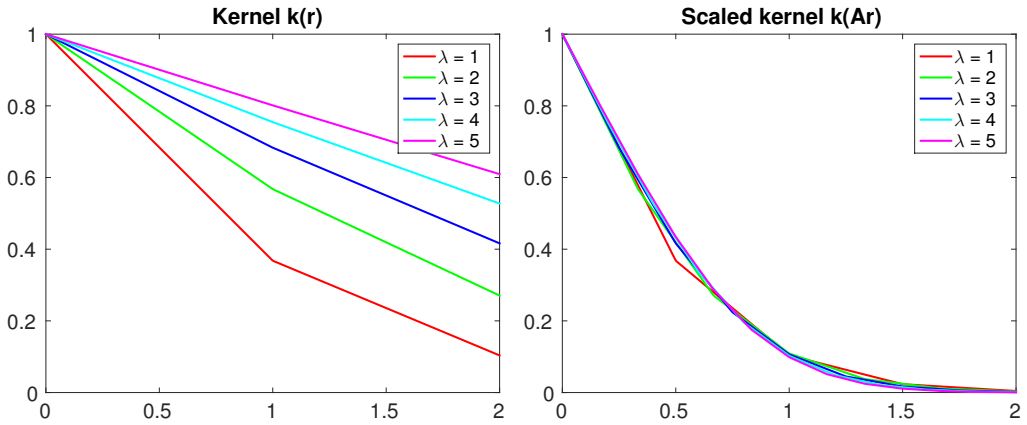
(f) mnist (g) acoustic (h) covtype

Figure 2: Regression/Classification performance as a function of sample size D. Top row: regression. Middle row: binary classification. Bottom row: multiclass classification.

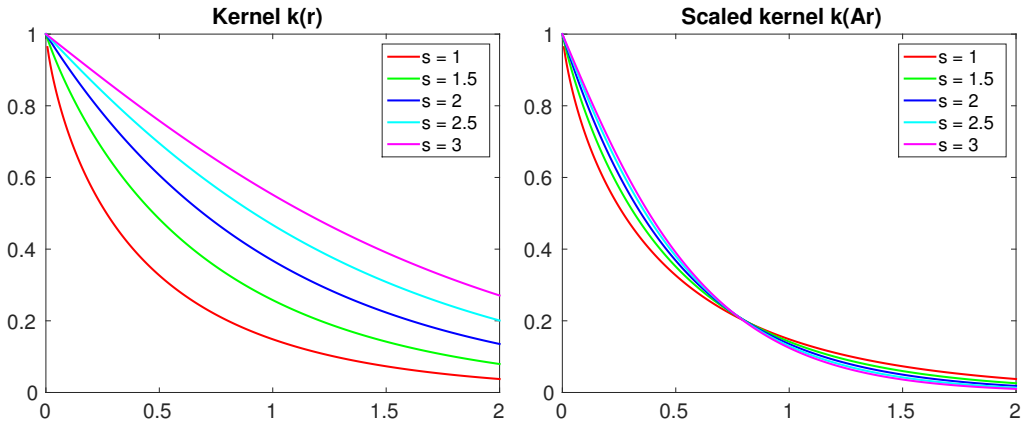


(a) cadata (b) ijcnn1 (c) acoustic

Figure 3: Matrix approximation error v.s. regression/classification performance. Laplace kernel. The two curves correspond to two methods for performing approximation.

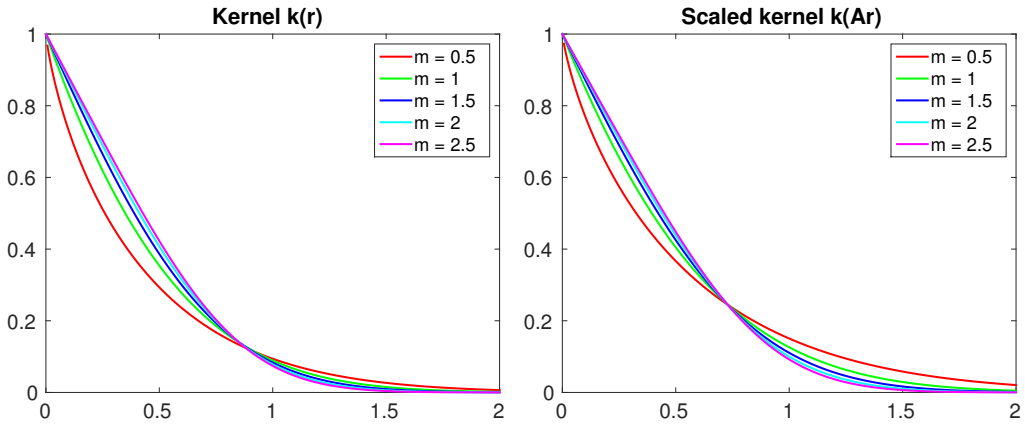


(a) $X \sim$ shifted Pois(μ)

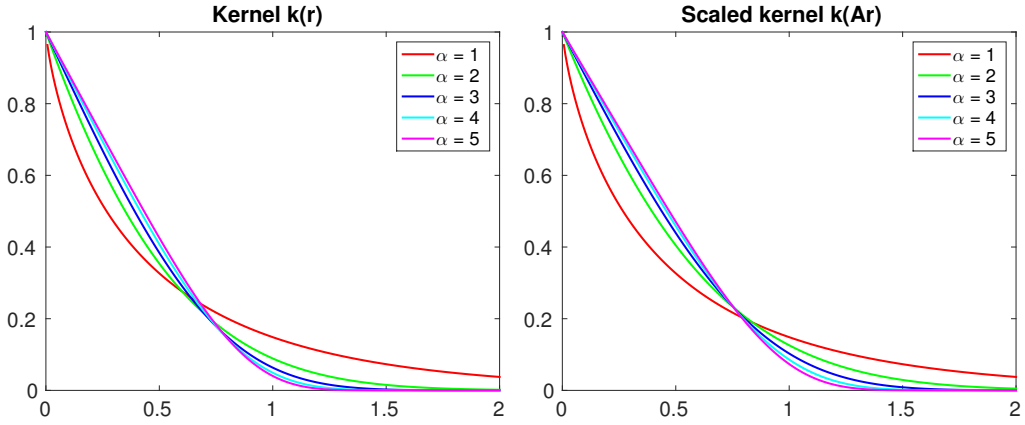


(b) $X \sim$ Gamma($s, 1$)

Figure 4: Constructed kernels from different probability distributions. Left: Unscaled. Right: Scaled by $A = E[X]$.



(a) $X \sim \text{Nakagami}(m, 1)$



(b) $X \sim \text{Weibull}(1, \alpha)$

Figure 5: (Continued from Figure 4) Constructed kernels from different probability distributions. Left: Unscaled. Right: Scaled by $A = E[X]$.

Table 3: Probability distributions and corresponding kernels

Distribution	pmf	$E[X]$	$k(r)$	$\mathcal{F}[k](t)$	Parameter choice/tuning
Shifted Pois(μ)	$\frac{\mu^{x-1}e^{-\mu}}{(x-1)!}, \quad x = 1, 2, \ldots$	$\mu + 1$	(12)	(13)	$\mu = 1, 2, 3, \ldots$
Distribution	**pdf**	$E[X]$	$k(r)$	$\mathcal{F}[k](t)$	**Parameter choice/tuning**
Gamma(s, θ)	$\frac{x^{s-1}e^{-x/\theta}}{\Gamma(s)\theta^s}$	θs	(14) $s > 1$ (16) $s = 1$	(15) $s > 1$ (17) $s = 1$	$s = \frac{1}{2}, 1, \frac{3}{2}, \ldots; \quad \theta = 1$
Nakagami(m, Ω)	$\frac{2m^m x^{2m-1} e^{-mx^2/\Omega}}{\Gamma(m)\Omega^m}$	$\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)}\left(\frac{\Omega}{m}\right)^{1/2}$	(27) $m > 1/2$ (29) $m = 1/2$	(28) $m > 1/2$ (30) $m = 1/2$	$m = \frac{1}{2}, 1, \frac{3}{2}, \ldots; \quad \Omega = 1$
Weibull(θ, α)	$\frac{\alpha}{\theta}\left(\frac{x}{\theta}\right)^{\alpha-1} e^{-(x/\theta)^\alpha}$	$\theta\Gamma(1 + 1/\alpha)$	(31) $\alpha > 1$ (16) $\alpha = 1$	(17) $\alpha = 1$	$\theta = 1; \quad \alpha = 1, 2, 3, \ldots$
Distribution	**pdf**	$E[X]$	$k(r)$	$\mathcal{F}[k](t)$	**Same as**
Exp(θ)	$\frac{1}{\theta}e^{-x/\theta}$	θ	(16)	(17)	Gamma($1, \theta$) Weibull($\theta, 1$)
χ^2_ν	$\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$	ν	(14)	(15)	Gamma($\nu/2, 2$)
χ_ν	$\frac{2^{1-\nu/2}}{\Gamma(\nu/2)}x^{\nu-1}e^{-x^2/2}$	$\sqrt{2}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}$	(18) $\nu > 1$ (20) $\nu = 1$	(19) $\nu > 1$ (21) $\nu = 1$	Nakagami($\nu/2, \nu$)
HN(σ)	$\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$	$\frac{\sigma\sqrt{2}}{\sqrt{\pi}}$	(20)	(21)	Nakagami($1/2, \sigma^2$)
Rayleigh(σ)	$\frac{x}{\sigma^2}e^{-x^2/(2\sigma^2)}$	$\sigma\sqrt{\frac{\pi}{2}}$	(25)	(26)	Nakagami($1, 2\sigma^2$) Rice($0, \sigma$)