Journal of Machine Learning Research 17 (2016) 1-28 Submitted 3/14; Revised 5/16; Published 9/16

Characteristic Kernels and Infinitely Divisible Distributions

Yu Nishiyama YNISHIYAM@GMAIL.COM
The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

Editor: Ingo Steinwart

Abstract

We connect shift-invariant characteristic kernels to infinitely divisible distributions on $\mathbb{R}^d$. Characteristic kernels play an important role in machine learning applications with their kernel means to distinguish any two probability measures. The contribution of this paper is twofold. First, we show, using the Lévy–Khintchine formula, that any shift-invariant kernel given by a bounded, continuous, and symmetric probability density function (pdf) of an infinitely divisible distribution on $\mathbb{R}^d$ is characteristic. We mention some closure properties of such characteristic kernels under addition, pointwise product, and convolution. Second, in developing various kernel mean algorithms, it is fundamental to compute the following values: (i) kernel mean values $m_P(x)$, $x \in \mathcal{X}$, and (ii) kernel mean RKHS inner products $\langle m_P, m_Q \rangle_{\mathcal{H}}$, for probability measures P, Q. If P, Q, and kernel k are Gaussians, then the computation of (i) and (ii) results in Gaussian pdfs that are tractable. We generalize this Gaussian combination to more general cases in the class of infinitely divisible distributions. We then introduce a *conjugate* kernel and a *convolution trick*, so that the above (i) and (ii) have the same pdf form, expecting tractable computation at least in some cases. As specific instances, we explore α-stable distributions and a rich class of generalized hyperbolic distributions, where the Laplace, Cauchy, and Student's t distributions are included.

Keywords: Characteristic Kernel, Kernel Mean, Infinitely Divisible Distribution, Conjugate Kernel, Convolution Trick

1. Introduction

Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space and $\mathcal{M}_1(\mathcal{X})$ be the set of probability measures. Let $\mathcal{H}$ be the real-valued reproducing kernel Hilbert space (RKHS) associated with a bounded and measurable positive-definite (p.d.) kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In machine learning, kernel methods provide a technique for developing nonlinear algorithms, by mapping data $X_1, \cdots, X_n$ in $\mathcal{X}$ to higher- or infinite-dimensional RKHS functions $k(\cdot, X_1), \ldots, k(\cdot, X_n)$ in $\mathcal{H}$ (Schölkopf and Smola, 2002; Steinwart and Christmann, 2008).

Recently, an RKHS representation of a probability measure $P \in \mathcal{M}_1(\mathcal{X})$, called kernel mean, $m_P := \mathbb{E}_{X\sim P}[k(\cdot, X)] \in \mathcal{H}$ (Smola et al., 2007; Fukumizu et al., 2013), or equivalently,

$$m_P(x) = \int k(x, y) dP(y), \quad x \in \mathcal{X} \tag{1}$$

has been used to handle probability measures in RKHSs. The kernel mean enables us to introduce a similarity and distance between two probability measures $P, Q \in \mathcal{M}_1(\mathcal{X})$, via the RKHS inner product $\langle m_P, m_Q \rangle_{\mathcal{H}}$ and the norm $||m_P - m_Q||_{\mathcal{H}}$, respectively. Using these quantities, different authors have proposed many algorithms, including density estimations (Smola et al., 2007; Song et al., 2008; McCalman et al., 2013), hypothesis tests (Gretton et al. 2012, Gretton et al. 2008, Fukumizu et al. 2008), kernel Bayesian inference (Song et al. 2009, Song et al. 2010, Song et al. 2011, Fukumizu et al. 2013, Song et al. 2013, Kanagawa et al. 2016, Nishiyama et al. 2016), classification (Muandet et al., 2012), dimension reduction (Fukumizu and Leng, 2012), and reinforcement learning (Grünewälder et al. 2012, Nishiyama et al. 2012, Rawlik et al. 2013, Boots et al. 2013).

In these applications, the characteristic property of a p.d. kernel k is important: a p.d. kernel is said to be *characteristic* if any two probability measures $P, Q \in \mathcal{M}_1(\mathcal{X})$ can be distinguished by their kernel means $m_P, m_Q \in \mathcal{H}$ (Fukumizu et al., 2004; Sriperumbudur et al., 2010, 2011). For a continuous, bounded, and shift-invariant p.d. kernel on $\mathbb{R}^d$ with $k(x, y) = \kappa(x - y)$, a necessary and sufficient condition for the kernel to be characteristic is known via the Bochner theorem (Sriperumbudur et al., 2010, Theorem 9).

As the first contribution of this paper, we show, using the Lévy–Khintchine formula (Sato, 1999; F. W. Steutel, 2004; Applebaum, 2009), that if κ is a continuous, bounded, and symmetric pdf of an infinitely divisible distribution P on $\mathbb{R}^d$, then k is a characteristic p.d. kernel. We call such kernels *convolutionally infinitely divisible* (CID) kernels. Examples of CID kernels are given in Example 3.4. In addition, we note some closure properties of the CID kernels with respect to addition, pointwise product, and convolution.

To describe the second contribution, we briefly explain what is essentially computed in kernel mean algorithms. In general kernel methods, the following computations are fundamental:

(i) RKHS function values: $f(x)$ for $f \in \mathcal{H}$, $x \in \mathcal{X}$,

(ii) RKHS inner products: $\langle f, g \rangle_{\mathcal{H}}$, $f, g \in \mathcal{H}$.

If $f \in \mathcal{H}$ is represented by $f := \sum_{i=1}^n w_i k(\cdot, X_i)$, $w \in \mathbb{R}^n$, then the function value (i) $f(x) = \sum_{i=1}^n w_i k(x, X_i)$ reduces to the evaluation of the kernel $k(x, y)$. Similarly, if two RKHS functions $f, g \in \mathcal{H}$ are both represented by $f := \sum_{i=1}^n w_i k(\cdot, X_i)$ and $g := \sum_{j=1}^l \tilde{w}_j k(\cdot, \tilde{X}_j)$, respectively, then the inner product (ii) $\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^n \sum_{j=1}^l w_i \tilde{w}_j k(X_i, \tilde{X}_j)$ reduces to the evaluation of the kernel $k(x, y)$, which is so-called the *kernel trick* $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} = k(x, y)$.

We consider a more general case in which $f, g \in \mathcal{H}$ are represented by $f := \sum_{i=1}^n w_i m_{P_i}$ and $g := \sum_{j=1}^l \tilde{w}_j m_{Q_j}$, respectively, where $\{m_{P_i}\}, \{m_{Q_j}\} \subset \mathcal{H}$ are kernel means of probability measures $\{P_i\}, \{Q_j\} \subset \mathcal{M}_1(\mathcal{X})$. Kernel algorithms involving kernel means use this type of RKHS functions explicitly or implicitly. If $\{P_i\}, \{Q_j\}$ are delta measures $\{\delta_{X_i}\}, \{\delta_{\tilde{X}_j}\}$,[1] then these functions are specialized to the above kernel trick case, where $m_{\delta_x} = k(\cdot, x)$. The computation of (i) $f(x) = \sum_{i=1}^n w_i m_{P_i}(x)$ and (ii) $\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^n \sum_{j=1}^l w_i \tilde{w}_j \langle m_{P_i}, m_{Q_j} \rangle_{\mathcal{H}}$ requires the following kernel mean evaluations:[2]

1. A probability measure $\delta_x(\cdot)$, $x \in \mathcal{X}$ is a delta measure; if $x \in B$, then $\delta_x(B) = 1$; otherwise, $\delta_x(B) = 0$ for $B \in \mathcal{B}(\mathcal{X})$.
2. If kernel means m_P, m_Q are also both expressed by a weighted sum, $m_P := \sum_{i=1}^{n_P} \eta_i k(\cdot, \dot{X}_i)$ and $m_Q := \sum_{i=1}^{n_Q} \tilde{\eta}_i k(\cdot, \ddot{X}_i)$, $\{\dot{X}_i\}, \{\ddot{X}_i\} \subset \mathcal{X}$, then the computation also reduces to the above kernel trick case.

(iii) kernel mean values: $m_P(x)$ for $P \in \mathcal{M}_1(\mathcal{X})$, $x \in \mathcal{X}$,

(iv) kernel mean inner products: $\langle m_P, m_Q \rangle_{\mathcal{H}}$, $P, Q \in \mathcal{M}_1(\mathcal{X})$.

Note that the kernel mean value (1) and the kernel mean inner product $\langle m_P, m_Q \rangle_{\mathcal{H}} = \int k(x,y) dP(x) dQ(y)$ involve an integral, and their rigorous computation is not tractable in general.

The second contribution of this paper is to provide some classes of p.d. kernels and parametric models $P, Q \in \mathcal{P}_\Theta := \{P_\theta | \theta \in \Theta\}$ such that the kernel computation of (iii) and (iv) can be reduced to a kernel evaluation, where tractable computation is considered. For a shift-invariant kernel $k(x,y) = \kappa(x-y)$, $x, y \in \mathbb{R}^d$ on $\mathbb{R}^d$, as shown in Lemma 2.5, the computation of (iii) and (iv) reduces to the following convolution:

(iii)' kernel mean values: $m_P(x) = (\kappa * P)(x)$,

(iv)' kernel mean inner products: $\langle m_P, m_Q \rangle_{\mathcal{H}} = (\kappa * \tilde{P} * Q)(0) = (\kappa * P * \tilde{Q})(0)$,

where $\tilde{P}$ and $\tilde{Q}$ are the dual of P and Q, respectively.[3] This convolution representation motivates us to explore a set of parametric distributions $\mathcal{P}_\Theta$ that is closed under convolution, namely, a convolution semigroup $(\mathcal{P}_\Theta, *) \subset \mathcal{M}_1(\mathbb{R}^d)$, where κ is a density function in $\mathcal{P}_\Theta$.

To illustrate the basic idea, let us consider Gaussian distributions $\mathcal{P}_\Theta$ as a parametric class, which is closed under convolution, and a Gaussian kernel. For simplicity, we consider the case of scalar variance matrices $\sigma^2 I_d$. Let $N_d(\mu, \sigma^2 I_d)$ and $f_d(x|\mu, \sigma^2 I_d)$ denote the d-dimensional Gaussian distribution with mean μ and variance-covariance matrix $\sigma^2 I_d$, and its pdf, respectively. If P and Q are Gaussian distributions $N_d(\mu_P, \sigma_P^2 I_d)$ and $N_d(\mu_Q, \sigma_Q^2 I_d)$, respectively, and k is given by the pdf $f_d(x-y|0, \tau^2 I)$, it is easy to see that $m_P(x) = f_d(x|\mu_P, (\sigma_P^2 + \tau^2) I_d)$ and $\langle m_P, m_Q \rangle_{\mathcal{H}} = f_d(\mu_P | \mu_Q, (\sigma_P^2 + \sigma_Q^2 + \tau^2) I_d)$. The kernel mean value and inner product are thus reduced to simply evaluating Gaussian pdfs, which are given by a parameter update following a specific rule. This type of computation appears in various applications: to list a few, Muandet et al. (2012) proposed a support measure classification by considering kernels $k(P,Q)$ between two input probability measures P, Q, including Gaussian models; Song et al. (2008) and McCalman et al. (2013) considered an approximation of a (target) probability measure P with a Gaussian mixture P_θ, via an optimization problem $\hat{\theta} = \mathrm{argmin}_\theta ||m_P - m_{P_\theta}||_{\mathcal{H}}^2$. The parametric expression of (iii) and (iv) is especially useful for the optimization of θ in the class of distributions. Other such applications are given in Section 5.

We generalize this closedness or "conjugacy[4]" of Gaussians with respect to kernel means and explore other cases in CID kernels. We then introduce a *conjugate* kernel k to parametric models $\mathcal{P}_\theta$ and a *convolution trick*, so that (iii) and (iv) have the same density form, i.e., there is some parameter update in the class. If P, Q are delta measures δ_x, δ_y, then the convolution trick simplifies to the kernel trick. See Proposition 4.2 for a description.

While a general perspective is obtained from the convolution semigroup $(\mathbb{I}(\mathbb{R}^d), *)$ of infinitely divisible distributions, the pdfs of $\mathbb{I}(\mathbb{R}^d)$ are not tractable in general. We then

3. A probability measure $\tilde{P} \in \mathcal{M}_1(\mathbb{R}^d)$ is called a *dual* of $P \in \mathcal{M}_1(\mathbb{R}^d)$ if $\tilde{P}(B) = P(-B)$ for every $B \in \mathcal{B}(\mathbb{R}^d)$, where $-B := \{-x : x \in B\}$ (Sato, 1999, p.8)

4. Here, the term "conjugacy" is an analogy of the conjugate prior in the Bayes' theorem, where the prior and posterior have the same pdf form in a probabilistic model.

explore smaller convolution sub-semigroups $(\mathcal{P}_\Theta, *) \subset (\mathbb{I}(\mathbb{R}^d), *)$ having a small number of parameters. In particular, we focus on the well-known α-stable distributions $\mathbb{S}_\alpha(\mathbb{R}^d)$ for each $\alpha \in (0, 2]$ and generalized hyperbolic (GH) distributions $\mathbb{GH}(\mathbb{R}^d)$, which include Laplace, Cauchy, and Student's t distributions. For each $\alpha \in (0, 2]$, the class $\mathbb{S}_\alpha(\mathbb{R}^d)$ is closed under convolution. The GH class has various convolutional properties, as given in Proposition 4.5. As in the Gaussian cases, the computation of (iii) and (iv) is realized by the evaluation of pdfs, i.e., evaluation of conjugate kernels, after a parameter update.

Unfortunately, these conjugate kernels are not generally tractable. However, we can find some subclasses of tractable conjugate kernels. See Section 6 for a discussion on the computation of conjugate kernels. Note that α-stable and GH distribution classes have many applications; applications of $\mathbb{S}_\alpha(\mathbb{R}^d)$ are listed in Nolan (2013a), and the GH distributions have been applied, e.g., to mathematical finance with the Lévy processes (Schoutens, 2003; Cont and Tankov, 2004; Barndorff-Nielsen and Halgreen, 1990; Madan et al., 1998; Barndorff-Nielsen, 1998; Barndorff-Nielsen and Prause, 2001; Carr et al., 2002). Note also that the Matérn kernel (Rasmussen and Williams, 2006, Section 4.2.1), often used in machine learning, is included in this GH class.

The rest of this paper is organized as follows. In Section 2, we review the notions of kernel means, characteristic kernels, and related matters. In Section 3, we show that the CID kernels are characteristic p.d. kernels on $\mathbb{R}^d$. In addition, we present the closedness property with respect to addition, pointwise product, and convolution. In Section 4, we introduce the absorbing, conjugate kernel and convolution trick for convolution semigroups of infinitely divisible distributions. Section 5 lists some motivating examples of kernel machine algorithms involving kernel means and parametric models. Section 6 notes the computation of the pdfs of conjugate kernels to realize the convolution trick.

2. Preliminaries: Kernel Means and Characteristic Kernels

In this section, we review kernel means and characteristic kernels restricted to $\mathbb{R}^d$.

Let $\mathbb{P}_d$ be the set of $d \times d$ p.d. matrices. Let $||x||_\Sigma = \sqrt{x^\top \Sigma x}$, $x \in \mathbb{R}^d$, and $\Sigma \in \mathbb{P}_d$. Let $L^1(\mathbb{R}^d)$ be the absolutely integrable function space on $\mathbb{R}^d$. Let $C_b(\mathbb{R}^d)$ be the continuous and bounded function space on $\mathbb{R}^d$.

A symmetric function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a *p.d. kernel* on $\mathbb{R}^d$ if, for any $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathbb{R}^d$, the matrix $G_{ij} = k(x_i, x_j)$, $i, j \in \{1, \ldots, n\}$ is positive-semidefinite. Throughout this paper, we assume a p.d. kernel k is on $\mathbb{R}^d$. It is known (Aronszajn, 1950) that every p.d. kernel k has the unique RKHS $\mathcal{H}$, which is a Hilbert space of functions $f : \mathbb{R}^d \to \mathbb{R}$, satisfying the following: (i) $k(\cdot, x) \in \mathcal{H}$, $\forall x \in \mathbb{R}^d$; (ii) $\mathrm{Span}\{k(\cdot, x) | x \in \mathbb{R}^d\}$ is dense in $\mathcal{H}$; and (iii) the *reproducing property* holds:

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}, \quad \forall x \in \mathbb{R}^d,$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product of $\mathcal{H}$. The map $\Phi : \mathbb{R}^d \to \mathcal{H}; x \mapsto k(\cdot, x)$ is called a *feature map*.

A p.d. kernel k is called *bounded* if $\sup_{x \in \mathbb{R}^d} k(x, x) < \infty$. A p.d. kernel k is bounded if and only if every $f \in \mathcal{H}$ is bounded (Steinwart and Christmann, 2008, Lemma 4.23). A p.d. kernel k is called *separately continuous* if $k(\cdot, x) : \mathbb{R}^d \to \mathbb{R}$ is continuous for all $x \in \mathbb{R}^d$. A p.d. kernel k is bounded and separately continuous if and only if every $f \in \mathcal{H}$ is a bounded and continuous function, i.e., $\mathcal{H} \subset C_b(\mathbb{R}^d)$, (Steinwart and Christmann, 2008, Lemma 4.28). A p.d. kernel k is called *continuous* if k is separately continuous and $x \mapsto k(x,x)$, $x \in \mathbb{R}^d$ is continuous (Steinwart and Christmann, 2008, Lemma 4.29). If a p.d. kernel k is continuous, the RKHS $\mathcal{H}$ is separable (Steinwart and Christmann, 2008, Lemma 4.33).

A p.d. kernel k is called *shift-invariant* if there exists a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that $k(x,y) = \kappa(x-y)$, $x, y \in \mathbb{R}^d$. The function κ is called a *p.d. function*. A p.d. function κ on $\mathbb{R}^d$ is characterized by the Bochner theorem:

Theorem 2.1 *(Bochner, 1959) (Wendland, 2005, Theorem 6.6) A continuous function $\kappa : \mathbb{R}^d \to \mathbb{C}$ is positive definite if and only if it is the Fourier transform $\mathcal{F}(\Lambda)$ of a finite nonnegative Borel measure Λ on $\mathbb{R}^d$:*

$$\kappa(x) = \int_{\mathbb{R}^d} e^{\sqrt{-1}w^\top x} d\Lambda(w), \quad x \in \mathbb{R}^d.$$

Let $\mathcal{K}_{cb}(\mathbb{R}^d) \subset C_b(\mathbb{R}^d)$ denote the set of continuous bounded p.d. functions.

A p.d. kernel k is called *radial* if there exists a function $\tilde{\kappa} : [0,\infty) \to \mathbb{R}$ such that $k(x,y) = \tilde{\kappa}(||x-y||)$, $x, y \in \mathbb{R}^d$. A radial kernel k is given by

$$k(x,y) = \tilde{\kappa}(||x-y||) = \int_{[0,\infty)} e^{-t||x-y||^2} d\nu(t), \ \ x, y \in \mathbb{R}^d, \tag{2}$$

where $\nu(t)$ is a finite nonnegative Borel measure on the Borel sets $\mathcal{B}([0,\infty))$. A p.d. kernel k is called *elliptical* if $k(x,y) = \tilde{\kappa}(||x-y||_\Sigma)$, $x, y \in \mathbb{R}^d$, $\Sigma \in \mathbb{P}_d$.

Let $\mathcal{M}_1(\mathbb{R}^d)$ be the set of Borel probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. An RKHS element $m_P \in \mathcal{H}$ with a p.d. kernel k is called a *kernel mean* of a probability measure $P \in \mathcal{M}_1(\mathbb{R}^d)$ if there exists an expectation of the feature map:

$$m_P := \mathbb{E}_{X\sim P}[\Phi(X)] = \mathbb{E}_{X\sim P}[k(\cdot, X)] \in \mathcal{H}, \quad P \in \mathcal{M}_1(\mathbb{R}^d).$$

If k is a bounded and continuous p.d. kernel, then the feature map $\Phi : \mathbb{R}^d \to \mathcal{H}$ is Bochner P-integrable for all $P \in \mathcal{M}_1(\mathbb{R}^d)$, since $\mathbb{E}_{X\sim P}[||k(\cdot,X)||_{\mathcal{H}}] = \mathbb{E}_{X\sim P}[\sqrt{k(X,X)}] < \infty$ for all $P \in \mathcal{M}_1(\mathbb{R}^d)$ (Steinwart and Christmann, 2008, p. 510). Throughout this paper, we assume a bounded and continuous p.d. kernel k. We write $m_{\mathcal{P}} := \{m_P | P \in \mathcal{P} \subset \mathcal{M}_1(\mathbb{R}^d)\}$.

As mentioned in the Introduction, there are many applications using m_P, since m_P enables us to introduce a similarity and distance between probability measures $P, Q \in \mathcal{M}_1(\mathbb{R}^d)$, via the Hilbert space inner product $\langle m_P, m_Q \rangle_{\mathcal{H}}$ and norm $||m_P - m_Q||_{\mathcal{H}}$, respectively, where the reproducing property is also exploited. In these applications, the characteristic kernel is important to distinguish any probability measures $P, Q \in \mathcal{M}_1(\mathbb{R}^d)$ by their kernel means $m_P, m_Q \in \mathcal{H}$. The following is the definition restricted to $\mathbb{R}^d$:

Definition 2.2 *(Fukumizu et al., 2004)(Sriperumbudur et al., 2010, Definition 6) A bounded and continuous p.d. kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called characteristic on $\mathbb{R}^d$ if the kernel mean map $\mathcal{M}_1(\mathbb{R}^d) \to \mathcal{H}$; $P \mapsto m_P$ is injective, i.e., $m_P = m_Q$ implies $P = Q$ for any $P, Q \in \mathcal{M}_1(\mathbb{R}^d)$.*

Sriperumbudur et al. (2010) showed a necessary and sufficient condition for a shift-invariant p.d. kernel $k(x,y) = \kappa(x-y)$, $x, y \in \mathbb{R}^d$, $\kappa \in \mathcal{K}_{cb}(\mathbb{R}^d)$, to be characteristic via the Bochner theorem:

Theorem 2.3 *(Sriperumbudur et al., 2010, Theorem 9) A shift-invariant p.d. kernel k with $\kappa \in \mathcal{K}_{cb}(\mathbb{R}^d)$ is characteristic if and only if the finite nonnegative measure Λ in Theorem 2.1 has the entire support,* $\mathrm{supp}(\Lambda) = \mathbb{R}^{\mathrm{d}}$.

Let $\mathcal{K}_{cb}^{ch}(\mathbb{R}^d) \subset \mathcal{K}_{cb}(\mathbb{R}^d)$ denote the set of such characteristic p.d. functions on $\mathbb{R}^d$.

The convolution $f * g$ of two functions f and g is defined by $f * g := \int_{\mathbb{R}^d} f(\cdot - y)g(y)dy$. The convolution $f * Q$ of a function f and a probability measure $Q \in \mathcal{M}_1(\mathbb{R}^d)$ is defined by $f * Q := \int_{\mathbb{R}^d} f(\cdot - y)dQ(y)$. The convolution $P * Q$ of two probability measures $P, Q \in \mathcal{M}_1(\mathbb{R}^d)$ is defined by the probability measure $(P * Q)(B) := \int_{\mathbb{R}^d} P(B - x)dQ(x)$, where $B - x := \{z - x : z \in B\}$, $B \in \mathcal{B}(\mathbb{R}^d)$.

Given a function $f(x)$, $x \in \mathbb{R}^d$, the function $\tilde{f}$ denotes $\tilde{f}(x) = f(-x)$, $x \in \mathbb{R}^d$. Given a probability measure $P \in \mathcal{M}_1(\mathbb{R}^d)$, a probability measure $\tilde{P} \in \mathcal{M}_1(\mathbb{R}^d)$ is called *dual* if $\tilde{P}(B) = P(-B)$, $B \in \mathcal{B}(\mathbb{R}^d)$, where $-B := \{-x : x \in B\}$ (Sato, 1999, p.8). A probability measure P is symmetric if $P = \tilde{P}$.

We have the following simple equalities:

Proposition 2.4 $\widetilde{f * g} = \tilde{f} * \tilde{g}$, $\widetilde{f * P} = \tilde{f} * \tilde{P}$, *and* $\widetilde{P * Q} = \tilde{P} * \tilde{Q}$.

Kernel mean m_P and RKHS inner product $\langle m_P, m_Q \rangle_{\mathcal{H}}$ have the following convolution representation:

Lemma 2.5 *Let k be a shift-invariant p.d. kernel with $\kappa \in C_b(\mathbb{R}^d)$. Then, we have the following:*

1. *Kernel mean m_P is given by the convolution*

$$m_P = \kappa * P \in \mathcal{H} \subset C_b(\mathbb{R}^d), \quad P \in \mathcal{M}_1(\mathbb{R}^d).$$

2. *The RKHS inner product $\langle m_P, m_Q \rangle_{\mathcal{H}}$ is given by the convolution*

$$\langle m_P, m_Q \rangle_{\mathcal{H}} = (\kappa * \tilde{P} * Q)(0) = (\kappa * P * \tilde{Q})(0), \quad P, Q \in \mathcal{M}_1(\mathbb{R}^d),$$

where $\tilde{P}$ and $\tilde{Q}$ are the dual of P and Q, respectively.

Proof 1. Kernel mean m_P has the following convolution representation:

$$m_P = \int_{\mathbb{R}^d} k(x, \cdot)dP(x) = \int_{\mathbb{R}^d} \kappa(\cdot - x)dP(x) = \kappa * P, \quad P \in \mathcal{M}_1(\mathbb{R}^d).$$

Kernel mean $m_P \in \mathcal{H} \subset C_b(\mathbb{R}^d)$ exists for all $P \in \mathcal{M}_1(\mathbb{R}^d)$ because, for $\kappa \in C_b(\mathbb{R}^d)$, the feature map $\Phi : x \mapsto k(x, \cdot)$ is Bochner P-integrable for all $P \in \mathcal{M}_1(\mathbb{R}^d)$, as given in the definition of m_P.

2. RKHS inner product $\langle m_P, m_Q \rangle_{\mathcal{H}}$ has the following convolution representation:

$$\langle m_P, m_Q \rangle_{\mathcal{H}} = \int_{\mathbb{R}^d} m_P(y)dQ(y) = \int_{\mathbb{R}^d} \tilde{m}_P(-y)dQ(y) = (\tilde{m}_P * Q)(0) = (\kappa * \tilde{P} * Q)(0),$$

where we have used Proposition 2.4 and $\tilde{\kappa} = \kappa$ in the last equality. Since $\langle m_P, m_Q \rangle_{\mathcal{H}}$ is symmetric with respect to P and Q, then $(\kappa * \tilde{P} * Q)(0) = (\kappa * P * \tilde{Q})(0)$. This is also obtained by $(\kappa * \tilde{P} * Q)(0) = (\kappa * \widetilde{\tilde{P} * Q})(0) = (\kappa * P * \tilde{Q})(0)$. ■

In this paper, we simply consider that κ is a pdf of a probability distribution.[5] Then, Lemma 2.5 motivates us to explore the set of probability distributions $\mathcal{P}_\Theta \subset \mathcal{M}_1(\mathbb{R}^d)$ that is closed under convolution, i.e., convolution semigroup $(\mathcal{P}_\Theta, *)$.

3. Characteristic Kernels and Infinitely Divisible Distributions

In this section, we introduce CID kernels, which are defined by infinitely divisible distributions, and show that they are characteristic (Section 3.1). In addition, we examine some closure properties of CID kernels with respect to addition, pointwise product, and convolution (Section 3.2).

3.1 Convolutionally Infinitely Divisible Kernels

We review the infinite divisibility of a probability measure (Sato, 1999; F. W. Steutel, 2004; Applebaum, 2009).

Definition 3.1 *(Sato, 1999, Definition 7.1, p. 31) A probability measure $P \in \mathcal{M}_1(\mathbb{R}^d)$ is called infinitely divisible if, for any integer $n \in \mathbb{N}$, there exists a probability measure $P_n \in \mathcal{M}_1(\mathbb{R}^d)$ such that $P = P_n^{*n}$.*

The support of every infinitely divisible distribution P is unbounded except for delta measures $\{\delta_x(\cdot) | x \in \mathbb{R}^d\}$ (Sato, 1999, Examples 7.2, p. 31). Let $\mathbb{I}(\mathbb{R}^d)$ denote the set of infinitely divisible distributions on $\mathbb{R}^d$. $\mathbb{I}(\mathbb{R}^d)$ is closed under convolution. Every infinitely divisible distribution $P \in \mathbb{I}(\mathbb{R}^d)$ has the following unique *Lévy–Khintchine representation* for the characteristic function. Let $x \wedge y = \min\{x, y\}$, $x, y \in \mathbb{R}$. Let 1_B denote the indicator function on $\mathbb{R}^d$ with $B \subset \mathbb{R}^d$.

Theorem 3.2 *(Sato, 1999, Theorem 8.1, p. 37) The characteristic function $\hat{P}(w)$ of an infinitely divisible distribution $P \in \mathbb{I}(\mathbb{R}^d)$ has the following unique representation:*

$$\hat{P}(w) = \exp\left(i w^\top \gamma - \frac{1}{2} w^\top A w + \int_{\mathbb{R}^d} \left(e^{i w^\top x} - 1 - i w^\top x 1_{\{|x| \le 1\}}(x) \right) \nu(dx) \right), \ w \in \mathbb{R}^d, \quad (3)$$

where $\gamma \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$, is a symmetric nonnegative-definite matrix and ν is a measure on $\mathbb{R}^d$ satisfying

$$\nu(\{\mathbf{0}\}) = 0 \quad \text{and} \quad \int_{\mathbb{R}^d} (|x|^2 \wedge 1) \nu(dx) < \infty. \quad (4)$$

5. In machine learning, normalized kernels $\bar{k}(x,y) := \frac{k(x,y)}{\sqrt{k(x,x)}\sqrt{k(y,y)}}$ are often used (e.g., Gaussian kernels $\bar{k}(x,y) := \exp(-\frac{||x-y||^2}{2\gamma^2})$) (Steinwart and Christmann, 2008, Lemma 4.55). However, we consider here pdf kernels (e.g., Gaussian kernels $k(x,y) := \frac{1}{\sqrt{(2\pi\gamma^2)^d}} \exp(-\frac{||x-y||^2}{2\gamma^2})$) for the closedness of the pdfs of P and m_P. A scalar multiplication ($c > 0$) changes as follows: $\bar{m}_P := \mathbb{E}_{X \sim P}[\bar{k}(\cdot, X)] = c\mathbb{E}_{X \sim P}[k(\cdot, X)] = c m_P$ and $\langle \bar{m}_P, \bar{m}_Q \rangle_{\bar{\mathcal{H}}} = c \langle m_P, m_Q \rangle_{\mathcal{H}}$, where $\langle f, g \rangle_{\bar{\mathcal{H}}} = \frac{1}{c} \langle f, g \rangle_{\mathcal{H}}$, $\forall f, g \in \mathcal{H}, \bar{\mathcal{H}}$ (Berlinet and Thomas-Agnan, 2004, p.37).

Conversely, for any $\gamma \in \mathbb{R}^d$, symmetric nonnegative-definite matrix $A \in \mathbb{R}^{d \times d}$, and measure ν satisfying (4), there exists an infinitely divisible distribution $P \in \mathbb{I}(\mathbb{R}^d)$.

(A, ν, γ) is called the *generating triplet* of $P \in \mathbb{I}(\mathbb{R}^d)$. A is called the covariance matrix of the Gaussian factor of $P \in \mathbb{I}(\mathbb{R}^d)$, and ν is called the *Lévy measure* of $P \in \mathbb{I}(\mathbb{R}^d)$. Gaussians correspond to the generating triplet $(A, 0, \gamma)$. α-Stable distributions, including Cauchy distributions, correspond to generating triplet $(0, \nu, \gamma)$, where ν is the corresponding nonzero Lévy measure. The Lévy measure of the α-stable distributions is shown in Appendix A.

An infinitely divisible distribution $P \in \mathbb{I}(\mathbb{R}^d)$ is symmetric if and only if $(A, \nu, \gamma) = (A, \nu_s, 0)$, where ν_s is a symmetric Lévy measure[6] (Sato, 1999, p.114). Let $\mathbb{IS}(\mathbb{R}^d)$ denote the set of symmetric and infinitely divisible distributions on $\mathbb{R}^d$. $\mathbb{IS}(\mathbb{R}^d)$ is closed under convolution. Let $\mathcal{K}_{cb}^{id}(\mathbb{R}^d)(\subset C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d))$ denote the set of continuous and bounded pdfs[7] of symmetric infinitely divisible distributions $\mathbb{IS}(\mathbb{R}^d)$:

$$\mathcal{K}_{cb}^{id}(\mathbb{R}^d) := \{\Xi(P_s) \in C_b(\mathbb{R}^d) | P_s \in \mathbb{IS}(\mathbb{R}^d)\},$$

where $\Xi : \mathcal{M}_1(\mathbb{R}^d) \to L^1(\mathbb{R}^d)$ is a function that maps a probability measure P to its pdf f if it exists.

The infinitely divisible pdf $\kappa \in \mathcal{K}_{cb}^{id}(\mathbb{R}^d)$ can be used for a characteristic kernel as follows.

Theorem 3.3 *The function $k(x, y) = \kappa(x - y)$, $x, y \in \mathbb{R}^d$, $\kappa \in \mathcal{K}_{cb}^{id}(\mathbb{R}^d)$ is a p.d. and characteristic kernel, i.e., $\mathcal{K}_{cb}^{id}(\mathbb{R}^d) \subset \mathcal{K}_{cb}^{ch}(\mathbb{R}^d)$.*

Proof A probability measure P on $\mathbb{R}^d$ is symmetric if and only if the characteristic function $\hat{P}(w)$, $w \in \mathbb{R}^d$ is real valued (Sato, 1999, p.67). If P is symmetric and infinitely divisible, $\hat{P}(w) > 0$ for every $w \in \mathbb{R}^d$ from the Lévy–Khintchine formula (3). Since $\hat{P}(w)$ is positive and has the entire support, $\mathrm{supp}(\hat{P}(w)) = \mathbb{R}^d$, then k is a p.d. and characteristic kernel from Theorem 2.3. ■

We call a p.d. kernel k in Theorem 3.3 a *convolutionally infinitely divisible* (CID) kernel[8]. CID kernels include the following examples:

Example 3.4 (CID p.d. kernels) *CID kernels include Gaussian kernels, Laplace kernels, Cauchy kernels, α-stable kernels for each $\alpha \in (0, 2]$ ($\alpha = 2$ corresponds to Gaussian kernels; $\alpha = 1$ corresponds to Cauchy kernels), sub-Gaussian α-stable kernels, Student's t kernels (Grosswald, 1976), GH kernels, normalized inverse Gaussian (NIG) kernels, variance gamma (VG) kernels (Matérn kernel is a special case of this), tempered α-stable (TαS) kernels (Rachev et al., 2011; Rosiński, 2007; Bianchi et al., 2010), etc.*

6. A symmetric Lévy measure is a Lévy measure such that $\nu_s(B) = \nu_s(-B)$ for $\forall B \in \mathcal{B}(\mathbb{R}^d)$.
7. A necessary and sufficient condition for $P \in \mathbb{IS}(\mathbb{R}^d)$ to have a pdf is not known (Sato, 1999, p.177). If the Gaussian factor $A \in \mathbb{R}^{d \times d}$ is full rank, then $P \in \mathbb{I}(\mathbb{R}^d)$ has the pdf. If $A = 0$, see some sufficient conditions (Sato, 1999, Theorem 27.7, 27.10). Every nondegenerate self-decomposable distribution on $\mathbb{R}^d$ has the pdf (Sato, 1999, Theorem 27.13).
8. The term "infinite divisibility" of a p.d. kernel is used in the pointwise product sense (Berg et al., 1984, Definition 2.6, p. 76), i.e., a p.d. kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ on a nonempty set $\mathcal{X}$ is called *infinitely divisible* if, for every $n \in \mathbb{N}$, there exists a p.d. kernel $k_n : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ such that $k = (k_n)^n$. The CID kernel considered here is the convolution sense $\kappa = (\kappa_n)^{*n}$.

3.2 Closure Property

In this subsection, we note some closure properties of CID and characteristic kernels with respect to addition, pointwise product, and convolution. The closure property is used, e.g., to generate a new CID and characteristic kernel. Example 3.8 shows such an example.

It is known that the set of continuous and bounded p.d. kernels $\mathcal{K}_{cb}(\mathbb{R}^d)$ is closed under addition and pointwise product as follows (Steinwart and Christmann, 2008, p. 114):

Proposition 3.5 *If $\kappa_1, \kappa_2 \in \mathcal{K}_{cb}(\mathbb{R}^d)$, then $\kappa_1 + \kappa_2 \in \mathcal{K}_{cb}(\mathbb{R}^d)$ and $\kappa_1\kappa_2 \in \mathcal{K}_{cb}(\mathbb{R}^d)$.*

Similarly, the set of characteristic kernels $\mathcal{K}_{cb}^{ch}(\mathbb{R}^d)$ is closed under addition and pointwise product as follows (Sriperumbudur et al., 2010, Corollary 11):

Proposition 3.6 *If $\kappa \in \mathcal{K}_{cb}^{ch}(\mathbb{R}^d)$, $\kappa_1, \kappa_2 \in \mathcal{K}_{cb}(\mathbb{R}^d)$, and $\kappa_2 \neq 0$, then $\kappa + \kappa_1, \kappa\kappa_2 \in \mathcal{K}_{cb}^{ch}(\mathbb{R}^d)$.*

The set of CID kernels $\mathcal{K}_{cb}^{id}(\mathbb{R}^d)$ is closed under convolution but not closed under addition or pointwise product.

Proposition 3.7 *Let $\kappa_1, \kappa_2 \in \mathcal{K}_{cb}^{id}(\mathbb{R}^d)$. Then, we have the following:*

1. *Convolution $\kappa_1 * \kappa_2 \in \mathcal{K}_{cb}^{id}(\mathbb{R}^d)$.*
2. *Addition $\kappa_1 + \kappa_2$ and product $\kappa_1\kappa_2$ do not necessarily belong to $\mathcal{K}_{cb}^{id}(\mathbb{R}^d)$, although they are characteristic, $\kappa_1 + \kappa_2, \kappa_1\kappa_2 \in \mathcal{K}_{cb}^{ch}(\mathbb{R}^d)$.*

Proof 1. Let $\kappa_1 = \Xi(P_1)$ and $\kappa_2 = \Xi(P_2)$. Then, $\kappa_1 * \kappa_2 = \Xi(P_1 * P_2)$. If $P_1, P_2 \in \mathbb{IS}(\mathbb{R}^d)$ are absolutely continuous and symmetric infinitely divisible measures, so is $P_1 * P_2 \in \mathbb{IS}(\mathbb{R}^d)$.

2. A mixture of two infinitely divisible distributions is not necessarily infinitely divisible. A product of two infinitely divisible distributions is not necessarily infinitely divisible. The counter-examples are as follows. Let $\kappa_1(x) = e^{-|x|}$ and $\kappa_2(x) = e^{-x^2}$, $x \in \mathbb{R}$, be p.d. functions of Laplace and Gaussian kernels, respectively. Then, the product $k(x) \propto e^{-|x|}e^{-x^2}$ is not infinitely divisible (F. W. Steutel, 2004, Example 11.13), although it is characteristic (Proposition 3.6). Let $\kappa_1(x) = \frac{1}{4\sqrt{\pi}}e^{-\frac{1}{4}x^2}$ and $\kappa_2(x) = \frac{1}{4\sqrt{2\pi}}e^{-\frac{1}{8}x^2}$, $x \in \mathbb{R}$, be Gaussian kernels; then, the addition $\kappa_1 + \kappa_2$ is not infinitely divisible (F. W. Steutel, 2004, Example 11.15), although it is characteristic (Proposition 3.6). Many examples can be found in F. W. Steutel (2004). ■

As given in Proposition 3.7, the infinite divisibility is not closed under mixing in general, although some special mixing cases preserve it (F. W. Steutel, 2004, Chapter 7). The *normal mean-variance mixture* with an infinitely divisible mixing distribution, given in Lemma 4.4, is one of them.

New CID kernels and characteristic kernels may be generated by using these closure properties. If $\kappa = \mathcal{F}(\hat{\kappa})$ is an infinitely divisible pdf with the characteristic function $\hat{\kappa}$, then symmetrization $\kappa^* := \kappa * \tilde{\kappa} = \mathcal{F}(|\hat{\kappa}|^2)$ and positive powers $(\kappa^*)^{*\lambda} = \mathcal{F}(|\hat{\kappa}|^{2\lambda})$ $(\lambda > 0)$ are also infinitely divisible pdfs. The following example shows that the Laplace and symmetric Gamma kernels are CID kernels generated from an exponential distribution.

Example 3.8 *(F. W. Steutel, 2004, Example 2.9) An exponential distribution P with the pdf $\kappa(x) = \alpha\exp(-\alpha x)1_{[0,\infty)}(x)$, $\alpha > 0$ is infinitely divisible. The dual is $\tilde{\kappa}(x) = \alpha\exp(\alpha x)1_{(-\infty,0)}(x)$.*

1. *The symmetrization* $\kappa^* = \kappa * \tilde{\kappa}$ *has the characteristic function* $\hat{\kappa}^*(w) = \hat{\kappa}(w)\hat{\tilde{\kappa}}(w) = \frac{\alpha}{\alpha - iw} \cdot \frac{\alpha}{\alpha + iw} = \frac{\alpha^2}{\alpha^2 + w^2}$. *This is a Laplace pdf* $\kappa^*(x) = \frac{\alpha}{2}\exp(-\alpha|x|)$.

2. *Positive powers* $(\kappa^*)^{*\lambda}$ $(\lambda > 0)$ *have the characteristic functions* $(\hat{\kappa}^*)^{\lambda}(w) = (\frac{\alpha^2}{\alpha^2+w^2})^{\lambda}$. *If* $\lambda = 1$, *the pdf is the above Laplace case. If* $\lambda = 2$, *the pdf is given by* $(\kappa^*)^{*2}(x) = \frac{\alpha}{4}(1 + \alpha|x|)\exp(-\alpha|x|)$. *For general* $\lambda > 0$, *the pdf is given by*

$$f(x) = \frac{\alpha^{2\lambda}}{\sqrt{\pi}(2\alpha)^{\lambda-\frac{1}{2}}\Gamma(\lambda)}|x-\mu|^{\lambda-\frac{1}{2}}K_{\lambda-\frac{1}{2}}(\alpha|x-\mu|), \;\; x \in \mathbb{R}$$

   where $\Gamma(\lambda)$ *is the Gamma function and* $K_\lambda(x)$ *is the modified Bessel function of the third kind with index* λ. *This is the pdf of the zero-skewed VG distribution* $VG_1(\lambda, \alpha, \beta = 0, \mu, 1)$ *on* $\mathbb{R}$, *as given in Section 4.3.*

The additions $(\kappa^*)^{*\lambda} + \tilde{\kappa}$, $\tilde{\kappa} \in \mathcal{K}_{cb}(\mathbb{R}^d)$, *and products* $(k^*)^{*\lambda}\tilde{\kappa}$, $\tilde{\kappa} \in \mathcal{K}_{cb}^{ch}(\mathbb{R}^d)$, *are characteristic kernels based on the closure properties.*

4. Kernel Means and Infinitely Divisible Distributions

In this section, we examine the kernel means of a parametric class of distributions $\mathcal{P}_\Theta \subset \mathbb{I}(\mathbb{R}^d)$. As mentioned in the Introduction, we wish to compute (iii) kernel mean values $m_P(x)$, $x \in \mathbb{R}^d$ and (iv) RKHS inner products $\langle m_P, m_Q \rangle_{\mathcal{H}}$ for parametric models $P, Q \in \mathcal{P}_\Theta$. These form a basic computation for establishing kernel machine algorithms combining kernel means and parametric models. In Section 4.1, we introduce absorbing, conjugate kernels, and convolution trick in the set of infinitely divisible distributions $\mathbb{I}(\mathbb{R}^d)$. In Sections 4.2 and 4.3, we focus on well-known subclasses of α-stable distributions and GH distributions, which include Laplace, Cauchy, and Student's t distributions.

4.1 Absorbing, Conjugate Kernels, and Convolution Trick

We begin by introducing the notion of *absorbing* and *conjugate* p.d kernels to particular sets of parametric models $\mathcal{P}_\Theta$ as follows:

Proposition 4.1 (absorbing & conjugate kernel) *Let* $\mathcal{P}_\Theta, \mathcal{Q}_{\Theta'} \subset \mathcal{M}_1(\mathbb{R}^d)$ *be two sets of parametric models such that* $\mathcal{P}_\Theta * \mathcal{Q}_{\Theta'} \subseteq \mathcal{P}_\Theta$, *where* Θ *and* Θ' *are finite or infinite index sets. Denote by* $\Xi(\mathcal{P}_\Theta)$ *and* $\Xi(\mathcal{Q}_{\Theta'})$ *the sets of pdfs. Let* $\kappa \in \mathcal{K}_{cb}(\mathbb{R}^d)$ *be a shift-invariant p.d. kernel. We have the following statements:*

1. *If* $\kappa \in \Xi(\mathcal{P}_\Theta)$, *then* $m_{\mathcal{Q}_{\Theta'}} \subset \Xi(\mathcal{P}_\Theta)$ *holds. RKHS inner products* $\langle m_P, m_Q \rangle_{\mathcal{H}}$, $P, Q \in \mathcal{Q}_{\Theta'}$ *are values of pdfs in* $\Xi(\mathcal{P}_\Theta)$.

2. *If* $\kappa \in \Xi(\mathcal{Q}_{\Theta'})$, *then* $m_{\mathcal{P}_\Theta} \subset \Xi(\mathcal{P}_\Theta)$ *holds. RKHS inner products* $\langle m_P, m_Q \rangle_{\mathcal{H}}$, $P, Q \in \mathcal{P}_\Theta$ *are not necessarily values of pdfs in* $\Xi(\mathcal{P}_\Theta)$.

Proof These statements are straightforward from Lemma 2.5 and assumptions. ■

Statements 1 and 2 indicate an *absorbing property* of k with respect to parametric models. If $\mathcal{P}_\Theta = \mathcal{Q}_{\Theta'}$ in Proposition 4.1, we call k (and, hence, its RKHS $\mathcal{H}$) a *conjugate* to $\mathcal{P}_\Theta$. A general perspective may be given by the CID kernels, where these kernels are conjugate to $\mathbb{I}(\mathbb{R}^d)$ as follows:

Proposition 4.2 *Let $k_{A,\nu_s}(x,y) = \kappa_{A,\nu_s}(x-y)$, $x, y \in \mathbb{R}^d$ be a CID kernel, where $\kappa_{A,\nu_s} \in \mathcal{K}_{cb}^{id}(\mathbb{R}^d)$ has a generating triplet $(A, \nu_s, 0)$, and let $\mathcal{H}_{A,\nu_s}$ be the RKHS given by κ_{A,ν_s}. Let $P, Q \in \mathbb{I}(\mathbb{R}^d)$ be infinitely divisible distributions with the generating triplets (A_P, ν_P, γ_P) and (A_Q, ν_Q, γ_Q), respectively. Then, we have the following:*

1. *Kernel mean m_P is given by an infinitely divisible pdf:*

$$\begin{aligned} m_P(\cdot) &= f(\cdot; A + A_P, \nu_s + \nu_P, \gamma_P), \quad f \in \Xi(\mathbb{I}(\mathbb{R}^d)) \\ &= k_{A+A_P, \nu_s+\nu_P}(\gamma_P, \cdot). \end{aligned}$$

2. *The RKHS inner product $\langle m_P, m_Q \rangle_{\mathcal{H}_{A,\nu_s}}$ is given by*

$$\begin{aligned} \langle m_P, m_Q \rangle_{\mathcal{H}_{A,\nu_s}} &= f(0; A + A_P + A_Q, \nu_s + \tilde{\nu}_P + \nu_Q, \gamma_Q - \gamma_P) \\ &= f(0; A + A_P + A_Q, \nu_s + \nu_P + \tilde{\nu}_Q, \gamma_P - \gamma_Q), \\ &= k_{A+A_P+A_Q, \nu_s+\nu_P+\tilde{\nu}_Q}(\gamma_P, \gamma_Q), \end{aligned}$$

where $\tilde{\nu}_P$ (respectively, $\tilde{\nu}_Q$) is the dual of the Lévy measure ν_P (respectively, ν_Q).

Proposition 4.2 indicates a general *convolution trick*. The computation of $\langle m_P, m_Q \rangle_{\mathcal{H}_{A,\nu_s}}$ is reduced to the computation of the same kernel $k_{A+A_P+A_Q, \nu_s+\nu_P+\tilde{\nu}_Q}$ with the updated parameters of the generating triplets. If Q is a delta measure δ_y (i.e., $A_Q = 0$, $\nu_Q = 0$, $\gamma_Q = y$), then statement 2 is specialized to statement 1. If P, Q are both delta measures δ_x, δ_y (i.e., $A_P = A_Q = 0$, $\nu_P = \nu_Q = 0$, $\gamma_P = x$, $\gamma_Q = y$), then statement 2 is specialized to the kernel trick $\langle k_{A,\nu_s}(\cdot, x), k_{A,\nu_s}(\cdot, y) \rangle_{\mathcal{H}_{A,\nu_s}} = k_{A,\nu_s}(x, y)$. If P, Q and k are all Gaussians (i.e., $\nu_P = \nu_Q = \nu_s = 0$), then statement 2 results in the computation of the same Gaussian kernel with increased variance $A + A_P + A_Q$, where the computation of Gaussian pdfs is tractable.

Although Proposition 4.2 gives us a theory that kernel means $m_{\mathcal{P}}$ and RKHS inner products $\langle m_P, m_Q \rangle$ are expressed with generating triplets (A, ν, γ), the computation of the general infinitely divisible pdfs may be intractable. We then systematically examine smaller subsemigroups of parametric models $(\mathcal{P}_\Theta, *) \subset (\mathbb{I}(\mathbb{R}^d), *)$ such that the computation of pdfs may be possible. We specifically examine well-known parametric classes of α-stable distributions and GH distributions on $\mathbb{R}^d$ in Sections 4.2 and 4.3, respectively.

4.2 α-stable distributions

α-Stable distributions $\mathbb{S}_\alpha(\mathbb{R}^d)$, $\alpha \in (0, 2]$, on $\mathbb{R}^d$ are a well-known convolution subsemigroup of infinitely divisible distributions (Zolotarev, 1986; Samorodnitsky and Taqqu, 1994).

$\alpha = 2$ implies Gaussian distributions $\mathbb{S}_2(\mathbb{R}^d) = \mathbb{G}(\mathbb{R}^d)$, which are closed under convolution; if P and Q are $N(\mu_P, R_P)$ and $N(\mu_Q, R_Q)$ with mean vectors μ_P, μ_Q and covariance matrices R_P, R_Q, respectively, then convolution $P * Q$ is $N(\mu_P + \mu_Q, R_P + R_Q)$.

For $\alpha \in (0, 2)$, α-stable distributions are heavy tailed, where there are many applications, as listed in Nolan (2013a). For each $\alpha \in (0, 2)$, a one-dimensional α-stable distribution $S_\alpha(\sigma, \beta, \mu)$ is specified by a scale parameter $\sigma > 0$, a skewness parameter $\beta \in [-1, 1]$, and a location parameter $\mu \in \mathbb{R}$. For each $\alpha \in (0, 2)$, the set $\mathbb{S}_\alpha(\mathbb{R})$ is closed under convolution;

if P and Q are two stable laws $S_\alpha(\sigma_P, \beta_P, \mu_P)$ and $S_\alpha(\sigma_Q, \beta_Q, \mu_Q)$, respectively, then $P * Q$ is $S_\alpha(\sigma, \beta, \mu) = S_\alpha((\sigma_P^\alpha + \sigma_Q^\alpha)^{1/\alpha}, \frac{\beta_P \sigma_P^\alpha + \beta_Q \sigma_Q^\alpha}{\sigma_P^\alpha + \sigma_Q^\alpha}, \mu_P + \mu_Q)$ (Samorodnitsky and Taqqu, 1994, Property1.2.1). See Appendix A.2 for more details.

For each $\alpha \in (0, 2)$, a d-dimensional α-stable distribution $S_\alpha(\mu, \Gamma)$ is specified by a location parameter $\mu \in \mathbb{R}^d$ and a spectral measure Γ on the unit sphere $S_{d-1} := \{s \in \mathbb{R}^d : ||s|| = 1\}$ (Samorodnitsky and Taqqu, 1994, Theorem 2.3.1, p.65). For each $\alpha \in (0, 2)$, the set $\mathbb{S}_\alpha(\mathbb{R}^d)$ is closed under convolution; if P and Q are two stable laws $S_\alpha(\mu_P, \Gamma_P)$ and $S_\alpha(\mu_Q, \Gamma_Q)$, respectively, then $P * Q$ is $S_\alpha(\mu_P + \mu_Q, \Gamma_P + \Gamma_Q)$. See Appendix A.1 for more details. α-Stable pdfs on $\mathbb{R}^d$ are intractable in general.

Sub-Gaussian α-stable distributions (equivalently, elliptically contoured α-stable distributions) $\mathbb{SG}_\alpha(\mathbb{R}^d)$ are a well-known subclass of $\mathbb{S}_\alpha(\mathbb{R}^d)$ (Samorodnitsky and Taqqu, 1994; Nolan, 2013b). For each $\alpha \in (0, 2)$, a sub-Gaussian α-stable distribution is specified by a location parameter $\mu \in \mathbb{R}^d$ and a p.d. matrix $R \in \mathbb{R}^{d \times d}$ (Samorodnitsky and Taqqu, 1994, Theorem 2.5.2, p.78). See Appendix A.4 for more details. Sub-Gaussian 1-stable distributions imply d-dimensional Cauchy distributions $\mathbb{CAU}(\mathbb{R}^d)$ (Samorodnitsky and Taqqu, 1994, Example 2.5.3, p.79). If $d = 1$, for each $\alpha \in (0, 2)$, sub-Gaussians $\mathbb{SG}_\alpha(\mathbb{R})$ are closed under convolution. If $d > 1$, for each $\alpha \in (0, 2)$, sub-Gaussians $\mathbb{SG}_\alpha(\mathbb{R}^d)$ are not closed under convolution. Let us decompose $\mathbb{SG}_\alpha(\mathbb{R}^d)$ into an equivalent class $\mathbb{SG}_\alpha(\mathbb{R}^d) = \bigcup_R \mathbb{SG}_\alpha(\mathbb{R}^d)[R]$, where

$$\mathbb{SG}_\alpha(\mathbb{R}^d)[R] := \{P \in \mathbb{SG}_\alpha(\mathbb{R}^d) \mid P = SG_\alpha(\mu, cR), \mu \in \mathbb{R}^d,\ c > 0\}.$$

For each $\alpha \in (0, 2)$ and a p.d. matrix $R \in \mathbb{P}^d$, the set $\mathbb{SG}_\alpha(\mathbb{R}^d)[R]$ is closed under convolution; if P and Q are $SG_\alpha(\mu_P, c_P R)$ and $SG_\alpha(\mu_Q, c_Q R)$, respectively, then $P * Q$ is $SG_\alpha(\mu_P + \mu_Q, (c_P^{\frac{\alpha}{2}} + c_Q^{\frac{\alpha}{2}})^{\frac{2}{\alpha}} R)$. Note that when $\alpha = 2$, the whole set $\mathbb{SG}_2(\mathbb{R}^d)$ is closed.

These convolution properties of α-stable distributions lead to the following conjugate pairs of α-stable kernels k and α-stable distributions $\mathcal{P}_\Theta$.

Example 4.3 *Conjugate pairs of α-stable kernels k and α-stable distributions on $\mathbb{R}^d$.*

1. *For $\alpha = 2$, let $k_R(x, y) = \frac{1}{\sqrt{(2\pi)^d |R|}} \exp(-\frac{1}{2}(x - y)^\top R^{-1}(x - y))$ be a Gaussian kernel and $\mathcal{H}_R$ be its RKHS. Let P, Q be two Gaussians $N(\mu_P, R_P)$ and $N(\mu_Q, R_Q)$, respectively. Then, the kernel mean is given by the Gaussian pdf $m_P = f_\alpha(\cdot|\mu_P, R + R_P)$ and the RKHS inner product is given by the Gaussian pdf $\langle m_P, m_Q \rangle_{\mathcal{H}_R} = f(\mu_P|\mu_Q, R + R_P + R_Q)$.*

2. *For each $\alpha \in (0, 2)$, let $k_{\alpha,\sigma}(x, y) = \kappa_{\alpha,\sigma}(x - y)$, $x, y \in \mathbb{R}$, be an α-stable kernel on $\mathbb{R}$ and $\mathcal{H}_{\alpha,\sigma}$ be its RKHS. Let P, Q be two α-stable laws $S_\alpha(\sigma_P, \beta_P, \mu_P)$ and $S_\alpha(\sigma_Q, \beta_Q, \mu_Q)$, respectively, on $\mathbb{R}$. Then, the kernel mean is given by the stable pdf $m_P = f_\alpha(\cdot|(\sigma_P^\alpha + \sigma^\alpha)^{1/\alpha}, \frac{\beta_P \sigma_P^\alpha}{\sigma_P^\alpha + \sigma^\alpha}, \mu_P)$ and the RKHS inner product is given by the stable pdf $\langle m_P, m_Q \rangle_{\mathcal{H}_{\alpha,\sigma}} = f_\alpha(\mu_P|(\sigma_P^\alpha + \sigma_Q^\alpha + \sigma^\alpha)^{1/\alpha}, \frac{\beta_Q \sigma_Q^\alpha - \beta_P \sigma_P^\alpha}{\sigma_Q^\alpha + \sigma_P^\alpha + \sigma^\alpha}, \mu_Q)$. If $\alpha = 1$ and $\beta = 0$, then $S_1(\sigma, 0, \mu)$ corresponds to the Cauchy distribution.*

3. *For each $\alpha \in (0, 2)$, let $k_{\alpha,\Gamma_s}(x, y) = \kappa_{\alpha,\Gamma_s}(x - y)$, $x, y \in \mathbb{R}^d$, be an α-stable kernel on $\mathbb{R}^d$, where Γ_s is a symmetric spectral measure, and let $\mathcal{H}_{\alpha,\Gamma_s}$ be its RKHS. Let P, Q be*

two α-stable laws $S_\alpha(\mu_P, \Gamma_P)$ and $S_\alpha(\mu_Q, \Gamma_Q)$, respectively, on $\mathbb{R}^d$. Then, the kernel mean is given by the stable pdf $m_P = f_\alpha(\cdot|\mu_P, \Gamma_P + \Gamma_s)$ and the RKHS inner product is given by the stable pdf $\langle m_P, m_Q\rangle_{\mathcal{H}_{\alpha,\sigma}} = f_\alpha(\mu_P|\mu_Q, \Gamma_Q + \tilde{\Gamma}_P + \Gamma_s)$.

4. *For each $\alpha \in (0, 2)$, let $k_{\alpha,R}(x, y) = \kappa_{\alpha,R}(x-y)$, $x, y \in \mathbb{R}^d$ be a sub-Gaussian α-stable kernel on $\mathbb{R}^d$ and let $\mathcal{H}_{\alpha,R}$ be its RKHS. Let $P, Q \in \mathbb{SG}_\alpha(\mathbb{R}^d)[R]$ be two sub-Gaussian α-stable laws $S_\alpha(\mu_P, c_P R)$ and $S_\alpha(\mu_Q, c_Q R)$, respectively, on $\mathbb{R}^d$. Then, the kernel mean is given by the sub-Gaussian pdf $m_P = f_\alpha(\cdot|\mu_P, (c_P^{\frac{\alpha}{2}}+1)^{\frac{2}{\alpha}} R)$ and the RKHS inner product is given by the sub-Gaussian pdf $\langle m_P, m_Q\rangle_{\mathcal{H}_{\alpha,R}} = f_\alpha(\mu_P|\mu_Q, (c_P^{\frac{\alpha}{2}}+c_Q^{\frac{\alpha}{2}}+1)^{\frac{2}{\alpha}} R)$. If $\alpha = 1$, then $S_1(\mu, R)$ corresponds to multivariate Cauchy distributions with pdf $f(x) \propto (1 + ||x-\mu||^2_{R^{-1}})^{-\frac{d+1}{2}}$.*

5. *Tempered stable distributions can also be considered as examples (Rachev et al., 2011, Table 3.2, p. 77).*

4.3 Generalized Hyperbolic Distributions

GH distributions on $\mathbb{R}^d$ are a rich model class that includes, e.g., NIGs, hyperbolic distributions, VG distributions, Laplace distributions, Cauchy distributions, and Student's t distributions, as special cases and limiting cases (Barndorff-Nielsen and Halgreen, 1977; Prause, 1999; v. Hammerstein, 2010). A list of parametric models is found in, e.g., Prause (1999, Table 1.1 p.4). The GH and related models are applied, e.g., to mathematical finance (Schoutens, 2003; Cont and Tankov, 2004; Barndorff-Nielsen and Halgreen, 1990; Madan et al., 1998; Barndorff-Nielsen, 1998; Barndorff-Nielsen and Prause, 2001; Carr et al., 2002). The Matérn kernel, often used in machine learning, is a special case of the VG distributions. A GH distribution is obtained by a *normal mean-variance mixture* of a generalized inverse Gaussian (GIG) distribution, which is a special case of the normal mean-variance mixture of the generalized Γ-convolution (Thorin, 1978). The pdfs of GIG, GH, NIG, and VG distributions are presented in Appendix B.

We start by introducing a normal mean-variance mixture distribution. Let $N_d(\mu, \Delta)$ be a Gaussian distribution with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Delta \in \mathbb{P}^d$. A *normal mean-variance mixture* distribution P on $\mathbb{R}^d$ is given by

$$P(dx) = \int_{\mathbb{R}^+} N_d(\mu + y\beta, y\Delta)(dx)G(dy), \ \ \beta \in \mathbb{R}^d,$$

where G is a mixing probability measure on $\mathbb{R}^+$ (v. Hammerstein, 2010, Definition 2.4, p. 78). $P = N_d(\mu + y\beta, y\Delta) \circ G$ denotes a simple notation. The closure properties of the convolution and the infinite divisibility of G are preserved as follows:

Lemma 4.4 *(v. Hammerstein, 2010, Lemma 2.5, p. 68) Let $\mathbb{G}$ be a class of probability distributions on $(\mathbb{R}^+, \mathcal{B}^+)$ and $G, G_1, G_2 \in \mathbb{G}$.*

1. *If $G = G_1 * G_2 \in \mathbb{G}$, then*

$$(N_d(\mu_1 + y\beta, y\Delta) \circ G_1) * (N_d(\mu_2 + y\beta, y\Delta) \circ G_2) = N_d(\mu_1 + \mu_2 + y\beta, y\Delta) \circ G.$$

2. *If G is infinitely divisible, then so is $N_d(\mu + y\beta, y\Delta) \circ G$.*

A GH distribution on $\mathbb{R}^d$ is given by a normal mean-variance mixture with the GIG distribution:

$$GH_d(\lambda, \alpha, \beta, \delta, \mu, \Delta) := N_d(\mu + y\Delta\beta, y\Delta) \circ GIG(\lambda, \delta, \sqrt{\alpha^2 - ||\beta||^2_\Delta}),$$

where the parameters imply $\lambda \in \mathbb{R}$, shape parameter $\alpha > 0$, skewness parameter β, scaling parameter δ, location parameter μ, and p.d. matrix $\Delta \in \mathbb{P}^d$ (see Appendices B.1 and B.2 for more details). A univariate GH distribution on $\mathbb{R}$ is given by letting $d = 1$ and $\Delta = 1$.

The GH distribution contains the following subclasses and limiting cases. Their pdfs are found in Appendices B.3, B.4, and v. Hammerstein (2010)).

1. If $\lambda = -\frac{1}{2}$, then $GH_d(-\frac{1}{2}, \alpha, \beta, \delta, \mu, \Delta))$ corresponds to the NIG distribution:

$$NIG_d(\alpha, \beta, \delta, \mu, \Delta) := N_d(\mu + y\Delta\beta, y\Delta) \circ GIG(-\frac{1}{2}, \delta, \sqrt{\alpha^2 - ||\beta||^2_\Delta}).$$

2. If $\lambda = \frac{d+1}{2}$, then $GH_d(\frac{d+1}{2}, \alpha, \beta, \delta, \mu, \Delta)$ corresponds to the hyperbolic distribution $HYP_d(\alpha, \beta, \delta, \mu, \Delta)$.

3. If $\lambda > 0$ and $\delta \to 0$, then $GH_d(\lambda > 0, \alpha, \beta, 0, \mu, \Delta)$ corresponds to the VG distribution

$$VG_d(\lambda, \alpha, \beta, \mu, \Delta) := N_d(\mu + y\Delta\beta, y\Delta) \circ Gamma(\lambda, \frac{\alpha^2 - ||\beta||^2_\Delta}{2}),$$

where $Gamma(\lambda, \gamma)$ is the Gamma distribution with the pdf $f(x) = \frac{\gamma^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\gamma x}$. Furthermore, if $\lambda = \frac{d+1}{2}$ (i.e., the above hyperbolic case), then $VG_d(\frac{d+1}{2}, \alpha, \beta, \mu, \Delta)$ corresponds to the skewed Laplace distribution

$$LAP_d(\alpha, \beta, \mu, \Delta) := N_d(\mu + y\Delta\beta, y\Delta) \circ Gamma(\frac{d+1}{2}, \frac{\alpha^2 - ||\beta||^2_\Delta}{2}),$$

with the pdf $f(x) \propto e^{-\alpha||x-\mu||_{\Delta^{-1}} + \langle \beta, x-\mu \rangle}$. We have seen the case of $d = 1$ in Example 3.8.

4. If $\lambda < 0$, $\alpha \to 0$, and $\beta \to \mathbf{0}$, then $GH_d(\lambda < 0, 0, \mathbf{0}, \delta, \mu, \Delta)$ corresponds to the scaled and shifted t distribution with $f = -2\lambda$ degrees of freedom:

$$t_d(\lambda, \delta, \mu, \Delta) := N_d(\mu, y\Delta) \circ iGamma(\lambda, \frac{\delta^2}{2}),$$

where $iGamma(\lambda, \delta)$ is the inverse Gamma distribution with the pdf $f(x) = \frac{x^{\lambda-1}}{\delta^\lambda \Gamma(-\lambda)} e^{-\frac{\delta}{x}}$. Furthermore, if $\lambda = -\frac{1}{2}$ (i.e., the above NIG case), then $t_d(-\frac{1}{2}, \delta, \mu, \Delta)$ corresponds to the multivariate Cauchy distribution

$$CAU(\delta, \mu, \Delta) := N_d(\mu, y\Delta) \circ iGamma(-\frac{1}{2}, \frac{\delta^2}{2}),$$

with the pdf $f(x) \propto (1 + \frac{||x-\mu||^2_{\Delta^{-1}}}{\delta^2})^{-\frac{d+1}{2}}$, which is also shown in Example 4.3.

These classes have the following convolution properties, by using Lemma 4.4 and Proposition B.1, which are the multivariate extensions of the univariate case (v. Hammerstein, 2010, eq. (1.9), p. 14).

Proposition 4.5 *For each $d \geq 1$, there are the following convolution properties in the d-dimensional GH distributions:*

1. $NIG_d(\alpha, \beta, \delta_1, \mu_1, \Delta) * NIG_d(\alpha, \beta, \delta_2, \mu_2, \Delta) = NIG_d(\alpha, \beta, \delta_1 + \delta_2, \mu_1 + \mu_2, \Delta)$,
2. $VG_d(\lambda_1, \alpha, \beta, \mu_1, \Delta) * VG_d(\lambda_2, \alpha, \beta, \mu_2, \Delta) = VG_d(\lambda_1 + \lambda_2, \alpha, \beta, \mu_1 + \mu_2, \Delta)$,
3. $NIG_d(\alpha, \beta, \delta_1, \mu_1, \Delta) * GH_d(1/2, \alpha, \beta, \delta_2, \mu_2, \Delta) = GH_d(1/2, \alpha, \beta, \delta_1 + \delta_2, \mu_1 + \mu_2, \Delta)$,
4. $GH_d(-\lambda, \alpha, \beta, \delta, \mu_1, \Delta) * GH_d(\lambda, \alpha, \beta, 0, \mu_2, \Delta) = GH_d(\lambda, \alpha, \beta, \delta, \mu_1 + \mu_2, \Delta)$,

where $\lambda, \lambda_1, \lambda_2 > 0$.

These convolution properties can also be obtained by looking up their characteristic functions and Lévy measures in v. Hammerstein (2010, Section 1.6.4, p. 46, Section 2.3, p. 79). Properties 1 and 2 imply a convolution semigroup. Property 3 implies an absorbing property. Property 4 implies another convolution property. By observing proposition 4.5, we obtain the following conjugate, absorbing, and related pairs in GH kernels and GH distributions. The parametric models in Proposition 4.5 contain p.d. kernels κ if and only if $\beta = \mathbf{0}$. Each example $(1-4)$ in the following corresponds to each property $(1-4)$ in Proposition 4.5.

Example 4.6 *Conjugate, absorbing, and related pairs in the GH class.*

1. *Let $k_{\alpha,\delta,\Delta}(x, y)$ be a shift invariant NIG p.d. kernel and $\mathcal{H}_{\alpha,\delta,\Delta}$ be the RKHS. Let P, Q be two NIG distributions $NIG(\alpha, \mathbf{0}, \delta_P, \mu_P, \Delta)$ and $NIG(\alpha, \mathbf{0}, \delta_Q, \mu_Q, \Delta)$, respectively. Then, the kernel mean is the NIG pdf $m_P = f(\cdot|\alpha, \mathbf{0}, \delta_P + \delta, \mu_P, \Delta)$ and the RKHS inner product is the NIG pdf $\langle m_P, m_Q \rangle_{\mathcal{H}_{\alpha,\delta,\Delta}} = f(\mu_P|\alpha, \mathbf{0}, \delta_P + \delta_Q + \delta, \mu_Q, \Delta)$. If $\alpha \to 0$, then these correspond to the Cauchy case.*

2. *Let $k_{\lambda,\alpha,\Delta}(x, y)$ be a shift invariant VG p.d. kernel*[9] *and $\mathcal{H}_{\lambda,\alpha,\Delta}$ be the RKHS. Let P, Q be two VG distributions $VG(\lambda_P, \alpha, \mathbf{0}, \mu_P, \Delta)$ and $VG(\lambda_Q, \alpha, \mathbf{0}, \mu_Q, \Delta)$, respectively. Then, the kernel mean is the VG pdf $m_P = f(\cdot|\lambda_P + \lambda, \alpha, \mathbf{0}, \mu_P, \Delta)$ and the RKHS inner product is the VG pdf $\langle m_P, m_Q \rangle_{\mathcal{H}_{\lambda,\alpha,\Delta}} = f(\mu_P|\lambda_P + \lambda_Q + \lambda, \alpha, \mathbf{0}, \mu_Q, \Delta)$. If $\lambda = \frac{d+1}{2}$, $\lambda_P = \frac{d+1}{2}$, or $\lambda_Q = \frac{d+1}{2}$, then these correspond to the Laplace case.*

3. *Let $k_{\alpha,\delta,\Delta}(x, y)$ be a NIG kernel and $\mathcal{H}_{\alpha,\delta,\Delta}$ be the RKHS. Let P be a GH distribution $GH(1/2, \alpha, \mathbf{0}, \delta_P, \mu_P, \Delta)$. Then, the kernel mean is the GH pdf $m_P = f(\cdot|1/2, \alpha, \mathbf{0}, \delta_P + \delta, \mu_P, \Delta)$. If $\alpha \to 0$, then the NIG kernel $k_{0,\delta,\Delta}(x, y)$ corresponds to the Cauchy kernel.*

   Let $k_{1/2,\alpha,\delta,\Delta}(x, y)$ be a GH kernel and $\mathcal{H}_{1/2,\alpha,\delta,\Delta}$ be the RKHS. Let P, Q be two NIG distributions $NIG(\alpha, 0, \delta_P, \mu_P, \Delta)$ and $NIG(\alpha, \mathbf{0}, \delta_Q, \mu_Q, \Delta)$, respectively. Then, the

9. The Matérn kernel corresponds to $\Delta = I$, and $\alpha = \frac{\sqrt{2\nu}}{\sigma}$ (Rasmussen and Williams, 2006, Section 4.2.1) (Sriperumbudur et al., 2010, p. 1533)

kernel mean is the GH pdf $m_P = f(\cdot|1/2, \alpha, \mathbf{0}, \delta_P + \delta, \mu_P, \Delta)$ *and the RKHS inner product is the GH pdf* $\langle m_P, m_Q \rangle_{\mathcal{H}_{1/2,\alpha,\delta,\Delta}} = f(\mu_P|1/2, \alpha, \mathbf{0}, \delta_P + \delta_Q + \delta, \mu_Q, \Delta)$. *If* $\alpha \to 0$*, then the NIG distributions,* P *and* Q*, correspond to the Cauchy distributions.*

4. *For* $\lambda > 0$*, let* $k_{-\lambda,\alpha,\delta,\Delta}(x, y)$ *be a GH kernel and* $\mathcal{H}_{-\lambda,\alpha,\delta,\Delta}$ *be the RKHS. Let* P *be a GH distribution* $GH(\lambda, \alpha, \mathbf{0}, 0, \mu_P, \Delta)$. *Then, the kernel mean is the GH pdf* $m_P = f(\cdot|\lambda, \alpha, \mathbf{0}, \delta, \mu_P, \Delta)$. *If* $\alpha \to 0$*, then* $k_{-\lambda,0,\delta,\Delta}(x, y)$ *corresponds to the Student's t kernel. Furthermore, if* $\lambda = \frac{1}{2}$*, then* $k_{-\frac{1}{2},0,\delta,\Delta}(x, y)$ *corresponds to the Cauchy kernel.*

   For $\lambda > 0$*, let* $k_{\lambda,\alpha,\Delta}(x, y)$ *be a GH kernel and* $\mathcal{H}_{\lambda,\alpha,\Delta}$ *be the RKHS. Let* P *be a GH distribution* $GH(-\lambda, \alpha, \mathbf{0}, \delta_P, \mu_P, \Delta)$. *Then, the kernel mean is the GH pdf* $m_P = f(\cdot|\lambda, \alpha, \mathbf{0}, \delta_P, \mu_P, \Delta)$. *If* $\alpha \to 0$*, then* P *is the Student's t distribution. Furthermore, if* $\lambda = -\frac{1}{2}$*, then* P *is the Cauchy distribution.*

5. Connection to Machine Learning

As mentioned in the Introduction, absorbing and conjugate kernels (Examples 4.3 and 4.6) provide a way to compute the RKHS values (i) $f(x)$, $x \in \mathbb{R}^d$, and the RKHS inner products (ii) $\langle f, g \rangle_{\mathcal{H}}$ when $f, g \in \mathcal{H}$ are expressed by the weighted sums of parametric kernel means, $f = \sum_{i=1}^n w_i m_{P_i}$ and $g = \sum_{j=1}^l \tilde{w}_j m_{Q_j}$ for $\{P_i\}, \{Q_j\} \subset \mathcal{P}_\Theta$. Many algorithms aim to use the convolution trick. Examples include as follows:

- The difference between a probability measure $P \in \mathcal{M}_1(\mathbb{R}^d)$ and a model $P_\theta \in \mathcal{P}_\Theta$ in the RKHS norm $||m_P - m_{P_\theta}||_{\mathcal{H}}$ needs to be computed, e.g., for the purpose of a goodness-of-fit test and model criticism (Lloyd and Ghahramani, 2015), based on the maximum mean discrepancy (MMD) (Gretton et al., 2012).

- Various kernels $k(P, P_\theta)$ between a probabilistic measure P and a model P_θ, e.g., $k(P, P_\theta) = \exp(-\frac{||m_P - m_{P_\theta}||^2_{\mathcal{H}}}{2\sigma^2})$ need to be computed, as in the support measure machine (Muandet et al., 2012).

- Song et al. (2008) and McCalman et al. (2013) studied an approximation of a target probability measure $P \in \mathcal{M}_1(\mathbb{R}^d)$ with a Gaussian mixture model $P_\theta = \sum_{i=1}^n \theta_i P_i$ via solving the following optimization problem:

$$\hat{\theta} = \mathrm{argmin}_\theta ||m_P - m_{P_\theta}||^2_{\mathcal{H}} + \Omega(\theta) = \mathrm{argmin}_\theta ||m_P - \sum_{i=1}^n \theta_i m_{P_i}||^2_{\mathcal{H}} + \Omega(\theta),$$

  where $\Omega(\theta)$ is a regularization term, $\frac{\lambda}{2}||\theta||^2$ $(\lambda > 0)$. This optimization is solved by a constrained quadratic program: $\min_\theta \frac{1}{2}\theta^\top(A + \lambda I_n)\theta - b^\top\theta$ subject to $\sum_{i=1}^n \theta_i = 1$ and $\theta \geq 0$, where we then need the computation of matrix $A \in \mathbb{R}^{n\times n}$ and vector $b \in \mathbb{R}^n$:

$$A_{ij} = \langle m_{P_i}, m_{P_j} \rangle_{\mathcal{H}}, \quad b_j = \langle \hat{m}_P, m_{P_j} \rangle_{\mathcal{H}}, \quad 1 \leq i, j \leq n,$$

  for parametric kernel means $\{m_{P_i}\}$.

- As mentioned in the Introduction, the kernel Bayesian inference (KBI), which employs Bayesian inference in kernel mean form, has been proposed (Fukumizu et al. 2013, Song et al. 2013). KBI is applied to, e.g., filtering and smoothing algorithms on state space models (Fukumizu et al. 2013 Kanagawa et al. 2016, Nishiyama et al. 2016) and policy learning in reinforcement learning (Grünewälder et al. 2012, Nishiyama et al. 2012, Rawlik et al. 2013, Boots et al. 2013). When we extend it to *semiparametric* KBI, which combines nonparametric inference and parametric inference, we may want to use the RKHS functions $f = \sum_{i=1}^{n} w_i m_{P_{\theta_i}} \in \mathcal{H}$ expressed by parametric kernel means $\{P_{\theta_i}\} \in \mathcal{P}_\Theta$, as is used in the model-based kernel sum rule (Mb-KSR) (Nishiyama et al., 2014).

- Preimage algorithms (Mika et al., 1999; Fukumizu et al., 2013) and kernel herding algorithms (Chen et al., 2010) can also be extended to estimators $f = \sum_{i=1}^{n} w_i m_{P_{\theta_i}}$ with parametric kernel means $\{P_{\theta_i}\}$.

6. Computation of Conjugate Kernels (Convolution Trick)

In Section 4, we mathematically investigated that several convolution tricks hold within a general convolution trick (Proposition 4.2): the computation of kernel mean values and RKHS inner products is the same as the computation of p.d. kernels having different parameters, if conjugate kernels are used. However, conjugate kernels do not provide a tractable computation in general. We then discuss the computation of the conjugate kernels: α-stable kernels and GH kernels.

- It is known that α-stable pdfs do not generally have a closed-form expression except for some special cases, Gaussians ($\alpha = 2$) and Cauchy ($\alpha = 1$), as given in Appendix A.3. Gaussian and Cauchy kernels may be used as tractable conjugate kernels. For other α-stable kernels ($\alpha \neq 2$ and $\alpha \neq 1$), some numerical elaborations or approximations may be needed for the computation of the pdfs. The STABLE 5.1[10] software allows the computation of α-stable pdfs when they are independent, isotropic, elliptical, or have discrete spectral measures Γ_d under some settings. More information can be found in the STABLE 5.1 software manual. For elliptically contoured α-stable sub-Gaussian kernels on any dimension $\mathbb{R}^d$, the computation of pdfs is sufficient only to compute a one-dimensional amplitude function $\tilde{\kappa}(r)$ in equation (2), which can be computed by, e.g., a one-dimensional numerical integration. The STABLE 5.1 software supports the computation of sub-Gaussian pdfs in dimension $d < 100$.

- GH kernels and their subclasses are also elliptical pdfs, and the computation of the kernels is sufficient only to compute a one-dimensional amplitude function $\tilde{\kappa}(r)$. VG kernels or Matérn kernels, which are a generalization of Laplace kernels, are used for covariance kernels in Gaussian processes. GH and NIG kernels are variants of Matérn kernels, all of which are expressed by the Bessel function of the third kind. For example, there is an R package software called 'ghyp' on the GH distributions (Breymann and Lüthi, 2013).

10. John Nolan's Page. `http://academic2.american.edu/~jpnolan/stable/stable.html`

In addition, random Fourier features (Rahimi and Recht, 2007) may be an approach to approximately compute conjugate kernels. From Proposition 4.2, we have an equality

$$\langle m_P, m_Q \rangle_{\mathcal{H}_{A,\nu_s}} = k_{A+A_P+A_Q,\nu_s+\nu_P+\tilde{\nu}_Q}(\gamma_P, \gamma_Q) = \mathbb{E}_\omega[\zeta_\omega(\gamma_P)\zeta_\omega(\gamma_Q)^*].$$

An RKHS inner product (l.h.s.) may be computed by approximating the expectation of $\zeta_\omega(\gamma_P)\zeta_\omega(\gamma_Q)^*$ (r.h.s.) with sampling ω from the characteristic function having the generating triplet $(A + A_P + A_Q, \nu_s + \nu_P + \tilde{\nu}_Q)$.

7. Conclusion

In this paper, we introduced a class of CID kernels that constitutes a large subclass in the set of shift-invariant characteristic kernels on $\mathbb{R}^d$, where CID kernels are closed under convolution but not closed under addition and pointwise product. We introduced absorbing, conjugate kernels, and convolution trick with respect to parametric models, where the basic computation of kernel mean values and RKHS inner products results in the computation of the same p.d. kernels with different parameters, which is an extension of kernel trick. Although the convolution trick may offer a mathematical view, the computation of conjugate kernels is not tractable in general. We then restrict convolution trick only to tractable cases or approximately compute intractable conjugate kernels. Future works include investigating the effectiveness of convolution trick in practice and developing approximation algorithms to efficiently compute intractable conjugate kernels.

Acknowledgments

We thank anonymous reviewers and the action editor for helpful comments. Y.N. thanks Prof. Tatsuhiko Saigo and Prof. Takaaki Shimura for a helpful discussion on infinitely divisible distributions. This work was supported in part by JSPS KAKENHI (grant nos. 26870821 and 22300098), the MEXT Grant-in-Aid for Scientific Research on Innovative Areas (no. 25120012), and by the Program to Disseminate Tenure Tracking System, MEXT, Japan.

Appendix A. α-Stable Distributions

We briefly review the α-stable distributions on $\mathbb{R}^d$.

A.1 α-Stable Distributions on $\mathbb{R}^d$

The α-stable distribution on $\mathbb{R}^d$ has the following characteristic function:

Theorem A.1 *(Samorodnitsky and Taqqu, 1994, Theorem 2.3.1, p. 65) Let $\alpha \in (0, 2)$. Then, $X = (X_1, \ldots, X_d)$ is an α-stable random vector in $\mathbb{R}^d$ if and only if there exists a finite measure Γ on the unit sphere $S_{d-1} = \{s \in \mathbb{R}^d : ||s|| = 1\}$ and a vector $\mu^0 \in \mathbb{R}^d$ such that*

$$\hat{P}(\theta) = \begin{cases} \exp\left(-\int_{S_{d-1}} |\theta^\top s|^\alpha \left(1 - i\mathrm{sgn}(\theta^\top s) \tan \frac{\pi\alpha}{2}\right) \Gamma(ds) + i\theta^\top \mu^0\right), & (\alpha \neq 1). \\ \exp\left(-\int_{S_{d-1}} |\theta^\top s|^\alpha \left(1 + i\frac{2}{\pi}\mathrm{sgn}(\theta^\top s) \ln |\theta^\top s|\right) \Gamma(ds) + i\theta^\top \mu^0\right), & (\alpha = 1). \end{cases}$$

The pair (Γ, μ^0) *is unique.*

The measure Γ is called the *spectral measure*. See Samorodnitsky and Taqqu (1994, Section 2.3) for some examples of spectral measures. The radial sub-Gaussian distribution has a uniform spectral measure. An α-stable random vector $X = (X_1, \ldots, X_d)$ has independent components if and only if its spectral measure Γ is discrete and concentrated on the intersection of the axes with the sphere S_{d-1}. It is known that any nondegenerate stable distribution on $\mathbb{R}^d$ has the C^∞ pdf (Sato, 1999, Example 28.2, p. 190). An α-stable distribution on $\mathbb{R}^d$ is symmetric if and only if $\mu^0 = 0$ and Γ is a symmetric measure on S_{d-1} (i.e., it satisfies $\Gamma(A) = \Gamma(-A)$ for any $A \in \mathcal{B}(S_{d-1})$) (Samorodnitsky and Taqqu, 1994, p.73).

For each $\alpha \in (0, 2)$, α-stable distributions on $\mathbb{R}^d$ have the generating triplet $(0, \nu, \gamma)$ with

$$\nu(B) = \int_{S_{d-1}} \Gamma(ds) \int_0^\infty 1_B(rs) \frac{dr}{r^{1+\alpha}}, \quad B \in \mathcal{B}(\mathbb{R}^d), \tag{5}$$

where Γ is the spectral measure on S_{d-1} (Sato, 1999, Theorem 14.3, p. 77). The sum of Lévy measures $\nu_1 + \nu_2$ implies the sum of spectral measures $\Gamma_1 + \Gamma_2$.

A.2 α-Stable Distributions on $\mathbb{R}$

As a special case, an α-stable distribution on $\mathbb{R}$ has the following characteristic function:

Theorem A.2 *(Samorodnitsky and Taqqu, 1994, Definition 1.1.6, p. 5) A random variable* X *is* α*-stable* ($\alpha \in (0, 2]$) *in* $\mathbb{R}$ *if and only if the parameters satisfy the conditions* $\sigma \geq 0$, $\beta \in [-1, 1]$, *and* $\mu \in \mathbb{R}$ *such that its characteristic function has the form*

$$\hat{P}(\theta) = \begin{cases} \exp\left(-\sigma^\alpha |\theta|^\alpha (1 - i\beta(\mathrm{sgn}\theta) \tan \frac{\pi\alpha}{2}) + i\mu\theta\right) & (\alpha \neq 1), \\ \exp\left(-\sigma|\theta|(1 + i\beta\frac{2}{\pi}(\mathrm{sgn}\theta) \ln|\theta|) + i\mu\theta\right) & (\alpha = 1), \end{cases}$$

where $\mathrm{sgn}\theta$ *is a sign function*

$$\mathrm{sgn}\theta = \begin{cases} 1 & \theta > 0, \\ 0 & \theta = 0, \\ -1 & \theta < 0. \end{cases}$$

When $\alpha \in (0, 2)$*, the parameters* σ*,* β*, and* μ *are unique. When* $\alpha = 2$*,* β *is irrelevant, and* σ *and* μ *are unique.*

An α-stable distribution on $\mathbb{R}$ is specified by the parameters (σ, β, μ), where σ is a scale parameter, β is a skewness parameter, and μ is a location parameter. $\sigma = 0$ implies a delta measure. For $\alpha \in (0, 2)$, an α-stable distribution is symmetric if and only if $\beta = \mu = 0$ (Samorodnitsky and Taqqu, 1994, Property 1.2.5, p. 11). A 2-stable distribution is symmetric if and only if $\mu = 0$. An α-stable density does not generally have a closed-form expression, except for some special cases. However, it is known that every nondegenerate stable distribution has the C^∞ pdf (Sato, 1999, Example 28.2, p. 190). Some known univariate α-stable pdfs, expressed by elementary functions and special functions, are given in Appendix A.3.

The Lévy measure ν of a univariate stable distribution is obtained by letting $d = 1$ in the Lévy measure (5). If $d = 1$, then $S_0 = \{-1, 1\}$ and $\Gamma = \Gamma(\{-1\})\delta_{-1} + \Gamma(\{1\})\delta_1$, where

$\Gamma(\{-1\}), \Gamma(\{1\}) \geq 0$ and $\Gamma(\{-1\}) + \Gamma(\{1\}) > 0$ (Samorodnitsky and Taqqu, 1994, Example 2.3.3, p. 67). By substituting this into equation (5), we can obtain the Lévy measure ν of a univariate stable distribution as

$$\nu(dx) = \Gamma(\{1\}) \frac{1}{x^{1+\alpha}} 1_{(0,\infty)}(x)dx + \Gamma(\{-1\}) \frac{1}{|x|^{1+\alpha}} 1_{(-\infty,0)}(x)dx.$$

A stable distribution $S_\alpha(\sigma, \beta, \mu)$ is given with the spectral measure as

$$\sigma = (\Gamma(\{1\}) + \Gamma(\{-1\}))^{\frac{1}{\alpha}} > 0, \;\; \beta = \frac{(\Gamma(\{1\}) - \Gamma(\{-1\}))}{\Gamma(\{1\}) + \Gamma(\{-1\})} \in [-1, 1].$$

The sum of Lévy measures $\nu_1 + \nu_2$ implies the sum of mass functions $\Gamma_1(\{-1\}) + \Gamma_2(\{-1\})$ and $\Gamma_1(\{1\}) + \Gamma_2(\{1\})$. We can see the convolution property $S_\alpha(\sigma_1, \beta_1, \mu_1) * S_\alpha(\sigma_2, \beta_2, \mu_2) = S_\alpha((\sigma_1^\alpha + \sigma_2^\alpha)^{\frac{1}{\alpha}}, \frac{\sigma_1^\alpha \beta_1 + \sigma_2^\alpha \beta_2}{\sigma_1^\alpha + \sigma_2^\alpha}, \mu_1 + \mu_2)$ of the univariate stable distribution from the viewpoint of the spectral measure.

A.3 Closed-Form and Special Function Form of α-Stable PDFs on $\mathbb{R}$

There are three cases where the α-stable pdf on $\mathbb{R}$ is expressed by elementary functions:

1. The 2-stable distribution $S_2(\sigma, \beta, \mu)$ is the Gaussian $N(\mu, 2\sigma^2)$, where β has no effect, with the pdf

$$f_{Gauss}(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{4\sigma^2}}, x \in \mathbb{R}.$$

2. The 1-stable distribution $S_1(\sigma, \beta = 0, \mu)$ is the Cauchy distribution with the pdf

$$f_{Cauchy}(x) = \frac{\sigma}{\pi((x-\mu)^2 + \sigma^2)}, x \in \mathbb{R}.$$

3. The 1/2-stable distribution $S_{1/2}(\sigma, \beta = \pm 1, \mu)$ is the Lévy distribution with the pdf

$$f_{Levy}(x) = \frac{\sqrt{\sigma}}{\sqrt{2\pi}(x-\mu)^{3/2}} e^{-\frac{\sigma}{2(x-\mu)}}, \mu < x < \infty.$$

There are some cases where the α-stable pdf is expressed by special functions. The following expression is found in Lee (2010). Note that kernel means m_P and RKHS inner products also take these expressions. For simplicity, we only show standardized stable pdfs $d_{stable}(x; \alpha, \sigma = 1, \beta, \mu = 0)$.

Fresnel integrals:

If $(\alpha, \sigma, \beta, \mu) = (1/2, 1, 0, 0)$,

$$d_{stable}(x; 1/2, 1, 0, 0) = \frac{|x|^{-\frac{3}{2}}}{\sqrt{2\pi}} \left(\sin\left(\frac{1}{4|x|}\right) \left(\frac{1}{2} - S\left(\sqrt{\frac{1}{2\pi|x|}}\right)\right) + \cos\left(\frac{1}{4|x|}\right) \left(\frac{1}{2} - C\left(\sqrt{\frac{1}{2\pi|x|}}\right)\right) \right),$$

where $C(z)$ and $S(z)$ are the Fresnel integrals

$$C(z) = \int_0^z \cos\left(\frac{\pi t^2}{2}\right) dt, \qquad S(z) = \int_0^z \sin\left(\frac{\pi t^2}{2}\right) dt.$$

This is a symmetric stable pdf. $k(x,y) = d_{stable}(x-y; 1/2, 1, 0, 0)$, $x, y \in \mathbb{R}$, gives a characteristic p.d. kernel.

Modified Bessel function:

If $(\alpha, \sigma, \beta, \mu) = (1/3, 1, 1, 0)$, the one-sided continuous density is

$$d_{stable}(x; 1/3, 1, 1, 0) = \frac{1}{\pi} \frac{2^{3/2}}{3^{7/4}} x^{-3/2} K_{1/3}\left(\frac{2^{5/2}}{3^{9/4}} x^{-1/2}\right), x \geq 0,$$

where $K_\nu(x)$ is a modified Bessel function of the third kind.

Hypergeometric function:

If $(\alpha, \sigma, \beta, \mu) = (4/3, 1, 0, 0)$,

$$\begin{aligned} d_{stable}(x; \frac{4}{3}, 1, 0, 0) \quad = \quad & \frac{3^{5/4}\Gamma(7/12)\Gamma(11/12)}{2^{5/2}\sqrt{\pi}\Gamma(6/12)\Gamma(8/12)} {}_2F_2\left(\frac{7}{12}, \frac{11}{12}; \frac{6}{12}, \frac{8}{12}; \frac{3^3 x^4}{2^8}\right) \\ & - \frac{3^{11/4}|x|^3\Gamma(13/12)\Gamma(17/12)}{2^{13/2}\sqrt{\pi}\Gamma(18/12)\Gamma(15/12)} {}_2F_2\left(\frac{13}{12}, \frac{17}{12}; \frac{18}{12}, \frac{15}{12}; \frac{3^3 x^4}{2^8}\right), x \in \mathbb{R}, \end{aligned}$$

where ${}_pF_q$ is the (generalized) hypergeometric function

$${}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; z) = \sum_{n=0}^{\infty} \frac{(a_1)_n \cdots (a_p)_n}{(b_1)_n \cdots (b_q)_n} \frac{z^n}{n!}$$

with the Pochhammer symbol $(a)_0 = 1$, $(a)_n = a(a+1)\ldots(a+n-1)$ for $n \in \mathbb{N}^+$. This is a symmetric stable pdf. $k(x,y) = d_{stable}(x-y; \frac{4}{3}, 1, 0, 0)$, $x, y \in \mathbb{R}$, gives a characteristic p.d. kernel.

If $(\alpha, \sigma, \beta, \mu) = (3/2, 1, 0, 0)$ (the Holtsmark distribution),

$$\begin{aligned} d_{stable}(x; \frac{3}{2}, 1, 0, 0) \quad = \quad & \frac{1}{\pi}\Gamma(5/3) {}_2F_3\left(\frac{5}{12}, \frac{11}{12}; \frac{1}{3}, \frac{1}{2}, \frac{5}{6}; -\frac{2^2 x^6}{3^6}\right) \\ & - \frac{x^2}{3\pi} {}_3F_4\left(\frac{3}{4}, 1, \frac{5}{4}; \frac{2}{3}, \frac{5}{6}, \frac{7}{6}, \frac{4}{3}; -\frac{2^2 x^6}{3^6}\right) \\ & + \frac{7x^4}{3^4\pi}\Gamma(4/3) {}_2F_3\left(\frac{13}{12}, \frac{19}{12}; \frac{7}{6}, \frac{3}{2}, \frac{5}{3}; -\frac{2^2 x^6}{3^6}\right), x \in \mathbb{R}. \end{aligned}$$

This is a symmetric stable pdf. The Holtsmark kernel $k(x,y) = d_{stable}(x-y; 3/2, 1, 0, 0)$, $x, y \in \mathbb{R}$, gives a characteristic p.d. kernel.

Whittaker function:

If $(\alpha, \sigma, \beta, \mu) = (2/3, 1, 0, 0)$,

$$d_{stable}(x; 2/3, 1, 0, 0) = \frac{1}{2\sqrt{3\pi}\,|x|} \exp\left(\frac{2}{27x^2}\right) W_{-1/2, 1/6}\left(\frac{4}{27x^2}\right), x \in \mathbb{R},$$

where $W_{\lambda,\mu}(z)$ is the Whittaker function defined as

$$\begin{aligned} W_{\lambda,\mu}(z) &= \frac{z^{\lambda}e^{-z/2}}{\Gamma(\mu-\lambda+1/2)}\int_0^{\infty}\mathrm{e}^{-t}t^{\mu-\lambda-1/2}\left(1+\frac{t}{z}\right)^{\mu-\lambda-1/2}dt, \\ &\quad \mathrm{Re}(\mu-\lambda)>-\frac{1}{2}, |\arg(z)|<\pi. \end{aligned}$$

This is a symmetric stable pdf. $k(x,y) = d_{stable}(x-y;2/3,1,0,0)$, $x,y\in\mathbb{R}$, gives a characteristic p.d. kernel.

If $(\alpha,\sigma,\beta,\mu) = (2/3,1,1,0)$, the one-sided density is

$$d_{stable}(x;2/3,1,1,0) = \sqrt{\frac{3}{\pi}}\frac{1}{|x|}\exp\left(-\frac{16}{27x^2}\right)W_{1/2,1/6}\left(\frac{32}{27x^2}\right), x\geq 0.$$

If $(\alpha,\sigma,\beta,\mu) = (3/2,1,1,0)$, the α-stable density is

$$d_{stable}(x;2/3,1,1,0) = \begin{cases} \sqrt{\frac{3}{\pi}}\frac{1}{|x|}\exp\left(\frac{x^3}{27}\right)W_{1/2,1/6}\left(-\frac{2}{27}x^3\right), & x<0 \\ \frac{1}{2\sqrt{3\pi}|x|}\exp\left(\frac{x^3}{27}\right)W_{-1/2,1/6}\left(\frac{2}{27}x^3\right), & x>0 \end{cases}$$

Lommel function:

If $(\alpha,\sigma,\beta,\mu) = (1/3,1,0,0)$,

$$d_{stable}(x;1/3,1,0,0) = \mathrm{Re}\left(\frac{2\exp(-i\pi/4)}{3\sqrt{3}\pi|x|^{3/2}}S_{0,1/3}\left(\frac{2\exp(i\pi/4)}{3\sqrt{3}|x|^{1/2}}\right)\right).$$

Here, the Lommel functions $s_{\mu,v}(z)$ and $S_{\mu,v}(z)$ are defined by

$$\begin{aligned} s_{\mu,v}(z) &= \frac{\pi}{2}\left(Y_{\nu}(z)\int_0^z z^{\mu}J_v(z)dz - J_{\nu}(z)\int_0^z z^{\mu}Y_v(z)dz\right), \\ S_{\mu,v}(z) &= s_{\mu,v}(z) - \frac{2^{\mu-1}\Gamma\left((1+\mu+\nu)/2\right)}{\pi\Gamma\left((\nu-\mu)/2\right)}\left(J_{\nu}(z)-\cos\left(\frac{\mu-\nu}{2}\pi\right)Y_v(z)\right), \end{aligned}$$

where $J_{\nu}(z)$ and $Y_{\nu}(z)$ are Bessel functions of the first and second kind, respectively. This is a symmetric stable pdf. $k(x,y) = d_{stable}(x-y;1/3,1,0,0)$, $x,y\in\mathbb{R}$, gives a characteristic p.d. kernel.

Landau distribution:

If $(\alpha,\sigma,\beta,\mu) = (1,1,1,0)$ (the Landau distribution),

$$d_{stable}(x;1,1,1,0) = \frac{1}{\pi}\int_0^{\infty}e^{-t\log t - xt}\sin(\pi t)dt.$$

A.4 Sub-Gaussian (Elliptically Contoured) α-Stable Distributions on $\mathbb{R}^d$

The sub-Gaussian α-stable distribution has the following characteristic function:

Proposition A.3 *(Samorodnitsky and Taqqu, 1994, Proposition 2.5.2, p. 78) Let $\alpha \in (0,2)$. The sub-Gaussian α-stable random vector X in $\mathbb{R}^d$ has the characteristic function*

$$E\exp\left[i\sum_{k=1}^{d}\theta_k X_k\right] = \exp\left(-\left|\frac{1}{2}\sum_{ij=1}^{d}\theta_i\theta_j R_{ij}\right|^{\frac{\alpha}{2}} + i(\theta,\mu^0)\right),$$

where R is a p.d. matrix and $\mu^0 \in \mathbb{R}^d$ is a shift vector.

$\alpha = 2$ and $\alpha = 1$ imply the multivariate Gaussian and Cauchy distribution, respectively.

For $\alpha \in (0,2)$, the radial sub-Gaussian $\mathbb{SG}_\alpha(\mathbb{R}^d)[I]$ (with identity matrix $R = I$) has the uniform spectral measure $\Gamma(B) = c|B|$, $\forall B \in \mathcal{B}(S_{d-1})$ in the Lévy measure (5) (Samorodnitsky and Taqqu, 1994, Proposition 2.5.5, p. 79). Sub-Gaussian $\mathbb{SG}_\alpha(\mathbb{R}^d)[R]$ with a p.d. matrix R is the elliptical version of the radial sub-Gaussians. Its spectral measure is given in Samorodnitsky and Taqqu (1994, Proposition 2.5.8, p. 82).

Appendix B. GH Classes on $\mathbb{R}^d$

A GH distribution on $\mathbb{R}^d$ is given by the normal mean-variance mixture with the GIG mixing distribution. See, e.g., v. Hammerstein (2010) for more information. We here reproduce some of them.

B.1 GIG Distributions on $\mathbb{R}^+$

A generalized inverse Gaussian (GIG) distribution $GIG(\lambda,\delta,\gamma)$ on $\mathbb{R}^+$ is given by the following pdf:

$$d_{GIG(\lambda,\delta,\gamma)}(x) = \left(\frac{\gamma}{\delta}\right)^{\lambda}\frac{1}{2K_\lambda(\delta\gamma)}x^{\lambda-1}\exp\left(-\frac{1}{2}\left(\frac{\delta^2}{x}+\gamma^2 x\right)\right)1_{(0,\infty)}(x),$$

where $K_\lambda(x)$ is the modified Bessel function of the third kind with index λ. The parameters (λ,δ,γ) take the following values:

$$\begin{cases} \delta \geq 0, \gamma > 0, & \text{if} \quad \lambda > 0, \\ \delta > 0, \gamma > 0, & \text{if} \quad \lambda = 0, \\ \delta > 0, \gamma \geq 0, & \text{if} \quad \lambda < 0, \end{cases}$$

where $\delta = 0$ and $\gamma = 0$ correspond to limiting cases,[11] which are the Gamma distribution and the inverse Gamma distribution, respectively. The GIG distributions have the following convolution properties:

Proposition B.1 *(v. Hammerstein, 2010, Proposition 1.11, p. 11) Within the class of GIG distributions, the following convolution properties hold:*

a) $GIG(-\frac{1}{2},\delta_1,\gamma) * GIG(-\frac{1}{2},\delta_2,\gamma) = GIG(-\frac{1}{2},\delta_1+\delta_2,\gamma)$,

b) $GIG(-\frac{1}{2},\delta_1,\gamma) * GIG(\frac{1}{2},\delta_2,\gamma) = GIG(\frac{1}{2},\delta_1+\delta_2,\gamma)$,

c) $GIG(-\lambda,\delta,\gamma) * GIG(\lambda,0,\gamma) = GIG(\lambda,\delta,\gamma), \qquad \lambda > 0$,

d) $GIG(\lambda_1,0,\gamma) * GIG(\lambda_2,0,\gamma) = GIG(\lambda_1+\lambda_2,0,\gamma), \qquad \lambda_1,\lambda_2 > 0$.

11. If $\lambda \neq 0$, then $K_\lambda(x) \sim \frac{1}{2}\Gamma(|\lambda|)\left(\frac{x}{2}\right)^{-|\lambda|}$ $(x \downarrow 0)$.

B.2 GH Distributions on $\mathbb{R}^d$

A GH distribution has the following pdf:

$$d_{GH_d(\lambda,\alpha,\beta,\delta,\mu,\Delta)}(x) = a(\lambda,\alpha,\beta,\delta,\mu,\Delta)\left(\sqrt{\delta^2+||x-\mu||^2_{\Delta^{-1}}}\right)^{\lambda-\frac{d}{2}} K_{\lambda-\frac{d}{2}}\left(\alpha\sqrt{\delta^2+||x-\mu||^2_{\Delta^{-1}}}\right) e^{\langle\beta,x-\mu\rangle},$$

where $a(\lambda,\alpha,\beta,\delta,\mu,\Delta)$ is the normalization constant:

$$a(\lambda,\alpha,\beta,\delta,\mu,\Delta) = \frac{(\alpha^2-||\beta||^2_\Delta)^{\lambda/2}}{(2\pi)^{d/2}|\Delta|^{\frac{1}{2}}\alpha^{\lambda-d/2}\delta^\lambda K_\lambda(\delta\sqrt{\alpha^2-||\beta||^2_\Delta})}.$$

The GH parameters $(\lambda,\alpha,\beta,\delta,\mu,\Delta)$ take the following values:

$$\lambda\in\mathbb{R},\quad \alpha,\delta\in\mathbb{R}_+,\quad \beta,\mu\in\mathbb{R}^d,\quad \Delta\in\mathbb{P}_d,\quad \begin{array}{ll} \delta\geq 0, 0\leq||\beta||_\Delta<\alpha, & \text{if } \lambda>0, \\ \delta>0, 0\leq||\beta||_\Delta<\alpha, & \text{if } \lambda=0, \\ \delta>0, 0\leq||\beta||_\Delta\leq\alpha, & \text{if } \lambda<0, \end{array}$$

where $\delta=0$ or $\alpha=||\beta||_\Delta$ is a limiting case. The GH distribution is symmetric if and only if $\beta=\mathbf{0}$ and $\mu=0$. The symmetric GH has the following elliptical pdf:

$$d_{SGH_d(\lambda,\alpha,\delta,\Delta)}(x) = \frac{\alpha^{\frac{d}{2}}}{(2\pi)^{\frac{d}{2}}|\Delta|^{\frac{1}{2}}\delta^\lambda K_\lambda(\delta\alpha)}\left(\sqrt{\delta^2+||x||^2_{\Delta^{-1}}}\right)^{\lambda-\frac{d}{2}} K_{\lambda-\frac{d}{2}}\left(\alpha\sqrt{\delta^2+||x||^2_{\Delta^{-1}}}\right),$$

where $\nu(t)$ in equation (2) is given by a GIG distribution.

B.3 NIG Distributions on $\mathbb{R}^d$

The NIG distribution $NIG_d(\alpha,\beta,\delta,\mu,\Delta)$ has the following pdf (v. Hammerstein, 2010, p.74):

$$d_{NIG_d(\alpha,\beta,\delta,\mu,\Delta)}(x) \propto \left(\sqrt{\delta^2+||x-\mu||^2_{\Delta^{-1}}}\right)^{-\frac{d+1}{2}} K_{\frac{d+1}{2}}\left(\alpha\sqrt{\delta^2+||x-\mu||^2_{\Delta^{-1}}}\right) e^{\langle\beta,x-\mu\rangle}.$$

B.4 VG Distributions on $\mathbb{R}^d$

The VG distribution $VG_d(\lambda,\alpha,\beta,\mu,\Delta)$ has the following pdf (v. Hammerstein, 2010, p.74):[12]

$$d_{VG_d(\lambda,\alpha,\beta,\mu,\Delta)}(x) \propto (||x-\mu||_{\Delta^{-1}})^{\lambda-\frac{d}{2}} K_{\lambda-\frac{d}{2}}(\alpha||x-\mu||_{\Delta^{-1}})\, e^{\langle\beta,x-\mu\rangle}.$$

References

D. Applebaum. *Lévy processes and stochastic calculus.* second edition, Cambridge University Press, 2009.

12. The VG pdf is bounded at $x=\mu$ if and only if $\lambda>\frac{d}{2}$.

N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

E. O. Barndorff-Nielsen. Processes of normal inverse gaussian type. *Finance and Stochastics*, 2:41–68, 1998.

E. O. Barndorff-Nielsen and K. Prause. Apparent scaling. *Finance and Stochastics*, 5: 103–113, 2001.

O. E. Barndorff-Nielsen and C. Halgreen. Infinite divisibility of the hyperbolic and generalized inverse gaussian distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 38:309–312, 1977.

O. E. Barndorff-Nielsen and C. Halgreen. The variance gamma (v.g.) model for share market returns. *Journal of Business*, 63:511–524, 1990.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, 1984.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publisher, 2004.

M. L. Bianchi, S.T. Rachev, Y.S. Kim, and F.J. Fabozzi. Tempered infinitely divisible distributions and processes. *Theory of Probability and Its Applications (TVP), Society for Industrial and Applied Mathematics (SIAM)*, 55(1):59–86, 2010.

S. Bochner. Lectures on fourier integrals. with an author's supplement on monotonic functions, stieltjes integrals, and harmonic analysis. In *Princeton University Press, Princeton, NJ*. 1959.

B. Boots, G. Gordon, and A. Gretton. Hilbert space embeddings of predictive state representations. *Uncertainty in Artificial Intelligence (UAI)*, 2013.

W. Breymann and D. Lüthi. ghyp: A package on generalized hyperbolic distributions. 2013.

P. Carr, H. Geman, D. B. Madan, and M. Yor. The fine structure of asset returns: an empirical investigation. *Journal of Business*, 75:305–332, 2002.

Y. Chen, M. Welling, and A. Smola. Super-Samples from Kernel Herding. In *Uncertainty in Artificial Intelligence (UAI)*. 2010.

R. Cont and P. Tankov. *Financial Modelling with Jump Processes*. Boca Raton: Chapman & Hall CRC Press, 2004.

K. v. Harn F. W. Steutel. *Infinite Divisibility of Probability Distributions on the Real Line*. Monogr. Textb. Pure Appl. Math., vol. 259, Marcel Dekker Inc., 2004.

K. Fukumizu and C. Leng. Gradient-based kernel method for feature extraction and variable selection. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2123–2131. 2012.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel Measures of Conditional Dependence. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 489–496. 2008.

K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, pages 3753–3783, 2013.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Annual Conference on Neural Information Processing Systems (NIPS)*. 2008.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, 2012.

E. Grosswald. The student t-distribution of any degree of freedom is infinitely divisible. *Zeit. Wahrsch. Verw. Gebiete*, 36:103–109, 1976.

S. Grünewälder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. In *International Conference on Machine Learning (ICML)*, pages 535–542, 2012.

M. Kanagawa, Y. Nishiyama, A. Gretton, and K. Fukumizu. Filtering with State-Observation Examples via Kernel Monte Carlo Filter. *Neural Computation*, 28:382–444, 2016.

W. H. Lee. Continuous and discrete properties of stochastic processes. *PhD thesis, The University of Nottingham*, 2010.

J. R. Lloyd and Z. Ghahramani. Statistical Model Criticism using Kernel Two Sample Test. In *Annual Conference on Neural Information Processing Systems (NIPS)*. 2015.

B. D. Madan, P. Carr, and E. C. Chang. The variance gamma process and option pricing. *European Finance Review*, 2:79–105, 1998.

L. McCalman, S. O'Callaghan, and F. Ramos. Multi-modal estimation with kernel embeddings for learning motion models. In *IEEE International Conference on Robots and Automation (ICRA)*, 2013.

S. Mika, B. Schölkopf, A. Smola, K. Müller, M. Scholz, and G. Rätsch. Kernel PCA and de-noising in feature spaces. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 536–542, 1999.

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from Distributions via Support Measure Machines. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 10–18. 2012.

Y. Nishiyama, A. Boularias, A. Gretton, and K. Fukumizu. Hilbert Space Embeddings of POMDPs. In *Uncertainty in Artificial Intelligence (UAI)*, pages 644–653, 2012.

Y. Nishiyama, M. Kanagawa, A. Gretton, and K. Fukumizu. Model-based Kernel Sum Rule. In *arXiv: 1409.5178*. 2014.

Y. Nishiyama, A. H. Afsharinejad, S. Naruse, B. Boots, and L. Song. The Nonparametric Kernel Bayes' Smoother. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2016.

J. Nolan. Bibliography on stable distributions, processes and related topics. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.295.9970&rep=rep1&type=pdf`, 2013a.

J. Nolan. Multivariate elliptically contoured stable distributions: theory and estimation. *Computational Statistics*, 28(5):2067–2089, 2013b.

K. Prause. *The generalized hyperbolic model: estimation, financial derivatives, and risk measures.* Ph.D. thesis University of Freiburg, 1999.

S. T. Rachev, Y. S. Kim, M. L. Bianchi, and F. J. Fabozzi. *Financial Models with Levy Processes and Volatility Clustering.* Wiley & Sons, 2011.

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In *Annual Conference on Neural Information Processing Systems (NIPS)*. 2007.

C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. *MIT Press, Cambridge, MA*, 2006.

K. Rawlik, M. Toussaint, and S. Vijayakumar. Path Integral Control by Reproducing Kernel Hilbert Space Embedding. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

J. Rosiński. Tempering stable processes. *Stochastic Processes and Their Applications*, 117 (6):677–707, 2007.

G. Samorodnitsky and M. S. Taqqu. *Stable non-Gaussian random processes : stochastic models with infinite variance.* Chapman & Hall, 1994.

K. Sato. Lévy processes and infinitely divisible distributions. *Cambridge University Press*, 1999.

B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, Cambridge, 2002.

W. Schoutens. *Lévy Processes in Finance: Pricing Financial Derivatives.* Chichester: Wiley, 2003.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring Density Estimation via Reproducing Kernel Moment Matching. *International Conference on Machine Learning (ICML)*, pages 992–999, 2008.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems. In *International Conference on Machine Learning (ICML)*, pages 961–968, 2009.

L. Song, B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola. Hilbert Space Embeddings of Hidden Markov Models. In *International Conference on Machine Learning (ICML)*, pages 991–998, 2010.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel Belief Propagation. *Journal of Machine Learning Research - Proceedings Track*, 15:707–715, 2011.

L. Song, K. Fukumizu, and A. Gretton. Kernel embedding of conditional distributions. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

O. Thorin. An extension of the notion of a generalized Γ-convolution. *Scandinavian Actuarial Journal*, pages 141–149, 1978.

E. A. F. v. Hammerstein. *Generalized hyperbolic distributions: Theory and applications to CDO pricing*. Ph.D. thesis University of Freiburg, 2010.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.

V.M. Zolotarev. *One-dimensional stable distributions*. Translations of mathematical monographs, American Mathematical Society, 1986.