

# Characterizing the function space for Bayesian kernel models

Natesh S. Pillai<sup>1</sup>

Qiang Wu<sup>1,2,3</sup>

Feng Liang<sup>1</sup>

Sayan Mukherjee<sup>1,2,3</sup>

Robert L. Wolpert<sup>1,4</sup>

NSP2@STAT.DUKE.EDU

QIANG@STAT.DUKE.EDU

FENG@STAT.DUKE.EDU

SAYAN@STAT.DUKE.EDU

WOLPERT@STAT.DUKE.EDU

<sup>1</sup>*Institute of Statistics and Decision Sciences*

*Duke University*

*Durham, NC 27708, USA*

<sup>2</sup>*Institute for Genome Sciences & Policy*

<sup>3</sup>*Department of Computer Science*

<sup>4</sup>*Nicholas School of the Environment and Earth Sciences*

*Duke University*

*Durham, NC 27708, USA*

**Editor:** Version: 1.9 Date: 01/25/2007

## Abstract

Kernel methods have been very popular in the machine learning literature in the last ten years, mainly in the context of Tikhonov regularization algorithms. In this paper we study a coherent Bayesian kernel model based on an integral operator whose domain is a space of signed measures. Priors on the signed measures induce prior distributions on their image functions under the integral operator. We study several classes of signed measures and their images, and identify general classes of measures whose images are dense in the reproducing kernel Hilbert space (RKHS) induced by the kernel. This gives a function-theoretic foundation for some nonparametric prior specifications commonly-used in Bayesian modeling, including Gaussian processes and Dirichlet processes, and suggests generalizations.

A general framework for the construction of priors on signed measures using Lévy processes is described. A computational approach for sampling from the posterior distributions is presented, and illustrated for a univariate regression and a high-dimensional classification problem.

**Keywords:** Reproducing Kernel Hilbert Space, non-parametric Bayesian methods, Lévy processes, Dirichlet processes, integral operator, Gaussian processes

## 1. Introduction

Kernel methods have a long history in statistics and applied mathematics (Schoenberg, 1942; Aronszajn, 1950; Parzen, 1963; de Boor and Lynch, 1966; Micchelli and Wahba, 1981; Wahba, 1990) and have had a tremendous resurgence in the machine learning literature in the last ten years (Poggio and Girosi, 1990; Vapnik, 1998; Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004). Much of this resurgence was due to the popularization of classification algorithms such as support vector machines (SVMs) (Cortes and Vapnik,

1995) that are particular instantiations of the method of regularization of Tikhonov (1963). Many machine learning algorithms and statistical estimators can be summarized by the following penalized loss functional (Evgeniou et al., 2000; Hastie et al., 2001, Section 5.8)

$$\hat{f} = \arg \min_{f \in \mathcal{H}} [L(f, \text{data}) + \lambda \|f\|_K^2], \quad (1)$$

where  $L$  is a loss function,  $\mathcal{H}$  is often an infinite-dimensional reproducing kernel Hilbert space (RKHS),  $\|f\|_K^2$  is the norm of a function in this space, and  $\lambda$  is a tuning parameter chosen to balance the trade-off between fitting errors and the smoothness of the function. The data is assumed to be drawn independently from a distribution  $\rho(x, y)$  with  $x \in \mathcal{X} \subset \mathbb{R}^d$  and  $y \in \mathcal{Y} \subset \mathbb{R}$ . Due to the representer theorem (Kimeldorf and Wahba, 1971) the solution of the penalized loss functional will be a kernel

$$\hat{f}(x) = \sum_{i=1}^n w_i K(x, x_i), \quad (2)$$

where  $\{x_i\}_{i=1}^n$  are the  $n$  observed input or explanatory variables. The statistical learning community as well as the approximation theory community has studied and characterized properties of the RKHS for various classes of kernels (DeVore et al., 1989; Zhou, 2003).

Probabilistic versions and interpretations of kernel estimators have been of interest going back to the work of Hájek (1961, 1962) and Kimeldorf and Wahba (1971). More recently a variety of kernel models with a Bayesian framework applied to the finite representation from the representer theorem have been proposed (Tipping, 2001; Sollich, 2002; Chakraborty et al., 2005). However, the direct adoption of the finite representation is not a fully Bayesian model since it depends on the (arbitrary) training data sample size (see remark 3 for more discussion). In addition, this prior distribution is supported on a finite-dimensional subspace of the RKHS. Our coherent fully Bayesian approach requires the specification of a prior distribution over the entire space  $\mathcal{H}$ .

A continuous, positive semi-definite kernel on a compact space  $\mathcal{X}$  is called a *Mercer* kernel. The RKHS for such a kernel can be characterized (Mercer, 1909; König, 1986) as

$$\mathcal{H}_K = \left\{ f \mid f(x) = \sum_{j \in \Lambda} a_j \phi_j(x) \text{ with } \sum_{j \in \Lambda} a_j^2 / \lambda_j < \infty \right\}, \quad (3)$$

where  $\{\phi_j\} \subset \mathcal{H}$  and  $\{\lambda_j\} \subset \mathbb{R}_+$  are the orthonormal eigenfunctions and the corresponding non-increasing eigenvalues of the integral operator with kernel  $K$  on  $L^2(\mathcal{X}, \mu(du))$ ,

$$\lambda_j \phi_j(x) = \int_{\mathcal{X}} K(x, u) \phi_j(u) \mu(du) \quad (4)$$

and where  $\Lambda := \{j : \lambda_j > 0\}$ . The eigenfunctions  $\{\phi_j\}$  depend on the measure  $\mu(du)$ , but the eigenvalues  $\{\lambda_j\}$  and the RKHS do not (König (1986), page 143). This suggests specifying a prior distribution over  $\mathcal{H}$  by placing one on the parameter space

$$\mathcal{A} = \left\{ \{a_j\} \mid \sum_{j \in \Lambda} a_j^2 / \lambda_j < \infty \right\}$$

as in (Johnstone, 1998; Wasserman, 2005, Section 7.2). There are serious computational and conceptual problems with specifying a prior distribution on this infinite-dimensional set. In particular, only in special cases are the eigenfunctions  $\{\phi_j\}$  and eigenvalues  $\{\lambda_j\}$  available in closed form.

Another approach, the *Bayesian kernel model*, is to study the class of functions expressible as kernel integrals

$$\mathcal{G} = \left\{ f \mid f(x) = \int_{\mathcal{X}} K(x, u) \gamma(du), \quad \gamma \in \Gamma \right\}, \quad (5)$$

for some space  $\Gamma \subseteq \mathcal{B}(\mathcal{X})$  of signed Borel measures. Any prior distribution on  $\Gamma$  induces one on  $\mathcal{G}$ . The natural question that arises in this Bayesian approach is:

For what spaces  $\Gamma$  of signed measures is the RKHS  $\mathcal{H}_K$  identical to the linear space  $\text{span}(\mathcal{G})$  spanned by the Bayesian kernel model?

The space  $\mathcal{G}$  is the range  $\mathcal{L}_K[\Gamma]$  of the integral operator  $\mathcal{L}_K : \Gamma \rightarrow \mathcal{G}$  given by

$$\mathcal{L}_K[\gamma](x) := \int_{\mathcal{X}} K(x, u) \gamma(du). \quad (6)$$

Informally (we will be more precise in Section 2) we can characterize  $\Gamma$  as the range of the inverse operator  $\mathcal{L}_K^{-1} : \mathcal{H}_K \rightarrow \Gamma$ . The requirements on  $\Gamma$  for the equivalence between  $\mathcal{L}_K[\Gamma]$  and  $\mathcal{H}_K$  is the primary focus of this paper and in Section 2 we formalize and prove the following proposition:

**Proposition 1** *For  $\Gamma = \mathcal{B}(\mathcal{X})$ , the space of all signed Borel measures,  $\mathcal{G} = \mathcal{H}_K$ .*

The proposition asserts that the Bayesian kernel model and the penalized loss model both operate in the same function space when  $\Gamma$  includes all signed measures.

This result lays a theoretical foundation from a function analytic perspective for the use of two commonly used prior specifications: Dirichlet process priors (Ferguson, 1973; West, 1992; Escobar and West, 1995; MacEachern and Müller, 1998; Müller et al., 2004) and Lévy process priors (Wolpert et al., 2003; Wolpert and Ickstadt, 2004).

## 1.1 Overview

In Section 2, we formalize and prove the above proposition. Prior distributions are placed on the space of signed measures in Section 4 using Lévy, Dirichlet, and Gaussian processes. In Section 5 we provide two examples using slightly different process prior distributions for a univariate regression problem and a high dimensional classification problem. This illustrates the use of these process priors for posterior inference. We close in Section 6 with a brief discussion.

**Remark 2** *Equation (5) is a Fredholm integral equation of the first kind (Fredholm, 1900). The problem of estimating an unknown measure  $\gamma$  for a specified element  $f \in \mathcal{H}_K$  is ill-posed (Hadamard, 1902) in the sense that small changes in  $f$  may give rise to large changes in estimates of  $\gamma$ . It was precisely the study of this problem that led Tikhonov (1963) to his regularization method, in a study of problems in numerical analysis such as interpolation or*

*Gauss quadrature. Bayesian methods for interpolation and Gauss quadrature were explored by Diaconis (1988). A Bayesian method using Lévy process priors to address numerically ill-posed problems was developed by Wolpert and Ickstadt (2004). We will return to this relation between robust statistical estimation and numerically stable methods in the discussion.*

**Remark 3** *Due to the relation between regularization and Bayes estimators the finite representation is a MAP (maximal a posterior) estimator (Wahba, 1999; Poggio and Girosi, 1990). However, functions in the RKHS having a posterior probability very close to that of the MAP estimator need not have a finite representation so building a prior only on the finite representation is problematic if one wants to estimate the full posterior on the entire RKHS. Thus the prior used to derive the MAP estimate is essentially the same as those used in Tipping (2001); Sollich (2002); Chakraborty et al. (2005). This will lead to serious computational and conceptual difficulties when the full posterior must be simulated.*

## 2. Characterizing the function space of the kernel model

In this section we formalize the relationship between the RKHS and the function space induced by the Bayesian kernel model.

### 2.1 Properties of the RKHS

Let  $\mathcal{X} \subset \mathbb{R}^d$  be compact and  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  a continuous, positive semi-definite (Mercer) kernel. Consider the space of functions

$$\mathcal{H} = \left\{ f \mid f(x) = \sum_{j=1}^n a_j K(x, x_j) : n \in \mathbb{N}, \{x_j\} \subset \mathcal{X}, \{a_j\} \subset \mathbb{R} \right\}$$

with an inner product  $\langle \cdot, \cdot \rangle_K$  extending

$$\langle K(\cdot, x_i), K(\cdot, x_j) \rangle_K := K(x_i, x_j).$$

The Hilbert space closure  $\mathcal{H}_K$  of  $\mathcal{H}$  in this inner-product is the RKHS associated with the kernel  $K$  (Aronszajn, 1950). The kernel is “reproducing” in the sense that each  $f \in \mathcal{H}_K$  satisfies

$$f(x) = \langle f, K_x \rangle_K$$

for all  $x \in \mathcal{X}$ , where  $K_x(\cdot) := K(\cdot, x)$ .

A well-known alternate representation of the RKHS is given by an orthonormal expansion (Aronszajn 1950, extended to arbitrary measures by König 1986; see Cucker and Smale 2001). Let  $\{\lambda_j\}$  and  $\{\phi_j\}$  be the nonincreasing eigenvalues and corresponding complete orthonormal set of eigenvectors of the operator  $\mathcal{L}_K$  of Equation (6), restricted to the Hilbert space  $L^2(\mathcal{X}, du)$  of measures  $\gamma(du) = \gamma(u)du$  with square-integrable density functions  $\gamma \in L^2(\mathcal{X}, du)$ . Mercer’s theorem (Mercer, 1909) asserts the uniform and absolute convergence of the series

$$K(u, v) = \sum_{j=1}^{\infty} \lambda_j \phi_j(u) \phi_j(v), \tag{7}$$

whereupon with  $\Lambda := \{j : \lambda_j > 0\}$  we have

$$\mathcal{H}_K = \left\{ f = \sum_{j \in \Lambda} a_j \phi_j \mid \sum_{j \in \Lambda} \lambda_j^{-1} a_j^2 < \infty \right\}.$$

## 2.2 Bayesian kernel models and integral operators

Recall the Bayesian kernel model was defined by

$$\mathcal{G} = \left\{ \mathcal{L}_K[\gamma](x) := \int_{\mathcal{X}} K(x, u) \gamma(du), \quad \gamma \in \Gamma \right\},$$

where  $\Gamma$  is a space of signed Borel measures on  $\mathcal{X}$ . We wish to characterize the space  $\mathcal{L}_K^{-1}(\mathcal{H}_K)$  of Borel measures mapped into the RKHS  $\mathcal{H}_K$  of Equation (3). A precise characterization is difficult and instead we will find a subclass  $\Gamma \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$  which will be large enough in practice, in the sense that  $\mathcal{L}_K(\Gamma)$  is dense in  $\mathcal{H}_K$ .

First we study the image under  $\mathcal{L}_K$  of four classes of measures: (1) those with square integrable (Lebesgue) density functions; (2) all finite measures with Lebesgue density functions; (3) the set of discrete measures; and (4) linear combinations of all of these. Then we will extend these results to the general case of Borel measures (see Appendix A for proofs).

We first examine the class  $L^2(\mathcal{X}, du)$ , viewed as the space of finite measures on  $\mathcal{X}$  with square-integrable density functions with respect to Lebesgue measure; in a slight abuse of notation we write  $\gamma(du) = \gamma(u)du$ , using the same letter  $\gamma$  for the measure and its density function. Since  $\mathcal{X}$  is compact and  $K$  bounded,  $\mathcal{L}_K$  is a positive compact operator on  $L^2(\mathcal{X}, du)$  with a complete ortho-normal system (CONS)  $\{\phi_j\}$  of eigenfunctions with nonincreasing eigenvalues  $\{\lambda_j\} \subset \mathbb{R}_+$  satisfying Equation (7). Each  $\gamma \in L^2(\mathcal{X}, du)$  admits a unique expansion  $\gamma = \sum_j a_j \phi_j$ , with  $\|\gamma\|_2^2 = \sum_j a_j^2 < \infty$ . The image under  $\mathcal{L}_K$  of the measure  $\gamma(du) := \gamma(u)du$  with Lebesgue density function  $\gamma$  may be expressed as the  $L^2$ -convergent sum

$$\mathcal{L}_K[\gamma](x) = \sum_j \lambda_j a_j \phi_j(x).$$

**Proposition 4** *For every  $\gamma \in L^2(\mathcal{X}, du)$ ,  $\mathcal{L}_K[\gamma] \in \mathcal{H}_K$  and*

$$\|\mathcal{L}_K[\gamma]\|_K^2 = \langle \mathcal{L}_K[\gamma], \gamma \rangle_2.$$

*Consequently,  $L^2(\mathcal{X}, du) \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$ .*

The following corollary illustrates that the space  $L^2(\mathcal{X}, du)$  is too small for our purpose—i.e., that important functions  $f \in \mathcal{L}_K^{-1}(\mathcal{H}_K)$  fail to lie in  $L^2(\mathcal{X}, du)$ .

**Corollary 5** *If the set  $\Lambda := \{j : \lambda_j > 0\}$  is finite, then  $\mathcal{L}_K(L^2(\mathcal{X}, du)) = \mathcal{H}_K$ ; otherwise  $\mathcal{L}_K(L^2(\mathcal{X}, du)) \subsetneq \mathcal{H}_K$ . The latter occurs whenever  $K$  is strictly positive definite and the RKHS is infinite-dimensional.*

Thus only for finite dimensional RKHS's is the space of square integrable functions sufficient to span the RKHS. In almost all interesting non-parametric statistics problems, the RKHS is infinite-dimensional.

Next we examine the space of integrable functions  $L^1(\mathcal{X}, du)$ , a larger space than  $L^2(\mathcal{X}, du)$  when  $\mathcal{X}$  is compact.

**Proposition 6** *For every  $\gamma \in L^1(\mathcal{X}, du)$ ,  $\mathcal{L}_K[\gamma] \in \mathcal{H}_K$ . Consequently,  $L^1(\mathcal{X}, du) \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$ .*

Another class of functions to be considered is the space of finite discrete measures,

$$\mathcal{M}_D = \left\{ \mu = \sum_j c_j \delta_{x_j} : \{c_j\} \subset \mathbb{R}, \{x_j\} \subset \mathcal{X}, \sum_j |c_j| < \infty \right\},$$

where  $\delta_x$  is the Dirac measure supported at  $x \in \mathcal{X}$  (the sum may be finite or infinite). This class will arise naturally when we examine Lévy and Dirichlet processes in Section 4.3.

**Proposition 7** *For every  $\mu \in \mathcal{M}_D$ ,  $\mathcal{L}_K[\mu] \in \mathcal{H}_K$ . Consequently,  $\mathcal{M}_D \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$ .*

By Proposition 6 and 7 the space  $\mathcal{M}$  spanned by  $L^1(\mathcal{X}, du) \cup \mathcal{M}_D$  is a subset of  $\mathcal{L}_K^{-1}(\mathcal{H}_K)$ . The range of  $\mathcal{L}_K$  on just the elements of  $\mathcal{M}_D$  with finite support is precisely  $\mathcal{H}$ , linear combinations of the  $\{K_{x_j}\}_{x_j \in \mathcal{X}}$ ; thus

**Proposition 8**  *$\mathcal{L}_K(\mathcal{M})$  is dense in  $\mathcal{H}_K$  with respect to the RKHS norm.*

Let  $\mathcal{B}_+(\mathcal{X})$  denote the cone of all finite nonnegative Borel measures on  $\mathcal{X}$  and  $\mathcal{B}(\mathcal{X})$  the set of signed Borel measures. Clearly every  $\mu \in \mathcal{B}(\mathcal{X})$  can be written uniquely as  $\mu = \mu_+ - \mu_-$  with  $\mu_+, \mu_- \in \mathcal{B}_+(\mathcal{X})$ . The set  $\mathcal{B} \setminus \mathcal{M}$  contains those Borel measures that are singular with respect to the Lebesgue measure. In this context, the set  $\mathcal{M} = \mathcal{M}_D \cup L^1(\mathcal{X}, du)$  contains the Borel measures that can be used in practice. The above results, Propositions 6 and 4, also hold if we replace the Lebesgue measure with a Borel measure. It is natural to compare  $\mathcal{B}(\mathcal{X})$  with  $\mathcal{L}_K^{-1}(\mathcal{H}_K)$ .

**Proposition 9**  *$\mathcal{B}(\mathcal{X}) \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$ .*

We close this section by showing that even  $\mathcal{B}(\mathcal{X})$  need not exactly characterize the class  $\mathcal{L}_K^{-1}(\mathcal{H}_K)$ . The proof of Proposition 6 implies that

$$\|\mathcal{L}_K[\gamma]\|_K^2 = \iint_{\mathcal{X} \times \mathcal{X}} K(x, u) \gamma(x) \gamma(u) dx du. \quad (8)$$

From the above it is apparent that  $\mathcal{L}_K[\gamma] \in \mathcal{H}_K$  holds only if  $\mathcal{L}_K[\gamma]$  is well defined and the quantity on the right hand side of (8) is finite. If the kernel is smooth and vanishes at certain  $x, u \in \mathcal{X}$ , then (8) can be finite even if  $\gamma \notin L^1(\mathcal{X}, du)$ . For example in the case of polynomial kernels  $\delta'_x$ , the functional derivatives of the Dirac measure  $\delta_x$ , are mapped into  $\mathcal{H}_K$ .

**Proposition 10**  *$\mathcal{B}(\mathcal{X}) \subsetneq \mathcal{L}_K^{-1}(\mathcal{H}_K(\mathcal{X}))$ .*

**Proof**

We construct an infinite signed measure  $\gamma$  satisfying  $\mathcal{L}_K[\gamma] \in \mathcal{H}_K$ . As in Example 1 below, let

$$K(x, u) := x \wedge u - xu$$

be the covariance kernel for the Brownian bridge on the unit interval  $\mathcal{X} = [0, 1]$  (as usual, “ $x \wedge u$ ” denotes the minimum of two real numbers  $x, u$ ). Consider the improper  $\mathbf{Be}(0, 0)$  distribution

$$\gamma(\mathrm{d}u) = \frac{\mathrm{d}u}{u(1-u)},$$

with image under the integral operator

$$f(x) := \mathcal{L}_K[\gamma](x) = -x \log(x) - (1-x) \log(1-x).$$

The function  $f(x)$  satisfies  $f(0) = 0 = f(1)$  and has finite RKHS norm

$$\|f\|_K^2 = -2 \int_0^1 \frac{\log(x)}{1-x} \mathrm{d}x = \frac{\pi^2}{3},$$

so  $f(x)$  is in the the RKHS (see Example 1). Thus the infinite signed measure  $\gamma(\mathrm{d}s)$  is in  $\mathcal{L}_K^{-1}[\mathcal{H}_K]$  but not in  $\mathcal{B}(\mathcal{X})$ , so  $\mathcal{L}_K^{-1}[\mathcal{H}_K]$  is larger than the space of finite signed measures. ■

### 3. Two Concrete Examples

In this section we construct two explicit examples to help illustrate the ideas of Section 2.

**Example 1 (Brownian bridge)** *On the space  $\mathcal{X} = [0, 1]$  consider the kernel*

$$K(x, u) := (x \wedge u) - xu,$$

*which is jointly continuous and the covariance function for the Brownian bridge (Rogers and Williams, 1987, §IV.40) and hence a Mercer kernel. The eigenfunctions and eigenvalues of Equation (4) for Lebesgue measure  $\mu(\mathrm{d}u) = \mathrm{d}u$  are*

$$\lambda_j = \frac{1}{j^2 \pi^2} \quad \phi_j(x) = \sqrt{2} \sin(j\pi x).$$

*The associate integral operator of Equation (6) is*

$$\begin{aligned} \mathcal{L}_K[\gamma](x) &:= \int_{\mathcal{X}} K(x, u) \gamma(\mathrm{d}u) \\ &= (1-x) \int_{[0, x]} u \gamma(\mathrm{d}u) + x \int_{[x, 1]} (1-u) \gamma(\mathrm{d}u), \end{aligned}$$

*mapping any  $\gamma(\mathrm{d}u) = \gamma(u)\mathrm{d}u$  with  $\gamma \in L^1(\mathcal{X}, \mathrm{d}u)$  to a function  $f(x) = \mathcal{L}_K[\gamma](x)$  that satisfies the boundary conditions  $f(0) = 0 = f(1)$  and, for almost every  $x \in \mathcal{X}$ ,*

$$\begin{aligned}
 f(x) &= (1-x) \int_0^x u \gamma(u) du + x \int_x^1 (1-u) \gamma(u) du \\
 f'(x) &= \int_x^1 \gamma(u) du - \int_0^x u \gamma(u) du \\
 f''(x) &= -\gamma(x)
 \end{aligned}$$

and hence, by Equation (8) and integration by parts,

$$\begin{aligned}
 \|f\|_K^2 &= \int_0^1 f(x) \gamma(x) dx \\
 &= \int_0^1 -f(x) f''(x) dx \\
 &= \int_0^1 f'(x)^2 dx.
 \end{aligned}$$

Evidently the RKHS is just

$$\begin{aligned}
 \mathcal{H}_K &= \left\{ f(x) = \sum_{j=1}^{\infty} a_j \sqrt{2} \sin(j\pi x) \mid \sum_{j=1}^{\infty} \pi^2 j^2 a_j^2 < \infty \right\} \\
 &= \{f \text{ in } L^2(\mathcal{X}, du) \mid f(0) = 0 = f(1) \text{ and } f' \in L^2(\mathcal{X}, du)\},
 \end{aligned}$$

the subspace of the Sobolev space  $H_{+1}(\mathcal{X})$  satisfying Dirichlet boundary conditions (Mazja, 1985, Section 1.1.4), and

$$\begin{aligned}
 \mathcal{L}_K^{-1}(\mathcal{H}_K) &= \left\{ \gamma(x) = \sum_{j=1}^{\infty} a_j \sqrt{2} \sin(j\pi x) \mid \sum_{j=1}^{\infty} \frac{a_j^2}{\pi^2 j^2} < \infty \right\} \\
 &= \{\gamma = f'' \mid f, f' \in L^2(\mathcal{X}, du), f(0) = 0 = f(1)\},
 \end{aligned}$$

a subspace of  $H_{-1}(\mathcal{X})$ .

**Example 2 (Splines on a circle)** The kernel function for first order splines on the real line is

$$K(x, u) := |x - u| \quad x, u \in \mathbb{R}$$

and the corresponding RKHS norm is

$$\|f\|_K^2 = \int_{-\infty}^{\infty} f'(x)^2 dx.$$

However, since the domain is not compact the spectrum of the associated integral operator on  $L^2(\mathbb{R}, du)$  is continuous rather than discrete, the approach of Section 2 does not apply.

Instead we consider the case of splines with periodic boundary conditions. On the space  $\mathcal{X} = [0, 1]$  we consider the kernel function

$$\begin{aligned}
 K(x, u) &= \sum_{j=1}^{\infty} \frac{1}{2\pi^2 j^2} \cos(2\pi j|u - x|) \\
 &= \frac{1}{2} \left( |x - u| - \frac{1}{2} \right)^2 - \frac{1}{24} \quad 0 < x, u < 1
 \end{aligned}$$

The eigenfunctions and eigenvalues of Equation (4) for Lebesgue measure  $\mu(du) = du$  are

$$\begin{aligned} \phi_{2j-1}(x) &:= \sqrt{2} \sin(2\pi jx) & \lambda_{2j-1} &= \frac{1}{4\pi^2 j^2} \\ \phi_{2j}(x) &:= \sqrt{2} \cos(2\pi jx) & \lambda_{2j} &= \frac{1}{4\pi^2 j^2} \end{aligned} \quad j \in \mathbb{N}.$$

The RKHS norm for this kernel is

$$\|f\|_K^2 = \int_0^1 f'(x)^2 dx$$

and the RKHS is

$$\mathcal{H}_K = \left\{ f(x) = \sum_{j=1}^{\infty} \sqrt{2} [a_j \sin(2\pi jx) + b_j \cos(2\pi jx)] \mid \sum_{j=1}^{\infty} 4\pi^2 j^2 (a_j^2 + b_j^2) < \infty \right\}$$

the subspace of the Sobolev space  $H_{+1}(\mathcal{X})$  satisfying periodic boundary conditions and orthogonal to the constants (Wahba, 1990, Section 2.1) and

$$\mathcal{L}_K^{-1}(\mathcal{H}_K) = \left\{ \gamma(x) = \sum_{j=1}^{\infty} \sqrt{2} [a_j \sin(\pi jx) + b_j \cos(\pi jx)] \mid \sum_{j=1}^{\infty} \frac{a_j^2 + b_j^2}{4j^2\pi^2} < \infty \right\},$$

a subspace of  $H_{-1}(\mathcal{X})$ .

Elements in each RKHS above that have the finite representation,

$$f(x) = \sum_{i=1}^m c_i K(x, x_i), \quad m < \infty$$

are splines. In the first example, these functions are linear splines that vanish at  $\{0, 1\}$ . In the second example if the coefficients sum to zero ( $\sum_{i=1}^m c_i = 0$ ), then these functions are linear splines with periodic boundary conditions; if the coefficients do not sum to zero then they are quadratic splines with periodic boundary conditions.

#### 4. Bayesian Kernel Models

Our goal from Section 1 is to present a coherent Bayesian framework for non-parametric function estimation in a RKHS. Suppose we observe data (with noise),  $\{(x_i, y_i)\} \subset \mathcal{X} \times \mathbb{R}$  from the linear regression model

$$y_i = f(x_i) + \varepsilon_i \tag{9}$$

where we assume  $\{\varepsilon_i\}$  are independent  $\mathbf{No}(0, \sigma^2)$  random variables with unknown variance  $\sigma^2$ , and  $f(\cdot)$  is an unknown function we wish to estimate. For a fixed kernel we assume  $f \in \mathcal{H}_K$ . Recall that the integral operator  $\mathcal{L}_K$  maps  $\mathcal{M}(\mathcal{X})$  into  $\mathcal{H}_K$  and in particular  $\mathcal{L}_K(\mathcal{M}(\mathcal{X}))$  is dense in  $\mathcal{H}_K$ . Therefore, we assume that

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du) \tag{10}$$

where  $Z(du) \in \mathcal{M}(\mathcal{X})$  is a signed measure on  $\mathcal{X}$ . If we put a prior on  $\mathcal{M}(\mathcal{X})$ , we are in essence putting a prior on the functions  $f \in \mathcal{G}$ .

Our measurement error model (9) gives us the following likelihood for the data  $D := \{(x_i, y_i)\}_{i=1}^n$

$$L(D|Z) \propto \prod_{i=1}^n \exp \left[ -\frac{1}{2\sigma^2} (y_i - f(x_i))^2 \right]. \quad (11)$$

With a prior distribution on  $Z$ ,  $\pi(Z)$ , we can obtain the posterior density function given data

$$\pi(Z|D) \propto L(D|Z) \pi(Z), \quad (12)$$

which implies a posterior distribution for  $f$  via the integral operator (10).

#### 4.1 Priors on $\mathcal{M}$

A random signed measure  $Z(du)$  on  $\mathcal{X}$  can be viewed as a stochastic process on  $\mathcal{X}$ . The practice of specifying a prior on  $\mathcal{M}(\mathcal{X})$  via a stochastic process is ubiquitous in non-parametric Bayesian analysis. Gaussian processes and Dirichlet processes are two commonly used stochastic processes to generate random measures.

We first apply the results of Section 2 to Gaussian process priors (Rasmussen and Williams, 2006, Section 6) and then to Lévy process priors (Wolpert et al., 2003; Tu et al., 2006). We also remark that Dirichlet processes can be constructed from Lévy process priors.

#### 4.2 Gaussian Processes

Gaussian processes are canonical examples of stochastic processes used for generating random measures. They have been used extensively in the machine learning and statistics community with promising results in practice and theory (Kimeldorf and Wahba, 1971; Chakraborty et al., 2005; Rasmussen and Williams, 2006; Ghosal and Roy, 2006).

We consider two modeling approaches using Gaussian process priors:

- i. Model I: Placing a prior directly on the space of functions  $f(x)$  by sampling from paths of the Gaussian process with its covariance structure defined via a kernel  $K$ ;
- ii. Model II: Placing a prior on the random signed measures  $Z(du)$  on  $\mathcal{X}$  by using a Gaussian process prior for  $Z(du)$  which implies a prior on the function space defined by the kernel model in Equation (10).

For both approaches we can characterize the function space spanned by the kernel model. The first approach is the more standard approach for non-parametric Bayesian inference using Gaussian processes while the later is an example of our Bayesian kernel model. However, as pointed out by (Wahba, 1990, Section 1.4) the random functions from the first approach will be almost surely outside the RKHS induced by the kernel.

We first state some classical results on the sample paths of Gaussian processes. We then use these properties and the results of Section 2 to characterize the function spaces of the two models.

## 4.2.1 SAMPLE PATHS OF GAUSSIAN PROCESSES

Consider a Gaussian process  $\{Z_u, u \in \mathcal{X}\}$  on a probability space  $\{\Omega, \mathcal{A}, \mathbf{P}\}$  having covariance functions determined by a kernel function  $K$ . Let  $\mathcal{H}_K$  be the corresponding RKHS and let the mean  $m$  be contained in the RKHS,  $m \in \mathcal{H}_K$ . Then the following zero-one law holds:

**Theorem 11** (*Kallianpur (1970), Theorem 5.1*) *If  $\{Z_u, u \in \mathcal{X}\}$  is a Gaussian process with covariance  $K$  and mean  $m \in \mathcal{H}_K$ , and  $\mathcal{H}_K$  is infinite dimensional, then*

$$\mathbf{P}(Z_\bullet \in \mathcal{H}_K) = 0.$$

*The probability measure is assumed to be complete.*

Thus the sample paths of the Gaussian process are almost surely outside  $\mathcal{H}_K$ . However, there exists a RKHS  $\mathcal{H}_R$  that is bigger than  $\mathcal{H}_K$  that contains the sample paths almost surely. To construct such an RKHS we first need to define nuclear dominance.

**Definition 12** *Given two kernel functions  $R$  and  $K$ ,  $R$  dominates  $K$  (written as  $R \succ K$ ) if  $\mathcal{H}_K \subseteq \mathcal{H}_R$ .*

Given the above definition of dominance the following operator can be defined:

**Theorem 13** (*Lukić and Beder, 2001*) *Let  $R \succ K$ . Then*

$$\|g\|_R \leq \|g\|_K, \quad \forall g \in \mathcal{H}_K.$$

*There exists a unique linear operator  $L : \mathcal{H}_R \rightarrow \mathcal{H}_R$  whose range is contained in  $\mathcal{H}_K$  such that*

$$\langle f, g \rangle_R = \langle Lf, g \rangle_K, \quad \forall f \in \mathcal{H}_R, \forall g \in \mathcal{H}_K.$$

*In particular*

$$LR_u = K_u, \quad \forall u \in \mathcal{X}.$$

*As an operator into  $\mathcal{H}_R$ ,  $L$  is bounded, symmetric, and positive.*

*Conversely, let  $L : \mathcal{H}_R \rightarrow \mathcal{H}_R$  be a positive, continuous, self-adjoint operator then*

$$K(s, t) = \langle LR_s, R_t \rangle_R, \quad s, t \in \mathcal{X}$$

*defines a reproducing kernel on  $\mathcal{X}$  such that  $K \leq R$ .*

$L$  is the dominance operator of  $\mathcal{H}_R$  over  $\mathcal{H}_K$  and this dominance is called nuclear if  $L$  is a nuclear or trace class operator (a compact operator for which a trace may be defined that is finite and independent of the choice of basis). We denote nuclear dominance as  $R \succ K$ .

#### 4.2.2 IMPLICATIONS FOR THE FUNCTION SPACES OF THE MODELS

Model I placed a prior directly on the space of functions using sample paths from the Gaussian process with covariance structure defined by the kernel  $K$ . Theorem 11 states that sample paths from this Gaussian process are not contained in  $\mathcal{H}_K$ . However, there exists another RKHS  $\mathcal{H}_R$  with kernel  $R$  which does contain the sample path if  $R$  has nuclear dominance over  $K$ .

**Theorem 14** (*Lukić and Beder, 2001*) *Let  $K$  and  $R$  be two reproducing kernels. Assume that the RKHS  $\mathcal{H}_R$  is separable. A necessary and sufficient condition for the existence of a Gaussian process with covariance  $K$  and mean  $m \in \mathcal{H}_R$  and with trajectories in  $\mathcal{H}_R$  with probability 1 is that  $R \succ K$ .*

The implication of this theorem is that we can find a function space  $\mathcal{H}_R$  that contains functions generated by the Gaussian process defined by covariance function  $K$ .

Model II places a prior on random signed measures  $Z(du)$  on  $\mathcal{X}$  by using a Gaussian process prior for  $Z(du)$ . This implies a prior of the space of functions spanned by the kernel model in Equation (10). This space  $\mathcal{G}$  is contained in  $\mathcal{H}_K$  by our results in Section 2. This is due to the fact that any sample path from a continuous Gaussian process on a compact domain  $\mathcal{X}$  is in  $L^1$  and therefore the corresponding function from the integral (10) is still in  $\mathcal{H}_K$ .

### 4.3 Lévy processes

Lévy processes offer an alternative to Gaussian processes in non-parametric Bayesian modeling. Dirichlet processes and Gaussian processes with a particular covariance structure can be formulated from the framework of Lévy processes. For the sake of simplicity in exposition, we will use the univariate setting  $\mathcal{X} = [0, 1]$  to illustrate the construction of random signed measures using Lévy processes. The extension to the multivariate setting is straightforward and outlined in Appendix B.

A stochastic process  $Z := \{Z_u \in \mathbb{R} : u \in \mathcal{X}\}$  is called a *Lévy process* if it satisfies the following conditions:

1.  $Z_0 = 0$  almost surely.
2. For any integer  $m \in \mathbb{N}$  and any  $0 = u_0 < u_1 < \dots < u_m$ , the random variables  $\{Z_{u_j} - Z_{u_{j-1}}\}$ ,  $1 \leq j \leq m$  are independent (Independent increments property).
3. The distribution of  $Z_{s+u} - Z_s$  does not depend on  $s$  (Temporal homogeneity or stationary increments property).
4. The sample paths of  $Z$  are almost surely right continuous and have left limits, i.e., are “càdlàg”.

Familiar examples of Lévy processes include Brownian motion, Poisson processes, and gamma processes. The following celebrated theorem characterizes Lévy processes.

**Theorem 15 (Lévy-Khintchine)**  *$Z$  is a Lévy process if and only if the characteristic function of  $Z_u : u \geq 0$  has the following form:*

$$\mathbb{E}[e^{i\lambda Z_u}] = \exp \left\{ u \left[ i\lambda a - \frac{1}{2}\sigma^2\lambda^2 + \int_{\mathbb{R}\setminus 0} [e^{i\lambda w} - 1 - i\lambda w 1_{\{|w|<1\}}(w)]\nu(dw) \right] \right\}, \quad (13)$$

where  $a \in \mathbb{R}$ ,  $\sigma^2 \geq 0$  and  $\nu$  is a nonnegative measure on  $\mathbb{R}\setminus 0$  with

$$\int_{\mathbb{R}\setminus 0} (1 \wedge |w|^2)\nu(dw) < \infty. \quad (14)$$

Note that (13) can be written as a product of two components,

$$\exp \left\{ iau\lambda - \frac{u\sigma^2}{2}\lambda^2 \right\} \times \exp \left\{ u \int_{\mathbb{R}\setminus 0} [e^{i\lambda w} - 1 - i\lambda w 1_{\{|w|<1\}}(w)] \nu(dw) \right\},$$

the characteristic functions of a Gaussian process and of a partially compensated Poisson process, respectively. This observation is the essence of the Lévy-Itô theorem (Applebaum, 2004, Theorem 2.4.16), which asserts that every Lévy process can be decomposed into the sum of two independent components: a “continuous process” (Brownian motion with drift) and a (possibly compensated) “pure jump” process. The three parameters  $(a, \sigma^2, \nu)$  in (13) uniquely determine a Lévy process where  $a$  denotes the drift term,  $\sigma^2$  denotes the variance (diffusion coefficient) of the Brownian motion, and  $\nu(dw)$  denotes the intensity of the jump process. The so-called “Lévy measure”  $\nu$  need not be finite, but (14) implies that  $\nu[(-\epsilon, \epsilon)^c] < \infty$  for each  $\epsilon > 0$  and so  $\nu$  is at least sigma-finite.

#### 4.3.1 PURE JUMP LÉVY PROCESSES

Pure jump Lévy processes are used extensively in non-parametric Bayesian statistics due to their computational amenability. In this section we first state an interpretation of these processes using Poisson random fields. We then describe Dirichlet and symmetric  $\alpha$ -stable processes.

#### 4.3.2 POISSON RANDOM FIELDS INTERPRETATION

Any pure jump Lévy process  $Z$  has a nice representation via a Poisson random field. Set  $\Delta Z_u := Z_u - \lim_{s \uparrow u} Z_s$ , the jump size at the location  $u$ . Set  $\Gamma = \mathbb{R} \times \mathcal{X}$ , the Cartesian product of  $\mathbb{R}$  with  $\mathcal{X}$ . For any sets  $A \subset \mathbb{R}\setminus 0$  bounded away from zero and  $B \subset \mathcal{X}$  we can define the counting measure

$$N(A \times B) := \sum_{s \in B} 1_A(\Delta Z_s). \quad (15)$$

The measure  $N$  defined above turns out to be a Poisson random measure on  $\Gamma$ , with mean measure  $\nu(dw)du$  where  $du$  is the uniform reference measure on  $\mathcal{X}$  (for instance the Lebesgue measure when  $\mathcal{X} = [0, 1]$ ). For any  $E \subset \Gamma$  with  $\mu = \int_E \nu(dw)du < \infty$  the random variable  $N(E)$  has a Poisson distribution with intensity  $\mu$ .

When  $\nu$  is a finite measure, the total number of jumps  $J \in \mathbb{N}$  of the process follows a Poisson distribution with finite intensity  $\mu(\Gamma)$ . When  $Z$  has a density with respect to the Lévy random field  $M$  with Lévy measure  $m$ ,  $Z_u$  has finite total variation and determines a finite measure  $Z(du) = dZ_u$ . In this case, any realization of  $Z(du)$  can be formulated as

$$Z(du) = \sum_{j=1}^J w_j \delta_{u_j}, \quad (16)$$

where  $(w_j, u_j) \in \Gamma$  are *i.i.d.* draws from  $\nu(dw)du$  representing the jump size and the jump location, respectively. Given a realization of  $Z(du) = \{u_j, w_j\}_{j=1}^J$ , Equation (10) reduces to

$$\int_{\mathcal{X}} K(x, u) Z(du) = \int_{\Gamma} K(x, u) N(dwdu) = \sum_{j=1}^J w_j K(x, u_j),$$

where  $N(dwdu)$  is a Poisson random measure as defined by (15). Then the likelihood for the data  $D := \{(x_i, y_i)\}_{i=1}^n$  is given by

$$L(D|Z) \propto \prod_{i=1}^n \exp \left[ -\frac{1}{2\sigma^2} \left( y_i - \sum_{j=1}^J w_j K(x_i, u_j) \right)^2 \right].$$

If the measure  $\nu(dw)du$  has a density function  $\nu(w, u)$  with respect to some finite reference measure  $m(dwdu)$ , then the prior density function for  $Z$  with respect to a Lévy( $m$ ) process is

$$\pi(Z) = \left[ \prod_{j=1}^J \nu(w_j, u_j) \right] e^{m(\Gamma) - \nu(\Gamma)}. \quad (17)$$

Using Bayes' theorem, we can calculate the posterior distribution for  $Z$  via (12).

When  $\nu$  is an infinite measure the number of jumps in the unit interval is countably infinite almost surely. However, if the Lévy measure satisfies

$$\int_{\mathbb{R}} (1 \wedge |w|) \nu(dw) < \infty, \quad (18)$$

then the sequence  $\{w_j\}$  is almost surely absolutely summable (i.e.  $\sum_{j=1}^{\infty} |w_j| < \infty$  a.s.) and we can still represent the process  $Z$  via the summation (16). Note that condition (18) is stronger than the integrability condition (14) in the Lévy-Khintchine theorem. This allows for the existence of Lévy processes with jumps that are not absolutely summable.

### 4.3.3 DIRICHLET PROCESS

The Dirichlet process is commonly used in non-parametric Bayesian analysis (Ferguson, 1973, 1974) mainly due to its analytical tractability. When passing from prior to posterior computations, it has been shown that the Dirichlet process is the only conjugate member of the whole class of normalized random measures with independent increments (James et al., 2005) so the posterior can be efficiently computed. Recently it has received much attention in the machine learning literature (Blei and Jordan, 2004; Xing et al., 2004, 2006). Though

Dirichlet processes are often defined via Dirichlet distributions, they can also be defined as a normalized Gamma process as noted by Ferguson (1973). A Gamma process is a pure jump Lévy process, which has the Lévy measure

$$\nu(dw) = aw^{-1} \exp\{-bw\}dw, \quad w > 0,$$

so at each location  $u$   $Z_u \sim \text{Gamma}(au, b)$ . Suppose  $Z_u$  is a  $\text{Gamma}(a, 1)$  process defined on  $\mathcal{X} = [0, 1]$ , then

$$\tilde{Z}_u = Z_u/Z_1$$

is the  $\mathbf{DP}(a du)$  Dirichlet process. Since the Dirichlet process is a random measure on probability distribution functions, it can be used when the target function  $f(x)$  is a probability density function. Dirichlet processes can also be used to model a general smooth function  $f(x)$  in combination with other random processes. For example, Liang et al. (2005) and Liang et al. (2006) consider a variation of the integral (10)

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du) = \int_{\mathcal{X}} w(u) K(x, u) F(du), \quad (19)$$

where the random signed measure  $Z(du)$  is modeled by a random probability distribution function  $F(du)$  and random coefficients  $w(u)$ . A Dirichlet process prior is specified for  $F$  and a Gaussian prior distribution is specified for  $w$ .

#### 4.3.4 SYMMETRIC $\alpha$ -STABLE PROCESS

Symmetric  $\alpha$ -stable processes are another class of Lévy processes, arising from symmetric  $\alpha$ -stable distributions. The symmetric  $\alpha$ -stable distribution has the following characteristic function:

$$\varphi(\eta) = \exp(-\gamma|\eta|^\alpha),$$

$\gamma$  is the dispersion parameter, and  $\alpha \in (0, 2]$  is the characteristic exponent. The case, when  $\gamma = 1$  is called the standard symmetric  $\alpha$ -stable (S $\alpha$ S) distribution. It has the following Lévy measure

$$\nu(dw) = \frac{\Gamma(\alpha + 1)}{\pi} \sin\left(\frac{\pi\alpha}{2}\right) |w|^{-1-\alpha} dw \quad \alpha \in (0, 2].$$

Two important cases of S $\alpha$ S distributions are the Gaussian when  $\alpha = 2$  and the Cauchy when  $\alpha = 1$ . Thus S $\alpha$ S processes allow us to model heavy or light tail processes by varying  $\alpha$ . One can verify that the Lévy measure is infinite for  $0 < \alpha \leq 2$  since  $\nu(\mathbb{R}) = \int_{\mathbb{R}} \nu(dw) = 2 \int_{(0, \infty]} \alpha w^{-1-\alpha} dw = \infty$ , and hence the process has an infinite number of jumps in any finite time interval. For implementation we omit jumps of negligible size (say, those smaller than some fixed  $\epsilon > 0$ ). Since the measure  $\nu(\cdot)$  is sigma-finite, there will be only finitely-many jumps of size greater than  $\epsilon$ , almost surely. Hence the support of our random measures reduces to

$$S_\epsilon \equiv (-\epsilon, \epsilon)^c \times [0, 1].$$

Given the jumps sizes and jump locations  $\{\{w_j, u_j\}, (w_j, u_j) \in S_\epsilon\}$ , and the number of jumps  $J$ , the prior probability density function (17) is

$$\pi(Z) = \left[ \prod_{j=1}^J |w_j| \right]^{1-\alpha} e^{2(\epsilon^{-1} - \epsilon^{-\alpha})} \alpha^J, \quad |w_j| \geq \epsilon \quad (20)$$

with respect to a Cauchy random field.

Using this prior is essentially the same as using a penalty term in a regularization approach. For the S $\alpha$ S process, we have

$$\log \pi(Z) \propto J \log \alpha + (1 - \alpha) \left( \sum_j \log |u_j| 1_{|u_j| > \epsilon} \right) + \text{constant}. \quad (21)$$

The first term is an AIC like penalty for the number of knots  $J$  and the second term is a LASSO-type penalty in log-scale. There is also a hidden penalty which shrinks all the coefficients with magnitude less than  $\epsilon$  to zero.

#### 4.4 Computational and modeling considerations

The computational and modeling issues involved in choosing process priors, especially in high dimensional settings, are at the heart of Bayesian non-parametrics. In this section we discuss these issues for the models discussed in the previous section.

A main challenge with Gaussian process models is that a finite dimensional representation of the sample path is required for computation. For low dimensional problems (say  $d \leq 3$ ), a reasonable approach is to place a grid on  $\mathcal{X}$ . Then we can approximate a continuous process  $Z$  by its values on the finitely many points  $\{u_j\}_{j=1}^m$  on the grid. Using this approximation, our kernel model (10) can be written as

$$f(x) = \sum_{j=1}^m w_j K(x, u_j),$$

and the implied prior distribution on  $(w_1, \dots, w_m)$  is a multivariate normal with mean and covariance structure as defined by the kernel  $K$  evaluated at points  $\{u_j\}$ . However, for high dimensional problems providing a grid is not feasible since the number of points in the grid grows exponentially with the dimension  $d$ . This is generally addressed by using the finite representation given by the representer theorem or by assuming a fixed effects model. In either case the set of knots are the input data points  $x_1, \dots, x_n$  and efficient computation of the posterior is possible. It is important to note however, that the prior being sampled in this model is not over  $\mathcal{X}$  but the restriction of  $\mathcal{X}$  to the data. Both the direct model and the kernel model will face this computational consideration and thus the computational cost will not differ significantly between models.

For pure jump processes discretization is not the bottleneck. The nature of the pure jump process ensures that the kernel model will have discrete knots. The key issue in using a pure jump processes to model multivariate data is that the knots of the model should be representative of samples drawn from the marginal distribution of the data  $\rho_X$ . This is a serious computational as well as modeling challenge, it is obvious that independently sampling each dimension will typically not be a good idea either in terms of computational time or modeling accuracy. In Section 5.2 we provide a kernel model that addresses this issue.

A theoretical and empirical comparison of the accuracy of the various process priors on a variety of function classes and data sets would be of interest, but is beyond the scope of this paper. Due to the extensive literature on Gaussian process models from theoretical as

well as practical perspectives (Rasmussen and Williams, 2006; Ghosal and Roy, 2006) our simulations will focus on two pure jump process models.

## 5. Posterior Inference

For the case of regression our model is

$$y_i = f(x_i) + \varepsilon_i \quad \text{for } x_i \in \mathcal{X}$$

with  $\{\varepsilon_i\}$  as i.i.d normal random variables and the unknown regression function  $f$  (which is assumed to be in  $\mathcal{H}_K$ ) is modeled as

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du).$$

In the case of binary regression we can use a probit model

$$\mathbf{P}(y_i = 1|x_i) = \Phi[f(x_i)], \quad (22)$$

where  $\Phi[\cdot]$  is the cumulative distribution function of the standard normal distribution.

In Section 4, we discussed specifying a prior on  $\mathcal{H}_K$  via the random measure  $Z(du)$ . The observed data add to our knowledge of both the “true function”  $f(\cdot)$  and the distribution of  $Z(du)$ . This information is used to update the prior and obtain the posterior density  $\pi(Z|D)$ . For pure jump measures  $Z(du)$  and most non-parametric models this update is computationally difficult because there is no closed-form expression for the posterior distribution. However, Markov chain Monte Carlo (MCMC) methods can be used to simulate the posterior distribution.

We will apply a Dirichlet process model to a high-dimensional binary regression problem and illustrate the use of Lévy process models on a univariate regression problem.

### 5.1 Lévy process model

Posterior inference for Lévy random measures has been less explored than Dirichlet and Gaussian processes. Wolpert et al. (2003) is a recent comprehensive reference on this topic. We use the methodology developed in this work for our model.

The random measure  $Z(du)$  is given by

$$Z(du) \sim \text{Lévy}(\nu(dw)du)$$

where

$$\nu(dw) = \frac{\Gamma(\alpha + 1)}{\pi} \sin\left(\frac{\pi\alpha}{2}\right) |w|^{-1-\alpha} 1_{\{|w|>\epsilon\}} dw \quad \alpha \in (0, 2]$$

is the Lévy measure (truncated) for the SaS process. Again (as in Section 4.3.4), since  $\nu(\mathbb{R}) = \infty$ , we omit jumps of size smaller than  $\epsilon$ . Any realization of the random measure  $Z(du)$  is an element of the parameter space

$$\Theta := \bigcup_{J=0}^{\infty} \left( (-\epsilon, \epsilon)^c \times [0, 1] \right)^J$$

with the prior probability density function given by Equation(20), with respect to a Cauchy random field.

## 5.1.1 TRANSITION PROBABILITY PROPOSAL

In this section, we describe an MCMC algorithm to simulate from  $\Theta$  according to the posterior distribution. We construct an irreducible transition probability distribution  $Q(d\theta^*|\theta)$  on the parameter space  $\Theta$  such whose stationary distribution is the desired posterior distribution.

Two different realizations from the parameter space  $\Theta$  may not have the same number of jumps, hence we use a birth-death process to model the number  $J$  of jumps. At any iteration step (indexed by  $t$ ) the state consists of  $J$  jump locations  $\{u_j\}$  of size  $\{w_j\}$ ,  $\theta_t = \{w_j, u_j\}_{j=1}^J$ . The transition probability algorithm, Algorithm 1, evaluates the transition probability to a new state  $\theta^*$  given the current state  $\theta$ .

---

**Algorithm 1:** Transition probability algorithm  $Q(\theta)$ .

---

**input :**  $0 < p_b, p_d < 1$ ,  $\tau > 0$ , current state  $\theta \in \Theta$

**return:** proposed new state  $\theta^*$  and its transition probability  $Q(\theta^*|\theta)$

Draw  $t \sim U[0, 1]$ ;

**if**  $t < 1 - p_b$  **then**

draw uniformly  $j \in \{1, \dots, J\}$ ; draw  $\gamma_1, \gamma_2 \sim \mathbf{No}(0, \tau^2)$ ;

$w_* \leftarrow w_j + \gamma_1$ ;  $u_* \leftarrow u_j + \gamma_2$ ;

**if**  $(|w_*| < \epsilon$  **or**  $t < p_d)$  **then**

$J \leftarrow J - 1$ ; delete  $(w_j, u_j)$ ;

$Q(\theta^*|\theta) \leftarrow \frac{(J+1)p_b}{2\epsilon^{-\alpha} \left( (1-p_b-p_d) \left[ \Phi\left(\frac{w_j+\epsilon}{\tau}\right) - \Phi\left(\frac{w_j-\epsilon}{\tau}\right) \right] + p_d \right)}$ ;

**else**

$Q(\theta^*|\theta) \leftarrow \left| \frac{w_*}{w_j} \right|$ ;  $w_j \leftarrow w_*$ ;  $u_j \leftarrow u_*$ ;

**else**

$J \leftarrow J + 1$ ;  $u_J \sim U[\mathcal{X}]$ ;  $w_J \sim \text{Birth}$ ;

$Q(\theta^*|\theta) \leftarrow \frac{2\epsilon^{-\alpha} \left( (1-p_d-p_b) \left[ \Phi\left(\frac{w_J+\epsilon}{\tau}\right) - \Phi\left(\frac{w_J-\epsilon}{\tau}\right) \right] + p_d \right)}{p_b J}$ ;

---

In the above algorithm,  $\mathbf{No}(0, \tau^2)$  denotes the normal distribution with mean 0 and variance  $\tau^2$  and  $\Phi(\cdot)$  denotes standard normal distribution function. The variables  $(p_b, p_d)$  represent the birth and death probabilities, respectively. There is an implicit update step, where a chosen point( $u_j$ ) is ‘updated’ with another point( $u_*$ ) with probability  $1 - p_b - p_d$ . In the birth step, a new point is sampled according to the density

$$\frac{\alpha|w|^{-1-\alpha}}{2\epsilon^{-\alpha}} \quad \epsilon > 0.$$

## 5.1.2 THE MCMC ALGORITHM

The MCMC algorithm, Algorithm 2, simulates draws from the posterior distribution. This is done by Metropolis-Hastings sampling using the transition probability algorithm above to generate a Markov chain whose equilibrium density is the posterior density.

---

**Algorithm 2:** MCMC algorithm
 

---

```

input : data  $D$ , number of iterations  $T$ , transition probability algorithm  $\mathcal{Q}(\theta)$ 

return: parameters drawn from the posterior  $\{\theta_i\}_{i=1}^T$ 

 $J \sim \mathbf{Po}(2\epsilon^{-\alpha})$ ; // initialize  $J$ 
for  $j \leftarrow 1$  to  $J$  do
    // initialize  $\theta(0)$ 
     $u_j \sim U[\mathcal{X}]$ ;  $w_j \sim \text{Birth}$ ;

for  $t \leftarrow 1$  to  $T$  do
    //  $t$ -th iteration of the Markov chain
     $\{\theta_*, Q(\theta_*|\theta_t)\} \leftarrow \mathcal{Q}(\theta(t))$ ; // call the transition probability algorithm
     $\log \pi(\theta_*|D) - \log \pi(\theta_t|D) = \log \frac{L(D|\theta_*)}{L(D|\theta_t)} + \log \frac{\pi(\theta_*)}{\pi(\theta_t)}$ ;
     $\zeta_* \leftarrow \log \pi(\theta_*|D) + \log Q(\theta_t|\theta_*) - \log \pi(\theta_t|D) - \log Q(\theta_*|\theta_t)$ ; // the
        Metropolis-Hastings log acceptance probability

     $e \sim \mathbf{Ex}(1)$ ;
    if  $e + \zeta_{t+1} > 0$  then  $\theta_{t+1} \leftarrow \theta_*$  else  $\theta_{t+1} \leftarrow \theta_t$ ;
    
```

---

The MCMC algorithm will provide us with  $T$  realizations of the jump parameters  $\{\theta_i\}_{i=1}^T$ . We assume that the chain reaches its stationary distribution after  $b$  iterations ( $b \ll T$ ). For each of the  $T - b$  realizations, we have a corresponding function

$$\hat{f}_t(x) = \sum_{i=1}^{J_t} w_{it} K(x, u_{it}),$$

where for the  $t$ -th realization  $J_t$  is the number of jumps,  $w_{it}$  is the magnitude of the  $i$ -th jump, and  $u_{it}$  is the position of the  $i$ -th jump. Point estimates can be made by averaging  $\hat{f}$  and credible intervals can be computed from the distribution of  $\hat{f}$  to provide an estimate of uncertainty.

### 5.1.3 ILLUSTRATION ON SIMULATED DATA

Data is generated from a noisy sinusoid

$$f(x_i) = \sin(2\pi x_i) + \varepsilon_i \quad \text{for } x \in [0, 1], \quad (23)$$

with  $\varepsilon_i \stackrel{iid}{\sim} \mathbf{No}(0, 0.01)$ ,  $\{x_i\}_{i=1}^{100}$  equally spaced in  $[0, 1]$ , and  $\{y_i\}_{i=1}^{100}$  computed by Equation (23). We applied the S $\alpha$ S model with  $\alpha = 1.5$  and a Gaussian kernel  $K(x, u) = \exp\{-(x - u)^2\}$  to this data. We set  $\epsilon = 0.01$  and  $(p_b, p_u, p_d) = (0.4, 0.2, 0.4)$ , in algorithms 1 and 2. In Figure 1a-d we plot the target sinusoid, the function realized at an iteration  $t$  of the Markov chain, and the jump locations and magnitudes of the random measure. In Figure 1e,f we provide a plot of the target function, realization of the data, and the 95% pointwise credible band – the 95% credible interval at each point  $x_i$ .

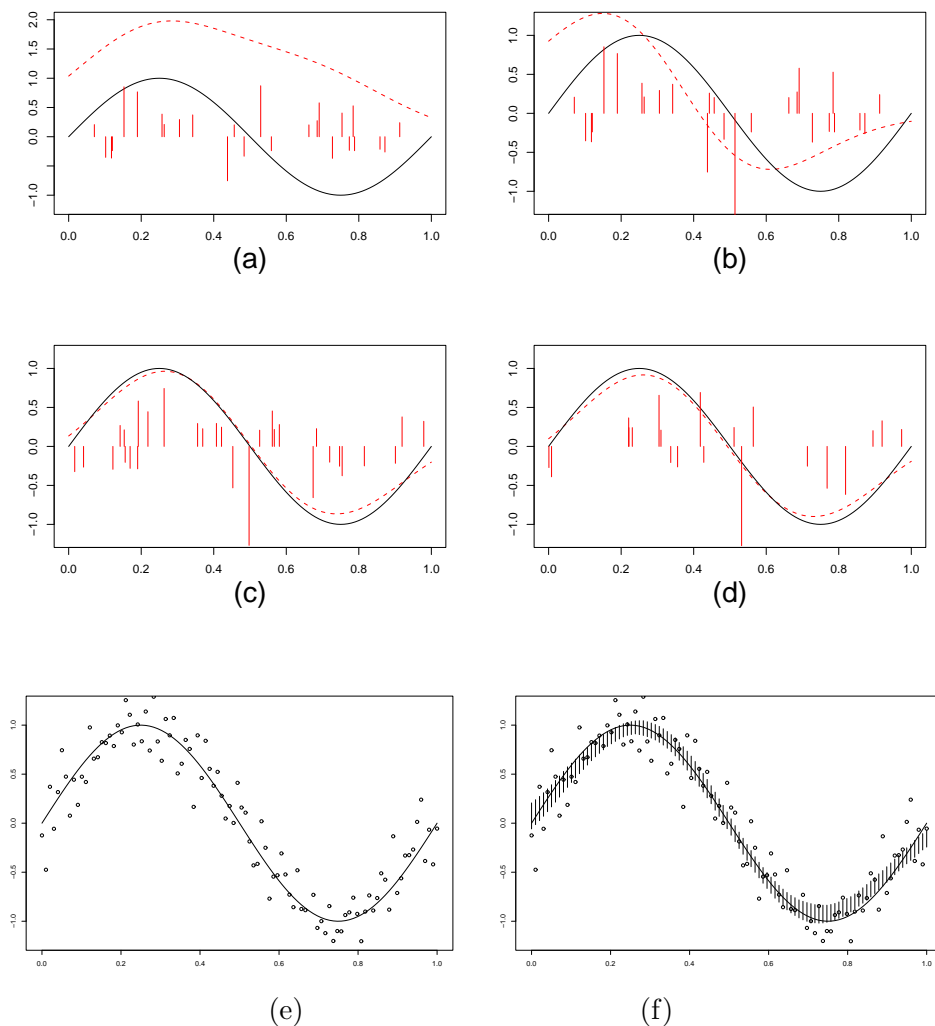


Figure 1: Plots of the target sinusoid (solid black line), the function realized at an iteration  $t$  of the Markov chain (dashed red line), and the jump locations and magnitudes of the measure (red spikes) for (a)  $t = 1$ , (b)  $t = 10$ , (c)  $t = 5 \times 10^3$ , and (d)  $t = 10^4$ . In (e) we present a realization of the simulated data (circles) and the underlying target sinusoid (solid line), and in (f) the 95% point-wise credible band for the data and the target sinusoid.

## 5.2 Classification of gene expression data

For Dirichlet processes there is extensive literature on exact posterior inference using MCMC methods (West, 1992; Escobar and West, 1995; MacEachern and Müller, 1998; Müller et al., 2004) as well as work on approximate inference using variational methods (Blei and Jordan, 2004). Recently Dirichlet process priors have been applied to a Bayesian kernel model

for high dimensional data. For example in Liang et al. (2006) and Liang et al. (2005) the Bayesian kernel model was used to classify gene expression data as well as digits, the MNIST database. We apply this model to gene expression data consisting of microarray gene expression profiles from 190 cancer samples and 90 normal samples (Ramaswamy et al., 2001; Mukherjee et al., 2003), over 16,000 genes.

The model is based upon the integral operator given in Equation (19)

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du) = \int_{\mathcal{X}} w(u) K(x, u) F(du),$$

where the random signed measure  $Z(du)$  is modeled by a random probability distribution function  $F(du)$  and a random weight function  $w(u)$ . We assume that the support of  $Z(du)$  and  $w(u)F(du)$  are equal. A key point in our model will be that if our estimate of  $F$  is discrete and puts masses  $w_i$  at support points (or “knots”)  $u_i$ , then the expression for  $f(\cdot)$  is simply

$$f(x) = \sum_i w(u_i) K(x, u_i).$$

In Liang et al. (2005) uncertainty about  $F$  is expressed using a Dirichlet process prior,  $\text{Dir}(\alpha, F_0)$ . The posterior after marginalization is also a Dirichlet distribution and given data  $(x_1, \dots, x_n)$  the posterior will have the following representation (Liang et al., 2005, 2006)

$$\hat{f}(x) = \frac{\alpha}{\alpha + n} \int w(u) K(x, u) F_0(du) + \frac{1}{\alpha + n} \sum_{i=1}^n w(x_i) K(x, x_i),$$

which can be approximated by the following discrete summation

$$\hat{f}(x) \approx \sum_{i=1}^n w_i K(x, x_i) \tag{24}$$

when  $\frac{\alpha}{n}$  is small and  $w_i = \frac{w(x_i)}{\alpha + n}$ . We specify a mixture-normal prior on the coefficients  $w_i$  as in Liang et al. (2005) and use the same MCMC algorithm to simulate the posterior.

Note that although Equation (24) has the same form as the representer theorem, it is derived from a very different formulation. In fact, when there is unlabeled data available –  $(x_{n+1}, \dots, x_{n+m})$  drawn from the margin  $\rho_X$  – our model has the following discrete representation

$$\hat{f}(x) = \sum_{i=1}^n w_i K(x, x_i) + \sum_{i=1}^m w_{i+n} K(x, x_{i+n}),$$

where  $w_\ell = \frac{w(x_\ell)}{\alpha + m + n}$ . The above form is identical to the one obtained via the manifold regularization framework (Belkin and Niyogi, 2004; Belkin et al., 2006). The two derivations are from different perspectives. This simple incorporation of unlabeled data into the model further illustrates the advantage of placing the prior over random measures in the Bayesian kernel model.

In our experiments we first applied a standard variation filter to reduce the number of genes to  $p = 2800$ . We then randomly assigned 20% of the samples from the cancer and non-cancer groups to training data and use the remaining 80% as test data. We performed two analyses on this data:

Analysis I – The training data were used in the model and the posterior probability was simulated for each point in the test set. A linear kernel was used.

Analysis II – The training and unlabeled test data were used in the model and the posterior probability was simulated for each point in the test set. A linear kernel was used.

The classification accuracy for Analyses I and II were 73% and 85%, respectively. The accuracy of the predictive models in Analysis I is comparable to that obtained for support vector machines in Mukherjee et al. (2003). Figure 2 displays boxplots of the posterior mean of the 72 the normal and 152 cancer samples for the two analyses.

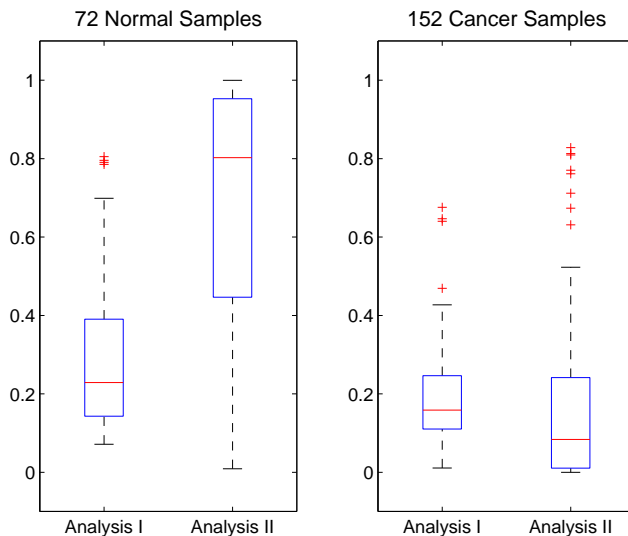


Figure 2: Boxplots of the posterior mean for normal and cancer samples with just the training data (Analysis I) and the training and unlabeled test data (Analysis II).

## 6. Discussion

The modeling objective underlying this paper is to formulate a coherent Bayesian perspective for regression using a RHKS model. This requires a firm theoretical foundation characterizing the function space that the Bayesian kernel model spans and the relation of this space to the RKHS. Our results in Section 2 are interesting in their own right, in addition to providing this foundation.

We examined the function class defined by the Bayesian kernel model, the integral of a kernel with respect to a signed Borel measure

$$\mathcal{G} = \left\{ f \mid f(x) = \int_{\mathcal{X}} K(x, u) \gamma(du), \quad \gamma \in \Gamma \right\}, \quad (25)$$

where  $\Gamma \subseteq \mathcal{B}(\mathcal{X})$ . We stated an equivalence under certain conditions of the function class  $\mathcal{G}$  and the RKHS induced by the kernel. This implies: (a) a theoretical foundation for the use of Gaussian processes, Dirichlet processes, and other jump processes for non-parametric Bayesian kernel models, (b) an equivalence between regularization approaches and the Bayesian kernel approach, and (c) an illustration of why placing a prior on the distribution is a more natural approach than placing a prior directly over functions.

Coherent non-parametric methods have been of great interest in the Bayesian community, however function analytic issues have not been considered. Conversely theoretical studies of RKHS have not approached the approximation and estimation problems from a Bayesian perspective (the exception to both of these are the works of Wahba (1990) and Diaconis (1988)). It is our view that the interface of these perspectives is a promising area of research for statisticians, computer scientists, and mathematicians and has both theoretical and practical implications.

A better understanding of this interface may lead to a better understanding of the following research problems:

1. Posterior consistency: It is natural to expect the posterior distribution to concentrate around the true function since the posterior distribution is a probability measure on the RKHS. A natural idea is to use the equivalence between the RKHS and our Bayesian model to exploit the well understood theory of RKHS in proving posterior consistency of the Bayesian kernel model. Tools such as concentration inequalities, uniform Glivenko-Cantelli classes, and uniform central limit theorems may be helpful.
2. Priors on function spaces: In this paper we discuss general function classes without concern for more subtle smoothness properties. An obvious question is can we use the same ideas to relate priors on measures and the kernel to specific classes of functions, such as Sobolev spaces. A study of the relation between integral operators and priors could lead to interesting and useful results for putting priors over specific function classes using the kernel model.
3. Comparison of process priors for modeling: A theoretical and empirical comparison of the accuracy of the various process priors on a variety of function classes and data sets would be of great practical importance and interest, especially for high dimensional problems.
4. Numerical stability and robust estimation: The original motivation for regularization methods was to provide numerical stability in solving Fredholm integral equation of the first kind. Our interest is that of providing robust non-parametric statistical estimates. A link between stability of operators and the generalization or predictive ability of regression estimates is known (Bousquet and Elisseeff, 2002; Poggio et al., 2004). Further developing this relation is a very interesting area of research and may be of importance for the posterior consistency of the Bayesian kernel model.

## Appendix A. Proofs of propositions

In this appendix we provide proofs for the propositions in Section 2.

### A.1 Proof for Proposition 4

It holds that

$$\|\mathcal{L}_K[\gamma]\|_K^2 = \left\| \sum_{j \in \Lambda} \lambda_j a_j \phi_j \right\|_K^2 = \sum_{j \in \Lambda} \frac{(\lambda_j a_j)^2}{\lambda_j} = \sum_{j \in \Lambda} \lambda_j a_j^2$$

which is upper bounded by  $\lambda_1 \sum_j a_j^2 < \infty$ . Hence  $\mathcal{L}_K[\gamma] \in \mathcal{H}_K$ . By direct computation, we have

$$\langle \mathcal{L}_K[\gamma], \gamma \rangle_2 = \left\langle \sum \lambda_j a_k \phi_j, \sum a_j \phi_j \right\rangle_2 = \sum \lambda_k a_j^2 = \|\mathcal{L}_K[\gamma]\|_K^2.$$

### A.2 Proof for Corollary 5

The first claim is obvious since both  $\mathcal{L}_K[L^2(\mathcal{X}, du)]$  and  $\mathcal{H}_K$  are the same finite dimensional space spanned by  $\{\phi_j\}_{j \in \Lambda}$ .

The second claim follows from the existence of the sequence  $(c_j)_{j \in \Lambda}$  such that

$$\sum_{j \in \Lambda} \frac{c_j^2}{\lambda_j} < \infty \quad \text{and} \quad \sum_{j \in \Lambda} \frac{c_j^2}{\lambda_j^2} = \infty.$$

For any such sequence, the function  $f = \sum_{j \in \Lambda} c_j \phi_j$  lies in  $\mathcal{H}_K$ . But by Proposition 4, one cannot find a  $\gamma \in L^2(\mathcal{X}, du)$  such that  $\mathcal{L}_K[\gamma] = f$ . A simple example is  $(c_j)_{j \in \Lambda} = (\lambda_j)_{j \in \Lambda}$ .

If  $K$  is strictly positive definite, then all its eigenvalues are positive. So the last claim holds.

### A.3 Proof for Proposition 6

Since  $K(u, v)$  is continuous on the compact set  $\mathcal{X} \times \mathcal{X}$ , it has a finite maximum  $\kappa^2 := \sup_{u, v} K(u, v) < \infty$ . Since  $L^2(\mathcal{X}, du)$  is dense in  $L^1(\mathcal{X}, du)$ , for every  $\gamma \in L^1(\mathcal{X}, du)$ , there exists a Cauchy sequence  $\{\gamma_n\}_{n \geq 1} \subset L^2(\mathcal{X}, du)$  which converges to  $\gamma$  in  $L^1(\mathcal{X}, du)$ . It follows from Proposition 4 that  $\mathcal{L}_K[\gamma_n] \in \mathcal{H}_K$  and

$$\|\mathcal{L}_K[\gamma_n]\|_K^2 = \int_{\mathcal{X}} \int_{\mathcal{X}} K(u, v) \gamma_n(u) du \gamma_n(v) dv \leq \kappa^2 \int_{\mathcal{X}} |\gamma_n(u)| du \int_{\mathcal{X}} |\gamma_n(v)| dv = \kappa^2 \|\gamma_n\|_1^2 < \infty.$$

Therefore we have  $\{\mathcal{L}_K[\gamma_n]\}_{n \geq 1} \subset \mathcal{H}_K$  and

$$\limsup_{n \rightarrow \infty} \sup_{m > n} \|\mathcal{L}_K[\gamma_n] - \mathcal{L}_K[\gamma_m]\|_K \leq \limsup_{n \rightarrow \infty} \sup_{m > n} \kappa \|\gamma_n - \gamma_m\|_1 = 0,$$

so  $\{\mathcal{L}_K[\gamma_n]\}_{n \geq 1}$  is a Cauchy sequence in  $\mathcal{H}_K$ . By completeness it converges to some  $f \in \mathcal{H}_K$ . The proof will be finished if we show  $\mathcal{L}_K[\gamma] = f$ .

By the reproducing property of  $\mathcal{H}_K$  convergence in the RKHS norm implies point-wise convergence for  $x \in \mathcal{X}$ , so  $L_K[\gamma_n](x) \rightarrow f(x)$  for every  $x$ .

In addition, for every  $x \in \mathcal{X}$ , we have

$$\lim_{n \rightarrow \infty} |\mathcal{L}_K[\gamma_n](x) - \mathcal{L}_K[\gamma](x)| \leq \int_{\mathcal{X}} |K(x, u)(\gamma_n(u) - \gamma(u))| du \leq \kappa^2 \|\gamma_n - \gamma\|_1 = 0,$$

which implies that  $\mathcal{L}_K[\gamma_n](x)$  also converges to  $\mathcal{L}_K[\gamma](x)$ . Hence  $\mathcal{L}_K[\gamma] = f \in \mathcal{H}_K$ .

#### A.4 Proof for Proposition 7

Let  $\gamma = \sum c_i \delta_{x_i} \in \mathcal{M}_D$ . Then  $\mathcal{L}_K[\gamma] = \sum c_i K_{x_i}$  and

$$\|\mathcal{L}_K[\gamma]\|_K^2 = \sum_{i,j} c_i K(x_i, x_j) c_j \leq \kappa^2 \left( \sum_i |c_i| \right)^2 < \infty.$$

Therefore, our conclusion holds.

#### A.5 Proof for Proposition 9

The arguments for Lebesgue measure hold if we replace the Lebesgue measure with any finite Borel measure. We denote the corresponding integral operator as  $\mathcal{L}_{K,\mu}$  and function space of integrable and square integrable functions as  $L_\mu^1(\mathcal{X})$  and  $L_\mu^2(\mathcal{X})$  respectively. Then

$$L_\mu^2(\mathcal{X}) \subset L_\mu^1(\mathcal{X}) \subset L_{K,\mu}^{-1}(\mathcal{H}_K).$$

Since the function  $1_{\mathcal{X}}(x) = 1$  lies in  $L_\mu^1(\mathcal{X})$  we obtain

$$\mathcal{L}_K(\mu) = \mathcal{L}_{K,\mu}(1_{\mathcal{X}}) = \int_{\mathcal{X}} K(\cdot, u) d\mu(u) \in \mathcal{H}_K.$$

This implies  $\mathcal{B}_+(\mathcal{X})$  lies in  $L_{K,\mu}^{-1}(\mathcal{H}_K)$  and so does  $\mathcal{B}(\mathcal{X})$ .

### Appendix B. Multivariate version of Lévy-Khintchine formula

Here we give the statement of the multivariate version of the Lévy-Khintchine formula (Applebaum, 2004, Corollary 2.4.20).

**Theorem 16 (Lévy-Khintchine)** *Let  $X$  be a  $d$ -dimensional Lévy process with characteristic function  $\phi_t(u) := \mathbb{E}(e^{i\langle u, X_t \rangle})$ ,  $u \in \mathbb{R}^d$ . Then there exists a unique vector  $a \in \mathbb{R}^d$ , a  $d \times d$  semi-positive definite matrix  $\sigma$ , and  $\nu$  a positive measure on  $\mathbb{R}^d \setminus 0$  with  $\int_{\mathbb{R}^d} (1 \wedge |u|^2) \nu(du) < \infty$  such that,*

$$\phi_t(u) = \exp \left\{ t \left[ i\langle u, a \rangle - \frac{1}{2} \langle u, \sigma u \rangle + \int_{\mathbb{R}^d \setminus 0} [e^{i\langle u, s \rangle} - 1 - i\langle u, s \rangle 1_{\{|s| < 1\}}(s)] \nu(ds) \right] \right\}$$

where  $\langle \cdot, \cdot \rangle$  denotes the standard inner product in  $\mathbb{R}^d$ .

The results we have presented extend to the multivariate case without complication. The simplest multivariate extension is to assume independence of the dimensions, however for small sample sizes and many dimensions this is not practical. This issue can be addressed by carefully inducing covariance structure in the model (Liang et al., 2005, 2006).

## References

- David Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, UK, 2004.
- Nachman Aronszajn. Theory of reproducing kernels. *T. Am. Math. Soc.*, 686:337–404, 1950.
- Mikhail Belkin and Partha Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(1-3):209–239, 2004.
- Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7: 2399–2434, 2006.
- David M. Blei and Michael I. Jordan. Variational methods for the Dirichlet process. In Brodley (2004). URL [http://www.aicml.cs.ualberta.ca/\\\_banff04/icml/pages/accepted.htm](http://www.aicml.cs.ualberta.ca/\_banff04/icml/pages/accepted.htm).
- Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.
- Carla E. Brodley, editor. *Machine Learning, Proceedings of the 21<sup>st</sup> International Conference (ICML 2004), Banff, Canada*. ACM Press, New York, NY, 2004. URL [http://www.aicml.cs.ualberta.ca/\\\_banff04/icml/pages/accepted.htm](http://www.aicml.cs.ualberta.ca/\_banff04/icml/pages/accepted.htm).
- Sounak Chakraborty, Malay Ghosh, and Bani K. Mallick. Bayesian non-linear regression for large  $p$  small  $n$  problems. *J. Am. Stat. Assoc.*, 2005. Under revision.
- Corinna Cortes and Vladimir N. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.
- Felipe Cucker and Stephen Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2001.
- Carl de Boor and Robert E. Lynch. On splines and their minimum properties. *J. Math. Mech.*, 15:953–969, 1966.
- Ronald A. DeVore, Ralph Howard, and Charles A. Micchelli. Optimal nonlinear approximation. *Manuskripta Mathematika*, 1989.
- Persi Diaconis. Bayesian numerical analysis. In Shanti S. Gupta and James O. Berger, editors, *Statistical decision theory and related topics, IV*, volume 1, pages 163–175. Springer-Verlag, New York, NY, 1988.
- Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, 90:577–588, 1995.
- Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

- Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Stat.*, 1: 209–230, 1973.
- Thomas S. Ferguson. Prior distributions on spaces of probability measures. *Ann. Stat.*, 2: 615–629, 1974.
- Erik Ivar Fredholm. Sur une nouvelle méthode pour la résolution du problème de Dirichlet. *Euvres complètes: publiées sous les auspices de la Kungliga svenska vetenskapsakademien par l'Institut Mittag-Leffler*, pages 61–68, 1900.
- Subhashis Ghosal and Anindya Roy. Posterior consistency of Gaussian process prior for nonparametric binary regression. *Ann. Stat.*, 34(5), October 2006. To appear.
- Jacques Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.
- Jaroslav Hájek. On linear statistical problems in stochastic processes. *Czechoslovak Math. J.*, 12(87):404–444, 1962.
- Jaroslav Hájek. On a property of normal distributions of any stochastic process. *Select. Transl. Math. Statist. and Probability*, 1:245–252, 1961.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.
- Lancelot F. James, Antonio Lijoa, and Igor Prünster. Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Stat.*, 33:105–120, 2005.
- Iain Johnstone. Function estimation in Gaussian noise: sequence models. Draft of a monograph, 1998.
- Gopinath Kallianpur. The role of reproducing kernel Hilbert spaces in the study of Gaussian processes. *Advances in Probability and Related Topics*, 2:49–83, 1970.
- Hermann König. *Eigenvalue distribution of compact operators*, volume 16 of *Operator Theory: Advances and Applications*. Birkhäuser, Basel, CH, 1986.
- George S. Kimeldorf and Grace Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2):495–502, 1971.
- Feng Liang, Sayan Mukherjee, and Mike West. Understanding the use of unlabelled data in predictive modeling. *Stat. Sci.*, 2006. To appear.
- Feng Liang, Sayan Mukherjee, Mike West, and Ming Liao. Nonparametric Bayesian kernel models. Discussion Paper 2005-09, Duke University ISDS, Durham, NC, 2005. URL [\emwww.stat.duke.edu/research/papers/](http://www.stat.duke.edu/research/papers/).
- Milan N. Lukić and Jay H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *T. Am. Math. Soc.*, 353(10):3945–3969, 2001.

- Stephen MacEachern and Peter Müller. Estimating mixture of Dirichlet process models. *J. Comput. Graph. Stat.*, pages 223–238, 1998.
- Vladimir G. Mazja. *Sobolev Spaces*. Springer-Verlag, New York, NY, 1985.
- Peter Müller, Fernando Quintana, and Gary Rosner. A method for combining inference across related nonparametric Bayesian models. *J. Am. Stat. Assoc.*, pages 735–749, 2004.
- James Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London A*, 209: 415–446, 1909.
- Charles A. Micchelli and Grace Wahba. Design problems for optimal surface interpolation. In Zvi Ziegler, editor, *Approximation Theory and Applications*, pages 329–348, 1981.
- Sayan Mukherjee, Pablo Tamayo, Simon Rogers, Ryan M. Rifkin, Anna Engle, Colin Campbell, Todd R. Golub, and Jill P. Mesirov. Estimating dataset size requirements for classifying DNA Microarray data. *Journal of Computational Biology*, 10:119–143, 2003.
- Emanuel Parzen. Probability density functionals and reproducing kernel Hilbert spaces. In Murray Rosenblatt, editor, *Proceedings of the Symposium on Time Series Analysis*, pages 155–169, New York, NY, 1963. John Wiley & Sons.
- Tomaso Poggio and Federico Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.
- Tomaso Poggio, Ryan M. Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, 2004.
- Sridhar Ramaswamy, Pablo Tamayo, Ryan M. Rifkin, Sayan Mukherjee, Chen-Hsiang Yeang, Michael Angelo, Christine Ladd, Michael Reich, Eva Latulippe, Jill P. Mesirov, Tomaso Poggio, William Gerald, Massimo Loda, Eric S. Lander, and Todd R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Nat. Aca. Sci.*, 98:149–54, 2001.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.
- L. Chris G. Rogers and David Williams. *Diffusions, Markov Processes, and Martingales*, volume 2. John Wiley & Sons, New York, NY, 1987. ISBN 0-471-91482-7.
- Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2001.
- Isaac J. Schoenberg. Positive definite functions on spheres. *Duke Mathematics Journal*, 9: 96–108, 1942.
- John S. Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

- Peter Sollich. Bayesian methods for Support Vector Machines: Evidence and predictive class probabilities. *Machine Learning*, 46(1-3):21–52, 2002.
- Andrei Nikolaevich Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Doklady*, 4:1035–1038, 1963.
- Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244, 2001.
- Chong Tu, Merlise A. Clyde, and Robert L. Wolpert. Lévy adaptive regression kernels. Discussion Paper 2006-08, Duke University ISDS, Durham, NC, 2006. URL <http://www.stat.duke.edu/research/papers/>.
- Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, NY, 1998.
- Grace Wahba. *Splines Models for Observational Data*, volume 59 of *Series in Applied Mathematics*. SIAM, Philadelphia, PA, 1990.
- Grace Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In Bernhard Schölkopf, Alexander J. Smola, Christopher J. C. Burges, and Rosanna Soentpiet, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 69–88. MIT Press, Cambridge, MA, 1999.
- Larry Wasserman. *All of Nonparametric Statistics*. Springer-Verlag, 2005.
- Mike West. Hyperparameter estimation in Dirichlet process mixture models. Discussion Paper 1992-03, Duke University ISDS, Durham, NC, 1992. URL <http://www.stat.duke.edu/research/papers/>.
- Robert L. Wolpert and Katja Ickstadt. Reflecting uncertainty in inverse problems: A Bayesian solution using Lévy processes. *Inverse Problems*, 20(6):1759–1771, 2004.
- Robert L. Wolpert, Katja Ickstadt, and Martin Bøgsted Hansen. A nonparametric Bayesian approach to inverse problems (with discussion). In José Miguel Bernardo, Maria Jesus Bayarri, James O. Berger, A. Phillip Dawid, David Heckerman, Adrian F. M. Smith, and Mike West, editors, *Bayesian Statistics 7*, pages 403–418, Oxford, UK, 2003. Oxford Univ. Press. ISBN 0-19-852615-6.
- Eric P. Xing, Roded Sharan, and Michael I. Jordan. Bayesian haplotype inference via the Dirichlet process. In Brodley (2004). URL <http://www.aicml.cs.ualberta.ca/~banff04/icml/pages/accepted.htm>.
- Eric P. Xing, Kyung-Ah Sohn, Michael I. Jordan, and Yee-Whye Teh. Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. In William Cohen and Andrew Moore, editors, *Machine Learning, Proceedings of the 23<sup>rd</sup> International Conference (ICML 2006)*, Pittsburgh, PA, New York, NY, 2006. ACM Press. URL <http://www.icml2006.org/icml2006/technical/accepted.html>.
- Ding-Xuan Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE T. Inform. Theory*, 49:1743–1752, 2003.